The PARA-suite: PAR-CLIP specific sequence read simulation and processing

Andreas Kloetgen [1, 2, 3], **Arndt Borkhardt** [2], **Jessica I Hoell** [2], **Alice C McHardy** Corresp. [1, 3]

[1] Department for Algorithmic Bioinformatics, Heinrich-Heine Universität Düsseldorf, Düsseldorf, Germany

[2] Department of Pediatric Oncology, Hematology and Clinical Immunology, Medical Faculty, Heinrich-Heine Universität Düsseldorf, Düsseldorf, Germany

[3] Computational Biology of Infection Research, Helmholtz Center for Infection Research, Braunschweig, Germany

Corresponding Author: Alice C McHardy
Email address: Alice.Mchardy@helmholtz-hzi.de

Background: Next-generation sequencing technologies have profoundly impacted biology over recent years. Experimental protocols, such as photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP), which identifies protein–RNA interactions on a genome-wide scale, commonly employ deep sequencing. With PAR-CLIP, the incorporation of photoactivatable nucleosides into nascent transcripts leads to high rates of specific nucleotide conversions during reverse transcription. So far, the specific properties of PAR-CLIP-derived sequencing reads have not been assessed in depth.

Methods: We here compared PAR-CLIP sequencing reads to regular transcriptome sequencing reads (RNA-Seq) to identify distinctive properties that are relevant for reference-based read alignment of PAR-CLIP datasets. We developed a set of freely available tools for PAR-CLIP data analysis, called the PAR-CLIP analyzer suite (PARA-suite). The PARA-suite includes error model inference, PAR-CLIP read simulation based on PAR-CLIP specific properties, a full read alignment pipeline with a modified Burrows–Wheeler Aligner algorithm and CLIP read clustering for binding site detection.

Results: We show that differences in the error profiles of PAR-CLIP reads relative to regular transcriptome sequencing reads (RNA-Seq) make a distinct processing advantageous. We examine the alignment accuracy of commonly applied read aligners on 10 simulated PAR-CLIP datasets using different parameter settings and identified the most accurate setup among those read aligners. We demonstrate the performance of the PARA-suite in conjunction with different binding site detection algorithms on several real PAR-CLIP and HITS-CLIP datasets. Our processing pipeline allowed the improvement of both alignment and binding site detection accuracy.

Availability: The PARA-suite toolkit and the PARA-suite aligner are available at https://github.com/akloetgen/PARA-suite and https://github.com/akloetgen/PARA-suite_aligner, respectively, under the GNU GPLv3 license.

The PARA-suite: PAR-CLIP specific sequence read simulation and processing

Andreas Kloetgen[1,2,3], Arndt Borkhardt[2], Jessica I. Hoell[2,§], Alice C. McHardy[1,3,§,*]

[1]Department of Algorithmic Bioinformatics, Heinrich Heine University, Düsseldorf, Germany.
[2]Department of Pediatric Oncology, Hematology and Clinical Immunology, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany.
[3]Computational Biology of Infection Research, Helmholtz Center for Infection Research, Braunschweig, Germany.
[§]These authors contributed equally to the work.
[*]To whom correspondence should be addressed.

Corresponding author: Alice C. McHardy; Inhoffenstr. 7, 38124 Braunschweig, Germany; AMC14@helmholtz-hzi.de

Short title: Simulating and processing PAR-CLIP data

Abstract

Background: Next-generation sequencing technologies have profoundly impacted biology over recent years. Experimental protocols, such as photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP), which identifies protein–RNA interactions on a genome-wide scale, commonly employ deep sequencing. With PAR-CLIP, the incorporation of photoactivatable nucleosides into nascent transcripts leads to high rates of specific nucleotide conversions during reverse transcription. So far, the specific properties of PAR-CLIP-derived sequencing reads have not been assessed in depth.

Methods: We here compared PAR-CLIP sequencing reads to regular transcriptome sequencing reads (RNA-Seq) to identify distinctive properties that are relevant for reference-based read alignment of PAR-CLIP datasets. We developed a set of freely available tools for PAR-CLIP data analysis, called the PAR-CLIP analyzer suite (PARA-suite). The PARA-suite includes error model inference, PAR-CLIP read simulation based on PAR-CLIP specific properties, a full read alignment pipeline with a modified Burrows–Wheeler Aligner algorithm and CLIP read clustering for binding site detection.

Results: We show that differences in the error profiles of PAR-CLIP reads relative to regular transcriptome sequencing reads (RNA-Seq) make a distinct processing advantageous. We examine the alignment accuracy of commonly applied read aligners on 10 simulated PAR-CLIP datasets using different parameter settings and identified the most accurate setup among those read aligners. We demonstrate the performance of the PARA-suite in conjunction with different binding site detection algorithms on several real PAR-CLIP and HITS-CLIP datasets. Our processing pipeline allowed the improvement of both alignment and binding site detection accuracy.

Availability: The PARA-suite toolkit and the PARA-suite aligner are available at https://github.com/akloetgen/PARA-suite and https://github.com/akloetgen/PARA-suite_aligner, respectively, under the GNU GPLv3 license.

Background

RNAs play a crucial role in cell survival and viability. Coding messenger RNAs (mRNAs), which are translated into proteins, and many other RNA species, such as small and long non-coding RNAs, ribosomal RNAs and transfer RNAs, are essential for the survival and proper functioning of the cells (Eddy 2001). Most RNAs maintain their function by working together with the so-called RNA-binding proteins (RBPs) (Glisovic, Bachorik et al. 2008). RBPs are involved in virtually all steps of the mRNA lifecycle, from polyadenylation, translocation and modification to translation (Hieronymus and Silver 2004). Thus it is not surprising that many RBPs that show aberrant functions or changes in expression patterns have been associated with disease progression or even with carcinogenesis (Lukong, Chang et al. 2008). For instance, the *FET* protein family, which consists of the three RBPs *FUS*, *EWSR1* and *TAF15*, is ubiquitously expressed and widely conserved in mammals. Genomic rearrangements, leading to mutant forms of these RBPs in humans, have been described as key players in sarcomas and leukemia (Tan and Manley 2009). More recently, two mutants of *FUS* causing amyotrophic lateral sclerosis have shown different RNA-binding patterns compared to their wild-type counterparts, supporting the importance of the function of *FUS* in mRNA processing (Hoell, Larsson et al. 2011).

Experimental protocols have been developed to analyze the functional network in which a particular RBP interacts. A promising method for this purpose is the photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP) technique (Hafner, Landthaler et al. 2010). When coupled with deep sequencing, it identifies the bound RNAs for a particular RBP on a genome-wide scale. First, the cells are supplied with a specific photoactivatable nucleoside, such as 4-thiouridine (4-SU), which is incorporated as an alternative to the respective nucleoside into nascent mRNA transcripts. Afterwards, the cells are treated with ultraviolet light at 365 nm to cross-link the amino acids of RBPs to the nucleotides of their bound RNA molecules. The incorporation of 4-SU instead of uridine results in nucleotide conversions from uridine to cytidine at all cross-linked sites containing a 4-SU during reverse transcription (a necessary step for preparing cDNA libraries for sequencing). This specific replacement is called a ‘T–C conversion’. T–C conversions can be used to distinguish between non-specifically bound RNA fragments (considered as contaminations) and those that are

specifically bound and cross-linked to the RBP of interest (Ascano, Hafner et al. 2012, Golumbeanu, Mohammadi et al. 2015). We recently published a detailed protocol for the PAR-CLIP procedure (Hoell, Hafner et al. 2014). Other CLIP protocols for the genome-wide identification of RBP targets are also frequently used, such as high-throughput sequencing of RNAs isolated by cross-linking and immunoprecipitation (HITS-CLIP, sometimes also called CLIP-seq) or the iCLIP protocol (Chi, Zang et al. 2009, König, Zarnack et al. 2010). The procedures, experimental designs and bioinformatic analysis of these different CLIP methods differ greatly and are still evolving. Recent reviews compare the strengths and weaknesses of the three methods in detail (Wang, Xiao et al. 2015, Danan, Manickavel et al. 2016). HITS-CLIP, for example, mainly introduces deletions of a single base at the cross-linked sites, whereas single nucleotide conversions do not seem to occur at a significant frequency (Zhang and Darnell 2011, Sugimoto, König et al. 2012).

