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

Andreas Kloetgen, Arndt Borkhardt, Jessica I Hoell, Alice C McHardy

Background: Next-generation sequencing (NGS) 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.

Methods: 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 describe a set of freely available tools for this purpose, which are called the PAR-CLIP Analyzer suite (PARA-suite). The PARA-suite includes error model inference, PAR-CLIP read simulation, a full read alignment pipeline with a modified Burrows-Wheeler Aligner (BWA) algorithm, and CLIP read clustering.

Results: We examined 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. Our processing pipeline allowed improvement of both alignment and binding site detection accuracy. 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.

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 have 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

Key words: next-generation sequencing, read simulation, read alignment, cross-linking and immunoprecipitation (CLIP), post-transcriptional regulation, RNA-binding proteins

Abstract

Background: Next-generation sequencing (NGS) 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.

Methods: 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 describe a set of freely available tools for this purpose, which are 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 (BWA) algorithm and CLIP read clustering for binding site detection.

Results: We examined 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. Our processing pipeline allowed improvement of both alignment and binding site detection accuracy. 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.

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 #310°. Most RNAs maintain their function by working together with the so-called RNA-binding proteins (RBPs) (Glisovic, Bachorik et al. 2008). RBPs are virtually involved in 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 which 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, consisting 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 amyotrophic lateral sclerosis causing mutants of *FUS* have shown different RNA-binding patterns compared to the wild-type counterpart, 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 within 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 to 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 (UV) 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 also called a ‘T–C conversion’. T–C conversions can be used to distinguish between unspecifically 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 the 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). HITS-CLIP mainly introduces deletions of a single base at the cross-linked sites, while 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 sequencing of mammalian transcriptome libraries with a high coverage. Nowadays, the most commonly used NGS platforms are 454, Illumina, IonTorrent or 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 template (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 were recently 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 predefined position specific scoring matrix to process the error-prone PAR-CLIP reads.

In general, sequencing error profiles of RNA-Seq datasets, including PAR-CLIP data, can vary between different sequencing runs, depending on the sequencing machine, experimental conditions or 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 data-set specific error estimation. The alignment pipeline thus automatically adjusts to the quality and error profiles of individual sequencing datasets. We compared PAR-CLIP sequencing reads to regular transcriptome sequencing reads (RNA-Seq) to identify distinctive properties relevant for the reference-based read alignment and RBP binding site detection from PAR-CLIP datasets. Generation of simulated PAR-CLIP datasets can be performed with the PAR-CLIP 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. It is licensed under GNU GPLv3 and implemented in the programming languages Java and C.

Methods

2.1 Datasets and read aligners

We downloaded PAR-CLIP data of the *FET* family from DRASearch database (https://trace.ddbj.nig.ac.jp/DRASearch/) with accession number SRA025082 (Hoell, Larsson et al. 2011), *HuR* dataset with accession number SRR248532, *MOV10* dataset with accession number SRR490650 and HITS-CLIP data on the Argonaute protein (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 with the accession numbers SRR896663 and SRR896664 of an NGS quality assessment study (SEQC/MAQC-III-Consortium 2014) from DRASearch and pooled the data.

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

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 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.

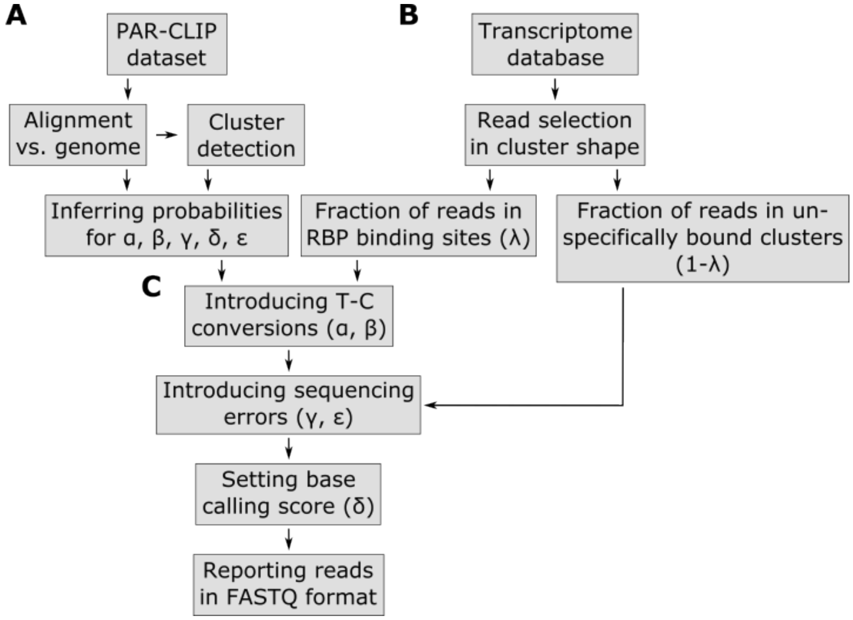

A
PAR-CLIP dataset
Alignment vs. genome
Cluster detection
Inferring probabilities for α, β, γ, δ, ε
B
Transcriptome database
Read selection in cluster shape
Fraction of reads in RBP binding sites (λ)
Fraction of reads in un-specifically bound clusters (1-λ)
C
Introducing T-C conversions (α, β)
Introducing sequencing errors (γ, ε)
Setting base calling score (δ)
Reporting reads in FASTQ format

Figure 1: Pipeline of the PAR-CLIP read simulator implemented in the PARA-suite. Part A describes the generation of the error profile and further 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, excluding PAR-CLIP specific T–C conversions, is estimated from the observed frequencies of all single nucleotide substitutions, except for T–C errors. 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 5 bases overlap of the genomic mapping positions as the clustering threshold. To identify high confidence clusters (sometimes referred to as binding sites) as defined in literature (Hafner, Landthaler et al. 2010), clusters which 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 reported as single nucleotide polymorphisms (SNP) 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 on also 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 (number of clusters 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 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 the distinctive alignment position shifts in the shape of a cluster are randomly introduced at this step (normal distribution with s.d. 1). A user-defined parameter $\lambda \in [0,1]$ specifies the fraction of clusters that are considered binding-sites, while the remaining clusters mimic contaminations of unbound RNAs which 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 based on the pre-computed probability matrix ε and 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 alignment (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. Per default, this is done with BWA, as our evaluations demonstrated it 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 the PARA-suite aligner (Section 2.4) against the reference sequence(s). All aligned reads are reported in a BAM file.