Current sequencing platforms allow for the sequencing of mammalian transcriptome libraries with high coverage. Nowadays, the most commonly used next-generation sequencing (NGS) platforms are 454, Illumina, IonTorrent and PacBio (van Dijk, Auger et al. 2014). Depending on the sequencing platform and the sample type, sequencing errors vary in type and frequency. The errors that most commonly occur are substitution errors and indels of a few bases between the sequencing read and the reference sequence (large rearrangements, such as those leading to chimeras, are also possible errors but are not discussed here) (Laehnemann, Borkhardt et al. 2015). In an RNA-Seq dataset, a single transcript will be covered by sequencing reads in all its expressed coding exons (apart from, for example, amplification errors or alternative splicing variants). For common sequencing data types, such as RNA-Seq and DNA-Seq, designated read aligners have recently been developed. These include short read aligners, such as BWA (Li and Durbin 2009) or Bowtie (Langmead, Trapnell et al. 2009), and read aligners such as TopHat (Trapnell, Pachter et al. 2009), STAR (Dobin, Davis et al. 2013) or Subjunc (Liao, Smyth et al. 2013), which can also handle longer sequencing reads spanning exon–exon junctions. Specific software for the evaluation and analysis of the PAR- and HITS-CLIP sequencing data is needed to accommodate their unique error profiles (Kloetgen, Münch et al. 2015). For instance, the read aligner BWA PSSM (Kerpedjiev, Frellsen et al. 2014) makes use of a pre-defined position-specific scoring matrix to process the error-prone PAR-CLIP reads.

In general, the sequencing error profiles of RNA-Seq datasets, including PAR-CLIP data, can vary between different sequencing runs, depending on the sequencing machine, the experimental conditions and the biological properties of the sample (Laehnemann, Borkhardt et al. 2015, Schirmer, Ijaz et al. 2015). Here, we describe the PAR-CLIP analyzer suite (PARA-suite), which includes a PAR-CLIP read simulator, an error estimation tool for CLIP datasets and an alignment pipeline based on a novel alignment algorithm performing on-the-fly dataset-specific error estimation. The alignment pipeline thus automatically adjusts to the quality and error profiles of individual sequencing datasets. We compare PAR-CLIP sequencing reads to regular transcriptome sequencing reads (RNA-Seq) to identify the distinctive properties that are relevant for reference-based read alignment and RBP binding site detection from PAR-CLIP datasets. Generation of simulated PAR-CLIP datasets can be performed with the PARA-suite´s read simulator. The PARA-suite toolkit is available at https://github.com/akloetgen/PARA-suite and https://github.com/akloetgen/PARA-suite_aligner, implemented as an extension of BWA (henceforth referred to as BWA PARA). It is licensed under GNU GPLv3, and can be implemented in the programming languages Java and C.

Methods

2.1 Datasets and read aligners

We downloaded PAR-CLIP data for the *FET* family (*EWSR1*, *FUS* and *TAF15*) from the DRASearch database (https://trace.ddbj.nig.ac.jp/DRASearch/) with the accession number SRA025082 (Hoell, Larsson et al. 2011), the *HuR* dataset with the accession number SRR248532, the *MOV10* dataset with the accession number SRR490650 and the HITS-CLIP data on the Argonaute2 protein (*AGO2*) (Chi, Zang et al. 2009) from http://ago.rockefeller.edu/. For estimating the error profiles of regular RNA-Seq runs, we downloaded two sequencing lanes from an NGS quality assessment study with the accession numbers SRR896663 and SRR896664 (SEQC/MAQC-III-Consortium 2014) from DRASearch and pooled the data. An overview of the analyzed datasets can be found in Table 1.

142 **Table 1:** Overview of the analyzed RNA-Seq and CLIP datasets.

Dataset	Published (year)	Sequencing method	Platform	Accession number/website
EWSR1	2011	PAR-CLIP	Illumina Genome Analyzer II	SRA025082
FUS	2011	PAR-CLIP	Illumina Genome Analyzer II	SRA025082
TAF15	2011	PAR-CLIP	Illumina Genome Analyzer II	SRA025082
HuR	2011	PAR-CLIP	Illumina Genome Analyzer	SRR248532
MOV10	2012	PAR-CLIP	Illumina Genome Analyzer II	SRR490650
AGO2	2009	HITS-CLIP	Illumina Genome Analyzer II	http://ago.rockefeller.edu/
Human reference RNA	2014	RNA-Seq	Illumina HiSeq 2000	SRR896663, SRR896664

143

144 We used the following read aligners and versions, shown in alphabetic order: Bowtie, version
145 0.12.7 (Langmead, Trapnell et al. 2009), Bowtie2, version 2.2.3 (Langmead and Salzberg 2012),
146 BWA, version 0.7.8 (Li and Durbin 2009), BWA PSSM, initial release version (Kerpedjiev,
147 Frellsen et al. 2014), MOSAIK, version 2.2.3 (Lee, Stromberg et al. 2014), STAR, version 2.3.0
148 (Dobin, Davis et al. 2013), Subjunc, version 1.4.2 (Liao, Smyth et al. 2013) and TopHat, version
149 2.0.13 (Trapnell, Pachter et al. 2009).

150

2.2 PAR-CLIP read simulator and hierarchical clustering

We developed a PAR-CLIP read simulator (Figure 1) that creates short RNA reads which mimic important PAR-CLIP specific properties (Section 3.1). First, the following probability distributions are obtained from real PAR-CLIP data: (a) a probability matrix ε representing the background error profile of sequencing errors, (b) a probability vector of T–C conversion frequencies α for ranked T–C conversion sites, (c) a probability vector β for the preferred read positions of T–C conversion sites within binding sites, (d) a probability vector μ for indel frequencies per read position and (e) a probability vector δ for the base-calling quality score distribution per read position. The probability matrix ε contains a probability distribution for each DNA base over the DNA bases {A, C, G, T}. For this purpose, a PAR-CLIP dataset is aligned against a reference genome sequence with an appropriate read aligner.

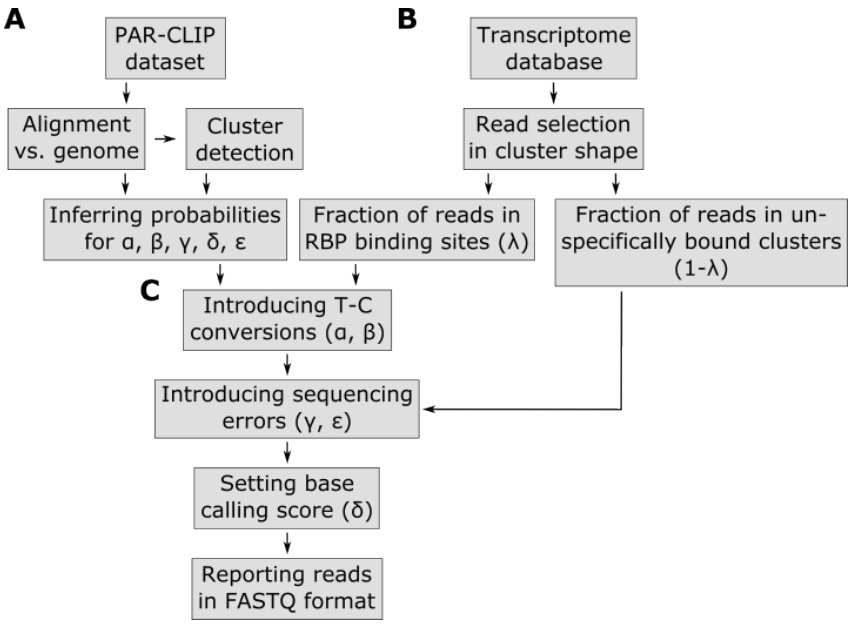


Figure 1: Pipeline of the PAR-CLIP read simulator implemented in the PARA-suite. Part A describes the process of generating the error profile and other parameters learned from a real PAR-CLIP dataset. Part B starts to generate reads mapping to RBP binding sites (clusters) on transcript regions from a given transcript database (e.g. Ensembl genes). In Part C, the pre-calculated profiles are used to introduce T–C conversions, sequencing errors, indels and base-calling quality scores to the defined reads.

Based on these alignments, the sequencing error profile ε is estimated from the observed frequencies of all single nucleotide substitutions, except for T–C errors, as these include PAR-CLIP specific T–C conversions. Standard T–C sequencing errors are approximated by the average over all the other sequencing error frequencies. The probability vectors μ and δ are also inferred from these alignments. Next, all aligned reads of the real dataset are clustered (stacked) using single-linkage hierarchical clustering based on their genomic mapping positions, using a 5-base overlap of the genomic mapping positions as the clustering threshold. To identify high confidence clusters (sometimes referred to as binding sites) as defined in the literature (Hafner, Landthaler et al. 2010), clusters that contain less than 10 reads, less than 25% T–C conversions per cluster, are longer than 75 bases and include only T–C conversion sites that are reported as single nucleotide polymorphism loci in the dbSNP database (version 142) (Sherry, Ward et al. 2001) are discarded. This implementation of hierarchical clustering is part of the PARA-suite

and will later be used for binding site detection. For the subsequent simulation, the positions and frequencies of highly mutated T–C sites within reads are determined to estimate α and β from the high confidence clusters (Figure S1A-B).

Next, the PAR-CLIP read simulation starts with the random selection of transcripts from a pre-selected database of annotated transcripts. One to at most three clusters (the number of clusters is randomly chosen from a uniform distribution) containing several reads are created for a selected transcript sequence. The starting positions of the clusters are randomly selected from a uniform distribution within the entire range of a transcript. The number of reads simulated for a single cluster is drawn from a normal distribution with a mean of 16 and a standard deviation of 10. This enables the simulation of a wide range of read coverages throughout the clusters. Furthermore, small shifts of the start and end site of each read leading to distinctive alignment position shifts in the shape of a cluster are randomly introduced at this step (normal distribution, standard deviation = 1). A user-defined parameter $\lambda \in [0,1]$ specifies the fraction of clusters that are considered to be binding-sites, whereas the remaining clusters mimic contaminations of unbound RNAs that occur in all PAR-CLIP experiments. We recommend values in the range of 0.5–0.7 (50–70%), as we observed this range of aligned sequencing reads stacking into clusters after hierarchical clustering and filtering (Table S1; similar values were previously reported by (Ascano, Hafner et al. 2012)). If more than one T–C site is simulated for a single cluster, a major T–C conversion site is selected according to the site-specific T–C conversion profile β and T–C conversion probabilities are drawn from α. Subsequently, background sequencing errors are introduced on the basis of the pre-computed probability matrix ε and the frequency vector μ for substitutions and indels, respectively. In the last step, every base receives a base-calling quality score, as specified by the position-specific quality score distribution δ. All generated reads are stored in the universal FASTQ format (Cock, Fields et al. 2010). The PAR-CLIP read simulator is available through the PARA-suite.

2.3 The PARA-suite: tools for error profile inference, read simulation, multiple database mapping and more

The PARA-suite is a toolkit for processing and aligning short and error-prone sequencing reads. It is implemented in Java using HTSjdk, a Java API for high-throughput sequencing data formats (https://github.com/samtools/htsjdk). The PARA-suite allows the user to estimate a sequencing run-specific error profile, combine the results of multiple reference database alignments, cluster an aligned sequencing read dataset (Section 2.2), run the PAR-CLIP read simulator, benchmark an alignment of simulated PAR-CLIP sequencing reads and run a full processing pipeline for error-prone short read alignments (Figure 2A). The alignment pipeline of the PARA-suite includes the calculation of an error profile for a particular sequencing run, applying the alignment algorithm described in the following section, and optionally combines the results of read mappings against multiple databases (Figure 2B–D). First, a read alignment against a reference sequence is performed with a fast short read aligner. By default, this is carried out with BWA, as our evaluations have demonstrated this to be a fast and accurate aligner (Section 3.3) on PAR-CLIP reads. However, other read aligners can also be used to produce the reference-based read alignment. This initial read alignment is used to estimate the underlying mismatch and indel probabilities M, I and D (as described in the next section) of the sequencing run. Once the error profile has been estimated, all sequencing reads can be aligned with BWA PARA (Section 2.4) against the reference sequence(s). All aligned reads are reported in a BAM file.

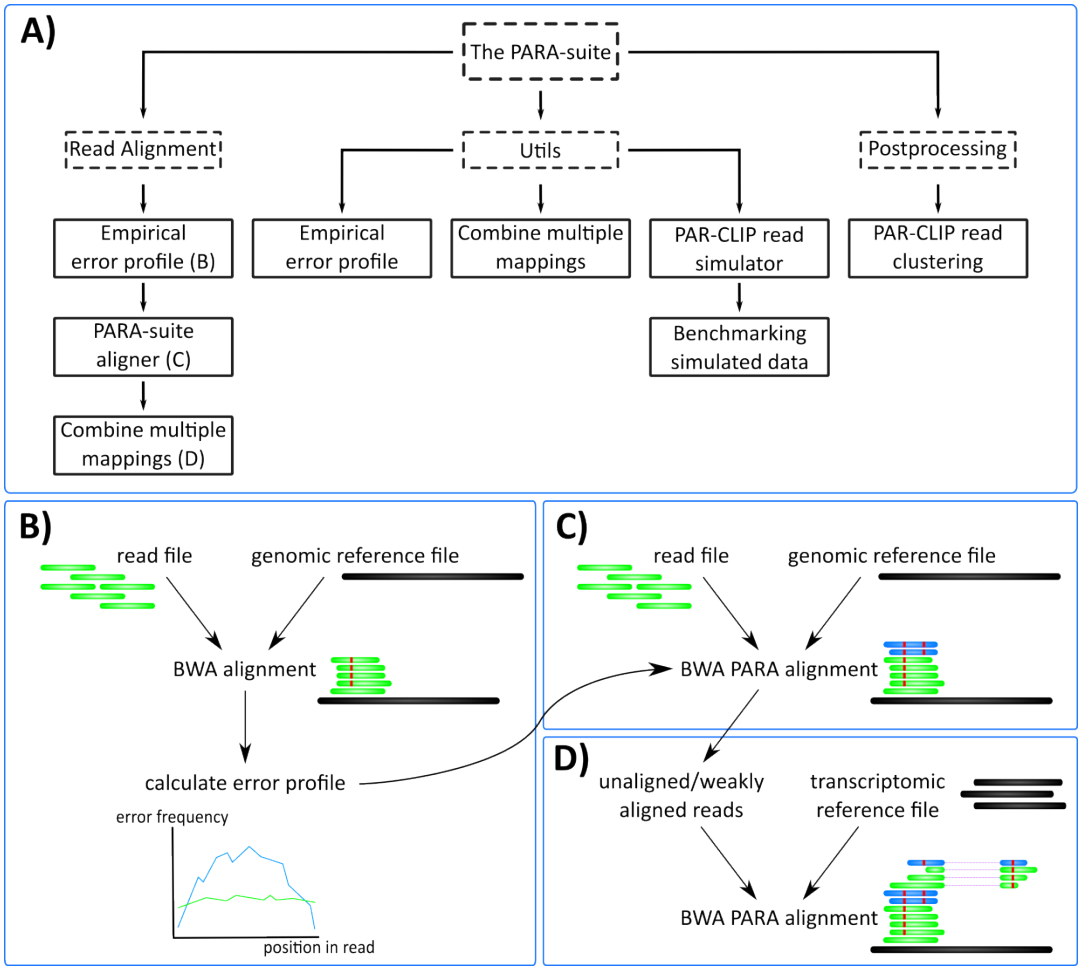


Figure 2: The PARA-suite. (A) The PARA-suite. Dashed boxes represent software packages; all other boxes represent executable programs. The Utils package includes tools for working with error-prone sequencing data and the postprocessing package contains a tool for clustering an aligned PAR-CLIP dataset to identify RBP-bound genomic regions. (B) Read alignment by a fast read aligner is necessary to infer the error profile for a particular read dataset (we selected BWA). (C) BWA PARA is applied to the entire dataset to map error-prone reads, indicated here by the additional mapping of the two reads (shown in blue). (D) An optional alignment versus a transcriptome reference database can be executed using BWA PARA to identify previously unmapped reads.

2.4 Algorithm of the PARA-suite aligner BWA PARA

The general BWA algorithm uses a Burrows–Wheeler transform (BWT) (Burrows and Wheeler 1994) to create an index for a reference genome sequence and applies a backward search to identify possible mapping positions in the genome for every single sequencing read. The backward search starts with the last base of a read proceeding to its front, searching the partly decompressed suffix trie using the auxiliary Ferragina and Manzini index (Ferragina and Manzini 2000) for a matching predecessor base of the read´s bases compared so far. Even if a match can be found for a single comparison, mismatches are introduced and all possible downstream paths within the suffix trie are considered until a pre-defined threshold of maximal mismatches is exceeded in a single path (Figure 3, red dotted line).

The principal idea of BWA PARA is the introduction of a probability estimate for each comparison of the backward search. This enables mismatches to be weighted according to their probabilities that they occur in the analyzed dataset. A sequencing run is initially characterized according to its underlying error probabilities. This allows us to determine specific error-profiles for experimental techniques, such as the frequent T–C conversions in PAR-CLIP data, which are more common than sequencing errors. The error profile M is a 4 × 4 probability matrix specifying substitution probabilities values ∈ [0..1] for each reference base ∈ {A, C, G, T} to the read bases {A, C, G, T} (Figure 4A). Indels are introduced during the alignment step separately, using the estimated probabilities $I \in [0,1]$ for insertions and $D \in [0,1]$ for deletions.

For each comparison between a read base at read position i (read[i]) and a reference base at position j (ref[j]) in the reference sequence, the algorithm recursively calculates a joint probability value p, which is used to examine the chance of incorporating a matching base or a suitable error, including indels, at the respective read positions (Figure 4D):

$$p_i = \begin{cases} p_{i+1} \cdot D, & if\ ref[j]\ is\ deleted \\ p_{i+1} \cdot I, & if\ read[i]\ is\ inserted \\ p_{i+1} \cdot M(read[i], ref[j]), & otherwise \end{cases}$$

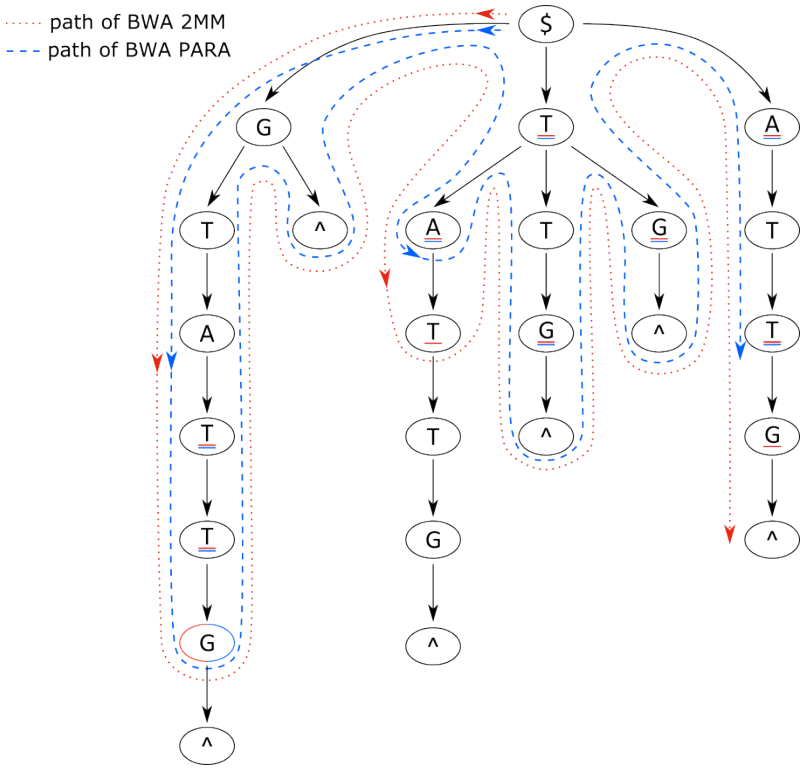
with $p_{|\text{read}|} = 1$, starting with $i = |\text{read}| - 1$ and decreasing i at each step, except in the case of a deletion (where i is left unchanged), for $i \geq 0$.

Before the alignment of a particular read, a minimal threshold T for the probability p is needed to decide whether a read is accepted as aligned or rejected. The calculation for T depends on a parameter X for the average number of mismatches. Note that this is not a maximal threshold in terms of absolute mismatches, as the number of the more frequent errors per aligned read can exceed X. The parameter X can be pre-defined by the user or is by default estimated as the expected number of mismatches for different read lengths based on the error profile M for a sequencing run. Next, the minimal threshold T is computed (Figure 4B&C):

$$T = avg(\text{match})^{|read| - X} \cdot avg(\text{mismatch})^{X},$$

where $avg(\text{match}) = \frac{1}{5}\left[\sum_{i \in \{0..3\}} M_{i,i} + (1 - (I + D))\right]$ and $avg(\text{mismatch}) = \frac{1}{14}$ $\left[\sum_{i,j \in \{0..3\}; i \neq j} M_{i,j} + I + D\right]$.

Both *avg(*match*)* and *avg(*mismatch*)* are normalized by the number of elements (four matches plus one for no indel occurring, and 12 mismatches plus 2 for either a insertion or a deletion). If p falls below the pre-calculated threshold T during read alignment, the path within the suffix trie is assumed not to match the read and is rejected (Figure 3, blue dashed line). The algorithm thus penalizes rare types of mismatches according to M, whereas frequent errors, such as T–C errors in PAR-CLIP reads, are the most favored substitutions in the alignment process (Figure 4B–D).



296

297 **Figure 3: Suffix trie paths for the BWA and PARA-suite.** Paths of the algorithms through the
298 suffix trie aligning the read sequence GCCATG$ against the reference sequence GTTATG$
299 (where $ means the end of a sequence). The red dotted line represents the algorithm of the BWA
300 aligner, allowing for two mismatches; the blue dashed line indicates the BWA PARA algorithm.
301 The underlined bases represent positions where the respective aligner introduces a mismatch.
302 The example shows that BWA PARA needs 14 comparisons but the basic BWA needs 16
303 comparisons. Indels are not shown for simplicity.

304

A)

error profile *M* (in %):

ref \ read	A	C	G	T
A	99.0	0.3	0.3	0.4
C	0.5	98.9	0.4	0.2
G	0.4	0.2	99.1	0.3
T	0.3	6.3	0.4	93.0

I = 0.254 %
D = 0.224 %

avg(match) = 97.904 %
avg(mm) = 0.748 %
X = 2

B)

```
read:   ... CACGGCCACG    (|read| = 26)
ref a): .. TACGGTTACG     (3 T-C)
ref b): .. CACAGCCA-G     (2 sequencing errors)
ref c): .. CACGCCCATG     (1 seq. error, 1 T-C)
```

C)

$p \geq 0.97904^{24} * 0.00748^{2} = 0.0000336522 = 3.36522 * 10^{-5} = T$

D)

a) $p_{26} = 1$; $p_{25} = 1 * 0.991$; $p_{24} = 0.991 * 0.989 = 0.980099$; ...;
$p_{22} = 0.06113$; $p_{21} = 0.003851$; ...; $p_{0} \approx 0.00020005 > T$ ✓

b) $p_{26} = 1$; $p_{25} = 0.991$; $p_{24} = 0.991 * 0.00254 = 0.002517$; ...;
$p_{19} = 0.000007247 < T$ ✗

c) $p_{26} = 1$; $p_{25} = 0.991$; $p_{24} = 0.991 * 0.063 = 0.062433$; ...;
$p_{0} = 0.0001534 > T$ ✓

Figure 4: The BWA PARA alignment approach. (A) The error profile probability matrix *M* and the indel probabilities *I* and *D*, which are used as input for the BWA PARA algorithm, as well as exemplary results of the intermediate calculations of the BWA PARA algorithm. In *M*, only T–C conversions have a higher probability (6.3%) than sequencing errors and indels. (B) The last characters of a particular read and three examples of mapping positions within a reference, called ref a–c. (C) The calculation of a maximum threshold *T* for the mapping probability *p* (see the Equation 2 in the main text, and values from (A) in this image). (D) The mapping probability calculation of the read when mapped to References a–c. The read fails to map against ref b with two sequencing errors, whereas ref a and ref c are suitable mapping positions, where the probability *p* is higher than the threshold *T*. For implementation, we worked with the open-source read aligner BWA (version 0.7.8) to extend its algorithm for the alignment of short and error-prone reads.

Results

3.1 Properties of PAR-CLIP reads

To assess the most important properties of the PAR-CLIP sequencing reads for read alignment, we systematically compared PAR-CLIP datasets for the three RBPs *EWSR1*, *FUS* and *TAF15* (the *FET* protein family) (Hoell, Larsson et al. 2011) to a recently published RNA-Seq run on human reference RNA (SEQC/MAQC-III-Consortium 2014). The 10 outermost bases of the SEQC/MAQC reads showed error rates with peaks at 1.5 and 2.2 errors per 100 reads (EPR). In contrast, the middle read length range showed an average of about 0.3 EPR (Figure S2A, red line). As the short reads of the *FET* PAR-CLIP datasets consisted only of these outermost bases, they exhibited a two- to threefold higher average sequencing error rate (about 0.7 EPR or even higher) than the SEQC/MAQC reads (Figure S2B, green line). When considering the T–C conversions only, we observed 1.319 EPR for *EWSR1*, 1.477 EPR for *FUS* and 1.051 EPR for *TAF15* on average. This is an approximately 20- to 30-fold increase in comparison to the SEQC/MAQC dataset with 0.051 EPR for T–C conversions on average (Figure S2). Moreover, we analyzed data from two further PAR-CLIP studies performed on the RBPs *HuR* (Mukherjee, Corcoran et al. 2011) and *MOV10* (Sievers, Schlumpf et al. 2012), which showed similar error profiles and EPRs to the *FET* PAR-CLIPs for T–C conversions (Figure S3).

Further analyses of the PAR-CLIP read datasets for *EWSR1*, *FUS*, *TAF15*, *MOV10* and *HuR* showed the PAR-CLIP reads (a) to be shorter than 30 bases, (b) to cover only short stretches of an expressed gene rather than the entire expressed RNA (these stretches are henceforth called clusters), (c) to exhibit a specific nucleotide conversion pattern with a strong enrichment of T–C conversions, where (d) such conversions occur in specific ‘conversion sites’ in the clusters. The two properties (a) and (b) are determined by treating the cells with RNAse T1 or the lysate during the PAR-CLIP experimental protocol. As only short RNA fragments that are not digested by the endonuclease (these are probably protected by the binding pocket of the RBP) are sequenced, the lengths of those fragments are usually short. However, the nucleotide composition of those reads is strongly affected by the digestion enzyme and can vary among different digestion enzymes (Kishore, Jaskiewicz et al. 2011). After quality trimming and adapter trimming of the five PAR-CLIP datasets, the average read lengths were 25.67 bases (*EWSR1*), 25.60 bases (*FUS*), 24.21 bases (*TAF15*), 25.20 bases (*HuR*) and 23.36 bases (*MOV10*). As the

transcript regions outside the bound RNA fragment are digested by the endonuclease, these are removed during immunoprecipitation and not sequenced, except for additional binding sites on the same transcript further up- or downstream. Thus the sequencing reads are stacked into short clusters covering short stretches of the gene and representing the RBP-bound regions of the transcripts (Figure S4A).

The two properties (c) and (d) were determined by incorporating photoactivatable nucleosides into the nascent transcripts during transcription. In the case of 4-SU, T–C conversions occur in the sequencing reads at all cross-linked sites, where the 4-SU is incorporated instead of the native uridine. These conversions can reach high rates in specific conversion sites within a cluster (Hafner, Landthaler et al. 2010). In the analyzed datasets, we observed an average frequency of about 70% T–C conversions in the main T–C conversion site (Figure S1A). This emphasizes that simulated read datasets with specific properties are necessary for the evaluation of common short read aligners for analyzing PAR-CLIP read data. However, this cannot be created by common sequencing read simulators, such as ART (Huang, Li et al. 2012) or GemSIM (McElroy, Luciani et al. 2012). These produce simulated reads with a continuous coverage over the entire transcript range and the introduced mutations are distributed randomly throughout the simulated reads. This is not the case for PAR-CLIP sequencing reads.

3.2 PAR-CLIP read simulation for performance evaluation

We simulated a total of 10 PAR-CLIP read datasets based on information learned from three previously published PAR-CLIP datasets of the *FET* protein family (Hoell, Larsson et al. 2011) (Table S2). We imitated Illumina GenomeAnalyzer II sequence data according to the real datasets used. The respective sequencing error and T–C conversion profiles were generated on the basis of alignments of all three datasets against the human reference genome sequence version 38 (GRCh38) (Lander, Linton et al. 2001). The error profile and additionally estimated distributions were similar to the ones from PAR-CLIP data on the two RBPs *HuR* and *MOV10*, indicating that these profiles represented a reasonable approximation for PAR-CLIP data in general. We selected human transcript sequences downloaded from Ensembl Genes version 77 (Cunningham, Amode et al. 2015) as our sequence database to simulate human transcript read sequences. We set λ, the parameter for the fraction of sequencing reads that stacked into clusters

bound by the RBP, to 65%. These true RBP binding sites showed high T–C conversion frequencies in different T–C conversion sites. The remaining 35% of the simulated sequencing reads were designated to represent non-specifically bound transcripts without an elevated T–C conversion rate, except for a few T–C sequencing errors. These reflected RNA contaminations that can occur during the PAR-CLIP experiment.

To assess the quality of the simulation, we then compared PAR-CLIP-specific properties between the 10 simulated datasets and the *FET* PAR-CLIP data. Within a cluster detected in a simulated dataset, shifts in the alignment positions of a few nucleotides at the beginning and the end of the simulated cluster could be seen between the reads (Figure S4B). According to the position-wise T–C conversion profile used, a T–C conversion site with a high conversion rate, as well as a few sites with lower conversion rates, were usually present in the detected clusters (e.g. Figure 1B). We compared the error profiles between one of the simulated datasets and the real datasets, and distinguished between T–C errors and all other errors; the latter represent all sequencing errors other than the T–C sequencing errors (Figure S2C). Similar to the real data, the distribution of the sequencing errors in the simulated dataset peaked at the beginning of the reads and dropped to a mean error rate of 0.6 EPR in the middle read length range. Error rates were slightly underestimated in the simulated data compared to the real PAR-CLIP data, presumably because of a small percentage of multiple mutations that occurred at individual sites. Apart from this, the simulated datasets appeared to be representative of real PAR-CLIP data in the relevant aspects.

3.3 Accuracy of common read aligners and the PARA-suite on simulated PAR-CLIP data

Using the simulated PAR-CLIP datasets, we analyzed the accuracy of state-of-the-art read aligners and common binding site detection algorithms, and compared these to the PARA-suite alignment pipeline. The aligners BWA and Bowtie have often been used in CLIP studies (Lebedeva, Jens et al. 2011, Ascano, Mukherjee et al. 2012, Sievers, Schlumpf et al. 2012). BWA PSSM was applied with the PSSM for PAR-CLIP provided by its authors because a PSSM estimated from the sequencing dataset revealed worse accuracy (data not shown). MOSAIK was executed, reporting only unique mappings, allowing for up to three mismatches between the read

and the reference sequence, and using a Smith–Waterman bandwidth of 5. The read aligners were used to align the simulated datasets to the reference sequence *GRCh38*. We also executed the PARA-suite on the Ensembl Genes transcriptome database (version 77) and combined the results with the genomic reference sequence alignments. These results are henceforth referred to as those of the "PARA-suite pipeline", whereas the results of the genomic alignment step using the PARA-suite only are referred to as those of "BWA PARA". For BWA PARA, the sequencing error and T–C conversion profiles for the simulated datasets were obtained on the basis of the BWA alignments, allowing for two mismatches (BWA 2MMs) for each of the simulated datasets separately (execution commands are outlined in the Supplementary Methods). For an overview of the performance, we estimated the average of the recall, precision and accuracy for each aligner over the 10 simulated datasets (our calculations are described in the Supplementary Methods). Unfortunately, BMix does not report negative clusters (contaminations) and thus we were able to neither calculate the recall nor the accuracy, but only the precision.

In terms of overall performance, the PARA-suite performed best, with an accuracy of 69.74% for BWA PARA and 73.14% for the entire pipeline, showing performance gains of 1.57% and 4.97% compared to the second-best aligner (BWA 2MM), respectively (Table 2, Table S3). Many prominent PAR-CLIP studies have used Bowtie 1MM or BWA 2MM for the read alignment step (Lebedeva, Jens et al. 2011, Mukherjee, Corcoran et al. 2011, Ascano, Mukherjee et al. 2012, Sievers, Schlumpf et al. 2012, Mukherjee, Jacobs et al. 2014). When we compared the PARA-suite pipeline with these two aligners, the PARA-suite pipeline showed an increase of 16.95% and 4.97% in the overall accuracy, respectively. Notably, 1.56% of the reads aligned by the PARA-suite pipeline on average spanned an exon–exon junction. These were not identified by the genomic reference mapping step but instead required alignment against the transcriptome reference sequences. Additionally, we compared the recall (the fraction of correctly aligned reads out of all simulated reads) and the precision (the fraction of correctly aligned reads out of all aligned reads) to assess the mapping ability of the read aligners (Table 2, Figure S5). Here, the PARA-suite pipeline and BWA PARA were ranked first and third regarding recall, and first and second regarding precision, respectively, out of 10 analyzed alignment scenarios (Table 2). Hence, the PARA-suite pipeline and BWA PARA offer notable performance increases over commonly applied alignment setups.

443 We then tested the accuracy of the binding site detection algorithms BMix, PARalyzer and the
444 hierarchical clustering of the PARA-suite using the read alignments of BWA PARA (Table S4).
445 The hierarchical clustering identified the most correct binding sites: 3.26% more correct sites
446 than BMix and 5.54% more correct binding sites than PARalyzer. However, BMix identified
447 fewer false binding sites than the hierarchical clustering (20.30% fewer) and PARalyzer (69.85%
448 fewer). Furthermore, we investigated whether BWA PARA increased the number of binding
449 sites detected, irrespective of the detection algorithm used. In conjunction with BMix, BWA
450 2MM (the second-best aligner) identified 7.17% fewer correct binding sites than BWA PARA.
451 With PARalyzer, BWA 2MM identified 2.97% fewer correct binding sites than BWA PARA.
452 Finally, the hierarchical clustering identified 7.52% more correct binding sites for BWA PARA
453 than for BWA 2MM. Overall, the combination of BMix and BWA PARA provided the most
454 accurate results on our simulated data.
455

Table 2: Alignment accuracy on simulated PAR-CLIP data. The most accurate alignment results were obtained for different parameter settings for each read aligner on 10 simulated PAR-CLIP datasets. The results are averaged per read aligner over all 10 datasets and are sorted by accuracy.

Aligner	Accuracy (in %)	Variance	Recall (in %)	Precision (in %)	Mapped overall	Mapped correctly	Real time (s)
PARA-suite pipeline	73.14	1.37E-06	84.49	71.85	1,024,792	969,948	396.8
BWA PARA	69.74	1.38E-06	82.16	68.24	975,672	924,802	153.7
BWA 2MMs	68.17	1.37E-06	82.31	64.98	959,171	904,034	359.2
Bowtie 2MMs	63.38	1.10E-06	77.91	60.93	886,512	840,540	120.6
BWA PSSM	59.80	1.18E-06	74.04	58.72	818,895	793,007	25.4
TopHat	59.69	8.35E-07	76.10	55.35	844,902	791,549	282.9
Bowtie2	56.22	1.11E-06	73.23	51.43	763,893	745,531	13.4
STAR	50.74	9.10E-07	69.57	43.02	826,871	672,920	248.6
MOSAIK	44.88	2.18E-04	62.83	37.16	897,679	595,220	12,128.18
Subjunc	35.42	9.03E-07	50.61	26.09	597,400	469,751	64.2

3.4 Analysis of *FET* PAR-CLIP datasets

To investigate the performance of the PARA-suite on real PAR-CLIP datasets, we applied it to the three *FET* PAR-CLIP datasets (Hoell, Larsson et al. 2011). The sequencing reads were preprocessed similarly to the method given in the original publication, and low quality ends and adapter sequences were trimmed using Cutadapt (Martin 2011). Afterwards, all remaining reads longer than 18 bases were aligned against *GRCh38* with Bowtie2, Bowtie 2MM, BWA 2MMs, BWA PSSM and BWA PARA (without the transcriptome mapping step to achieve comparable results). Selection of the read aligners (i.e. Bowtie2, Bowtie 2MM, BWA 2MM, BWA PSSM and BWA PARA) was based on the results of the previous section, as these represent the most accurate read aligners on PAR-CLIP data. We measured the fraction of aligned reads for all the

aligners on the three datasets (Table S5). BWA PARA generated the largest fraction of aligned reads over all three datasets in comparison to BWA 2MM and BWA PSSM. Next, we stacked (clustered) all the aligned reads using BMix and the hierarchical clustering tool of the PARA-suite (Table 3). BWA 2MM identified fewer binding sites than BWA PSSM or BWA PARA for read alignments prior to either BMix or hierarchical clustering. Using the hierarchical clustering, BWA PARA reported the largest number of binding sites for two out of the three datasets. BWA PSSM identified 6.90% more clusters than BWA PARA for the *FUS* dataset whereas BWA PARA identified 3.98% more clusters for the *EWSR1* dataset and 19.21% more clusters for the *TAF15* dataset than BWA PSSM. In comparison to the values reported in the original publication, the use of BWA PARA and hierarchical clustering increased the number of binding sites by 33.71% for *EWSR1* and 16.77% for *FUS*, and decreased them by 12.56% for *TAF15*. After extracting distinct genes from all binding sites identified by BWA PARA (10,631 genes in total), 26.90% additional genes were found for all three RBPs, in comparison to the original publication (7,771 genes in total). As expected for three RBPs from the same family, there was a substantial overlap in terms of the identified genes, with 2,702 genes targeted by all three RBPs (Figure S6).

Table 3: Binding sites detected for the *FET* protein family. The number of binding sites for the *FET* protein family identified by the aligners BWA 2MM, BWA PSSM, BWA PARA, Bowtie 2MM and Bowite2 in combination with BMix and the hierarchical clustering of the PARA-suite. Filters were applied according to Section 2.2.

	EWSR1	***FUS***	***TAF15***
BWA 2MM BMix	20,703	14,768	5,086
BWA 2MM Clustering	22,760	36,861	5,810
BWA PSSM BMix	24,639	19,628	5,238
BWA PSSM Clustering	27,550	51,606	6,130
BWA PARA BMix	25,478	19,006	5,862
BWA PARA Clustering	28,692	48,042	7,588
Bowtie 2MM BMix	19,173	13,902	4,582
Bowtie 2MM Clustering	21,082	35,490	5,254
Bowtie2 BMix	12,384	8,078	3,558

Bowtie2 Clustering	13,338	20,398	3,710

493

494 ## 3.5 Analysis of PAR-CLIP data on *HuR*

495

496 We next applied the PARA-suite to a PAR-CLIP dataset on *HuR*, an RBP promoting RNA
497 stabilization (Mukherjee, Corcoran et al. 2011). Adapters and low-quality ends within the *HuR*
498 dataset were trimmed using Cutadapt and reads shorter than 14 bases were discarded. The
499 binding motif of *HuR* is well-studied and is AU-rich, with a consensus motif described as
500 AUUUA, AUUUUA or AUUUUUA (Nabors, Suswam et al. 2003, Lebedeva, Jens et al. 2011),
501 showing potentially more T–C conversions within each binding site than other RBPs. As the
502 generated error profile of the dataset was similar to those of the *FET* PAR-CLIP data (Section
503 3.1), the data quality seemed comparable. However, we noted a slight increase in T–C
504 conversions (Figure S3). The AU-rich binding motif might explain the higher T–C conversion
505 rate of 1.684 EPR compared to the conversion rate of 1.477 EPR e.g. for *FUS*.

506

507 We used the same read aligners as described in the previous section (Bowtie2, Bowtie 2MM,
508 BWA 2MM, BWA PSSM and BWA PARA) to align the pre-processed dataset against the
509 human genome reference *GRCh38*. We applied BMix and the hierarchical clustering of the
510 PARA-suite to determine the binding sites of *HuR* derived by using the different read aligners.
511 BWA PSSM, in conjunction with BMix, identified the most RBP binding sites within the
512 genome, which was 3.69% more than BWA PARA (Table 4). When we compared the binding
513 sites detected by BMix and the PARA-suite hierarchical clustering for alignments created by
514 BWA PARA (binding site positions overlapping by at least 13 bases), the difference was only
515 marginal, with an overlap of more than 98.25% for the two methods. A recent study of this
516 dataset reported binding sites using Bowtie 2MM for the alignment step and PARalyzer for the
517 binding site detection. We found that the use of either BWA PSSM or BWA PARA in
518 conjunction with either BMix or hierarchical clustering increased the number of binding sites
519 detected by 2.87–7.84%.

520 **Table 4:** Binding sites detected by BMix and the hierarchical clustering based on read
521 alignments performed by BWA 2MM, BWA PSSM, BWA PARA, Bowtie 2MM and Bowtie2
522 on the *HuR* dataset.

	BMix	Hierarchical clustering
BWA 2MM	136,775	137,697
BWA PSSM	147,883	148,985
BWA PARA	141,365	141,867
Bowtie 2MM	125,592	125,067
Bowtie2	88,369	87,400

We searched for the exact binding motifs of *HuR* (ATTTA, ATTTTA and ATTTTTA) within the binding sites detected by BMix within 3' untranslated region (UTR) or introns for all the read aligners tested. We found that all aligners performed comparably, with motifs present in 42–44% of all binding sites detected. The largest fraction was achieved using read alignments with BWA PSSM (44.33%), whereas BWA PARA in combination with BMix found 42.53% of the binding sites that were most likely correct. Bowtie 2MM in combination with BMix had the lowest fraction of binding sites containing the reported binding motif (42.44%). We also compared the previously reported *HuR* binding sites to the binding sites determined by the full PARA-suite pipeline with BMix for clustering and detected 13 out of 15 sites, namely 3' UTR *PTGS2*, 3'UTR *CDKN1A*, 3'UTR *VEGFA*, 3'UTR *TNF*, 3' UTR *SLC7A1*, 3'UTR *CCND1*, 3'UTR *MYC*, 3' UTR *XIAP*, 3'UTR *CELF1*, TTS *CSF2*, 3'UTR *CCNB1*, intron *NCL* and 3' UTR *KRAS*. The binding information for this comparison was taken from the Ingenuity knowledge base (Calvano, Xiao et al. 2005). The original study on the *HuR* dataset (Mukherjee, Corcoran et al. 2011) only reported 12 out of these 15 genes having confirmed binding site.

Discussion

We provided a detailed characterization of the error profiles of PAR-CLIP reads and an in-depth performance assessment of short read aligners in combination with binding site detection tools. We characterized some of the unique properties of PAR-CLIP sequence datasets, including the preferred read positions for T–C conversion sites and their frequencies per read position. We observed higher frequencies of sequencing errors in PAR-CLIP data than in the human reference RNA-Seq data. A likely reason for this behavior could be that PAR-CLIP reads are much shorter than common RNA-Seq reads, which reach lengths of 200 bases and show high-quality regions

in the middle read length range (Laehnemann, Borkhardt et al. 2015, Schirmer, Ijaz et al. 2015). We used these observations for the design of a PAR-CLIP read simulator that embeds PAR-CLIP specific information within the simulation process and the PARA-suite pipeline for error-aware read alignment and processing. The read simulator mimics PAR-CLIP datasets with error profiles drawn from real PAR-CLIP datasets.

Based on the simulated PAR-CLIP datasets, we determined the parameter settings that delivered the best performance for commonly used aligners (Mukherjee, Corcoran et al. 2011, Ascano, Mukherjee et al. 2012, Sievers, Schlumpf et al. 2012, Mukherjee, Jacobs et al. 2014). Our analysis showed that the read alignment is crucial for detecting RBP binding sites in PAR-CLIP datasets. The PAR-CLIP specific read properties make it nearly impossible to identify splice junctions covered by PAR-CLIP reads with RNA-Seq read aligners such as TopHat, STAR or Subjunc, as their algorithms are based on unmet assumptions, such as a similar read coverage across all exons or long reads, in order to achieve high confidence k-mer spectra. Accordingly, these three aligners were outperformed by the other methods (Table S3–4). Interestingly, MOSAIK, an error-aware aligner based on hash queries that has been shown to be more robust on RNA-Seq reads than BWT-based aligners (Lee, Stromberg et al. 2014), was also outperformed by most of the other tested methods. Although it was robust on longer RNA-Seq reads, MOSAIK seemed to struggle with the very short PAR-CLIP reads. The PARA-suite alignment pipeline allowed us to increase the fraction of aligned reads in comparison to other aligners, including the alignment of reads spanning exon–exon junctions, both for PAR-CLIP datasets and data from a HITS-CLIP study (Supplementary Results). We observed this improvement irrespective of the binding site detection algorithm applied downstream. Importantly, unlike the error-aware short read aligner BWA PSSM, our short read alignment algorithm does not need the manual input of an error profile, which is instead inferred *de novo* within individual sequencing runs. The aligner thus automatically adapts to varying qualities of individual (PAR-)CLIP sequencing runs and is specifically adjusted to each sequence dataset. To our knowledge, it is the first tool for simultaneous *de novo* error model inference and short read alignment based on the BWA algorithm. Another difference from the BWA PSSM algorithm is that the latter introduces mismatches while considering the base calling quality scores and a probabilistic background model for matching bases in addition to the input error profile. In

578 contrast, the generic error profile estimation of the PARA-suite is not limited to any specific
579 input profile. Further applications of our software could thus be used to analyze other types of
580 error-prone sequencing data such as bisulphite sequencing data, which introduces a high amount
581 of C–T mutations (Frommer, McDonald et al. 1992) or data from low-quality ancient DNA
582 samples (Briggs, Stenzel et al. 2007).
583
584 Common read simulators such as ART or GemSim do not allow simulating PAR-CLIP reads
585 with their specific error profiles. When comparing the PAR-CLIP read simulator to the recently
586 developed CSeq simulator for CLIP data (Kassuhn, Ohler et al. 2016), both have different
587 strengths. CSeq takes an exact binding motif and T–C conversion profile that is specific for the
588 respective binding motif as input, thus restricting the reads' base composition and T–C
589 conversion sites. This allows to mimic PAR-CLIP reads for a specific RBP, but not to generalize
590 evaluations on these datasets to all kinds of RBPs. In comparison, the PAR-CLIP reads simulated
591 with the PARA-suite are based on data that have been inferred from three different PAR-CLIP
592 datasets to simulate heterogenic reads, which represent a broader spectrum of RBP binding sites.
593 In addition, the read selection is not restricted to sequences containing the actual RBP binding
594 motif. Thus CSeq and the PARA-suites' read simulator have slightly different applications: CSeq
595 allows one to simulate reads to optimize parameters for a specific dataset and the PARA-suite
596 allows one to simulate reads for general tool evaluation and algorithmic improvements.
597
598 Our analysis of combinations of read aligners and binding site detection algorithms on simulated
599 and real datasets indicated that no single software performed best in terms of detecting binding
600 sites on the available PAR-CLIP datasets. This observation was recently also made on other
601 datasets (Kassuhn, Ohler et al. 2016). Our analysis of the *HuR* and *FUS* datasets revealed that U-
602 rich binding sites tended to show higher rates of T–C conversions per read and were best aligned
603 by BWA PSSM. RBPs with a more heterogeneous nucleotide distribution within the binding site
604 (e.g. *EWSR1* and *TAF15*) are better assessed by BWA PARA. This is supported by an analysis of
605 uridylate-rich sequences from our simulated data aligned by BWA PSSM and BWA PARA
606 (Supplementary Results and Supplementary Table S6). Therefore, a preliminary analysis of the
607 error profile using the PARA-suite error profiler could allow one to determine the best approach
608 to analyze sequencing data of a novel yet uncharacterized RBP.

Acknowledgements

The authors thank Johannes Droege, David Laehnemann and Cristina della Beffa for their critical comments on the manuscript.

References

Ascano, M., Hafner, M., Cekan, P., Gerstberger, S. and Tuschl, T. (2012). Identification of RNA–protein interaction networks using PAR-CLIP. *Wiley Interdiscip. Rev. RNA* **3**(2): 159-177.

Ascano, M., Mukherjee, N., Bandaru, P., Miller, J. B., Nusbaum, J. D., Corcoran, D. L., Langlois, C., Munschauer, M., Dewell, S. and Hafner, M. (2012). FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature* **492**(7429): 382-386.

Briggs, A. W., Stenzel, U., Johnson, P. L., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T. and Lachmann, M. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U S A* **104**(37): 14616-14621.

Burrows, M. and Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm. *Digital Equipment Corporation: Palo Alto, CA* (Technical Report 124).

Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P. and Tschoeke, S. K. (2005). A network-based analysis of systemic inflammation in humans. *Nature* **437**(7061): 1032-1037.

Chi, S. W., Zang, J. B., Mele, A. and Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature* **460**(7254): 479-486.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. and Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**(6): 1767-1771.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Kahari, A. K., Keenan, S., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A.,

638 Wilder, S. P., Zadissa, A., Aken, B. L., Birney, E., Harrow, J., Kinsella, R., Muffato, M.,
639 Ruffier, M., Searle, S. M., Spudich, G., Trevanion, S. J., Yates, A., Zerbino, D. R. and Flicek,
640 P. (2015). Ensembl 2015. *Nucleic Acids Res.* **43**(Database issue): D662-669.
641 Danan, C., Manickavel, S. and Hafner, M. (2016). PAR-CLIP: A Method for Transcriptome-Wide
642 Identification of RNA Binding Protein Interaction Sites. *Post-Transcriptional Gene*
643 *Regulation*: 153-173.
644 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and
645 Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1):
646 15-21.
647 Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature reviews. Genetics*
648 **2**(12): 919-929.
649 Ferragina, P. and Manzini, G. (2000). Opportunistic data structures with applications.
650 *Proceedings of the 41st Symposium on Foundations of Computer Science*: 390-398.
651 Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L. and
652 Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-
653 methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U S A* **89**(5):
654 1827-1831.
655 Glisovic, T., Bachorik, J. L., Yong, J. and Dreyfuss, G. (2008). RNA-binding proteins and post-
656 transcriptional gene regulation. *FEBS Lett.* **582**(14): 1977-1986.
657 Golumbeanu, M., Mohammadi, P. and Beerenwinkel, N. (2015). BMix: probabilistic modeling of
658 occurring substitutions in PAR-CLIP data. *Bioinformatics*: btv520.
659 Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A.,
660 Ascano, M., Jr., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S.,
661 Zavolan, M. and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding
662 protein and microRNA target sites by PAR-CLIP. *Cell* **141**(1): 129-141.
663 Hieronymus, H. and Silver, P. A. (2004). A systems view of mRNP biology. *Genes Dev.* **18**(23):
664 2845-2860.
665 Hoell, J. I., Hafner, M., Landthaler, M., Ascano, M., Farazi, T. A., Wardle, G., Nusbaum, J., Cekan,
666 P., Khorshid, M. and Burger, L. (2014). Transcriptome-Wide Identification of Protein

667 Binding Sites on RNA by PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced
668 Crosslinking and Immunoprecipitation). *Handbook of RNA Biochemistry: Second,*
669 *Completely Revised and Enlarged Edition*. A. B. R.K. Hartmann, A. Schön, and E. Westhof.
670 Weinheim, Wiley-VCH Verlag GmbH & Co. KGaA. **II:** 877-898.
671 Hoell, J. I., Larsson, E., Runge, S., Nusbaum, J. D., Duggimpudi, S., Farazi, T. A., Hafner, M.,
672 Borkhardt, A., Sander, C. and Tuschl, T. (2011). RNA targets of wild-type and mutant FET
673 family proteins. *Nat. Struct. Mol. Biol.* **18**(12): 1428-1431.
674 Huang, W., Li, L., Myers, J. R. and Marth, G. T. (2012). ART: a next-generation sequencing read
675 simulator. *Bioinformatics* **28**(4): 593-594.
676 Kassuhn, W., Ohler, U. and Drewe, P. (2016). *Cseq-simulator: a data simulator for CLIP-Seq*
677 *experiments*. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing.
678 Kerpedjiev, P., Frellsen, J., Lindgreen, S. and Krogh, A. (2014). Adaptable probabilistic mapping
679 of short reads using position specific scoring matrices. *BMC Bioinformatics* **15**(1): 100.
680 Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M. and Zavolan, M. (2011). A
681 quantitative analysis of CLIP methods for identifying binding sites of RNA-binding
682 proteins. *Nat. Methods* **8**(7): 559-564.
683 Kloetgen, A., Münch, P. C., Borkhardt, A., Hoell, J. I. and McHardy, A. C. (2015). Biochemical and
684 bioinformatic methods for elucidating the role of RNA-protein interactions in
685 posttranscriptional regulation. *Brief. Funct. Genomics* **14**(2): 102-114.
686 König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M. and
687 Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual
688 nucleotide resolution. *Nat. Struct. Mol. Biol.* **17**(7): 909-915.
689 Laehnemann, D., Borkhardt, A. and McHardy, A. C. (2015). Denoising DNA deep sequencing
690 data—high-throughput sequencing errors and their correction. *Brief. Bioinform.*:
691 bbv029.
692 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar,
693 K., Doyle, M. and FitzHugh, W. (2001). Initial sequencing and analysis of the human
694 genome. *Nature* **409**(6822): 860-921.

695 Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat.*
696 *Methods* **9**(4): 357-359.
697 Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009). Ultrafast and memory-efficient
698 alignment of short DNA sequences to the human genome. *Genome Biol.* **10**(3): R25.
699 Lebedeva, S., Jens, M., Theil, K., Schwanhäusser, B., Selbach, M., Landthaler, M. and Rajewsky,
700 N. (2011). Transcriptome-wide analysis of regulatory interactions of the RNA-binding
701 protein HuR. *Mol. Cell* **43**(3): 340-352.
702 Lee, W.-P., Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P. and Marth, G. T. (2014).
703 MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read
704 mapping. *PLoS One* **9**(3): e90581.
705 Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler
706 transform. *Bioinformatics* **25**(14): 1754-1760.
707 Liao, Y., Smyth, G. K. and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read
708 mapping by seed-and-vote. *Nucleic Acids Res.* **41**(10): e108.
709 Lukong, K. E., Chang, K.-w., Khandjian, E. W. and Richard, S. (2008). RNA-binding proteins in
710 human genetic disease. *Trends Genet.* **24**(8): 416-425.
711 Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing
712 reads. *EMBnet. J.* **17**(1): 10-12.
713 McElroy, K. E., Luciani, F. and Thomas, T. (2012). GemSIM: general, error-model based simulator
714 of next-generation sequencing data. *BMC Genomics* **13**(1): 74.
715 Mukherjee, N., Corcoran, D. L., Nusbaum, J. D., Reid, D. W., Georgiev, S., Hafner, M., Ascano,
716 M., Tuschl, T., Ohler, U. and Keene, J. D. (2011). Integrative regulatory mapping indicates
717 that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability.
718 *Mol. Cell* **43**(3): 327-339.
719 Mukherjee, N., Jacobs, N. C., Hafner, M., Kennington, E. A., Nusbaum, J. D., Tuschl, T.,
720 Blackshear, P. J. and Ohler, U. (2014). Global target mRNA specification and regulation
721 by the RNA-binding protein ZFP36. *Genome Biol.* **15**(1): R12.

722 Nabors, L. B., Suswam, E., Huang, Y., Yang, X., Johnson, M. J. and King, P. H. (2003). Tumor
723 Necrosis Factor α Induces Angiogenic Factor Up-Regulation in Malignant Glioma Cells A
724 Role for RNA Stabilization and HuR. *Cancer Res.* **63**(14): 4181-4187.
725 Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T. and Quince, C. (2015). Insight into
726 biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform.
727 *Nucleic Acids Res.*: gku1341.
728 SEQC/MAQC-III-Consortium (2014). A comprehensive assessment of RNA-seq accuracy,
729 reproducibility and information content by the Sequencing Quality Control Consortium.
730 *Nat. Biotechnol.* **32**(9): 903-914.
731 Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. and Sirotkin, K.
732 (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**(1): 308-311.
733 Sievers, C., Schlumpf, T., Sawarkar, R., Comoglio, F. and Paro, R. (2012). Mixture models and
734 wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-
735 CLIP data. *Nucleic Acids Res.* **40**(20): e160.
736 Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M. and Ule, J. (2012). Analysis of
737 CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions.
738 *Genome Biol.* **13**(8): R67.
739 Tan, A. Y. and Manley, J. L. (2009). The TET family of proteins: functions and roles in disease. *J.*
740 *Mol. Cell. Biol.* **1**(2): 82-92.
741 Trapnell, C., Pachter, L. and Salzberg, S. L. (2009). TopHat: discovering splice junctions with
742 RNA-Seq. *Bioinformatics* **25**(9): 1105-1111.
743 van Dijk, E. L., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014). Ten years of next-generation
744 sequencing technology. *Trends Genet.* **30**(9): 418-426.
745 Wang, T., Xiao, G., Chu, Y., Zhang, M. Q., Corey, D. R. and Xie, Y. (2015). Design and
746 bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Res.* **43**(11):
747 5263-5274.
748 Zhang, C. and Darnell, R. B. (2011). Mapping in vivo protein-RNA interactions at single-
749 nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol.* **29**(7): 607-614.

750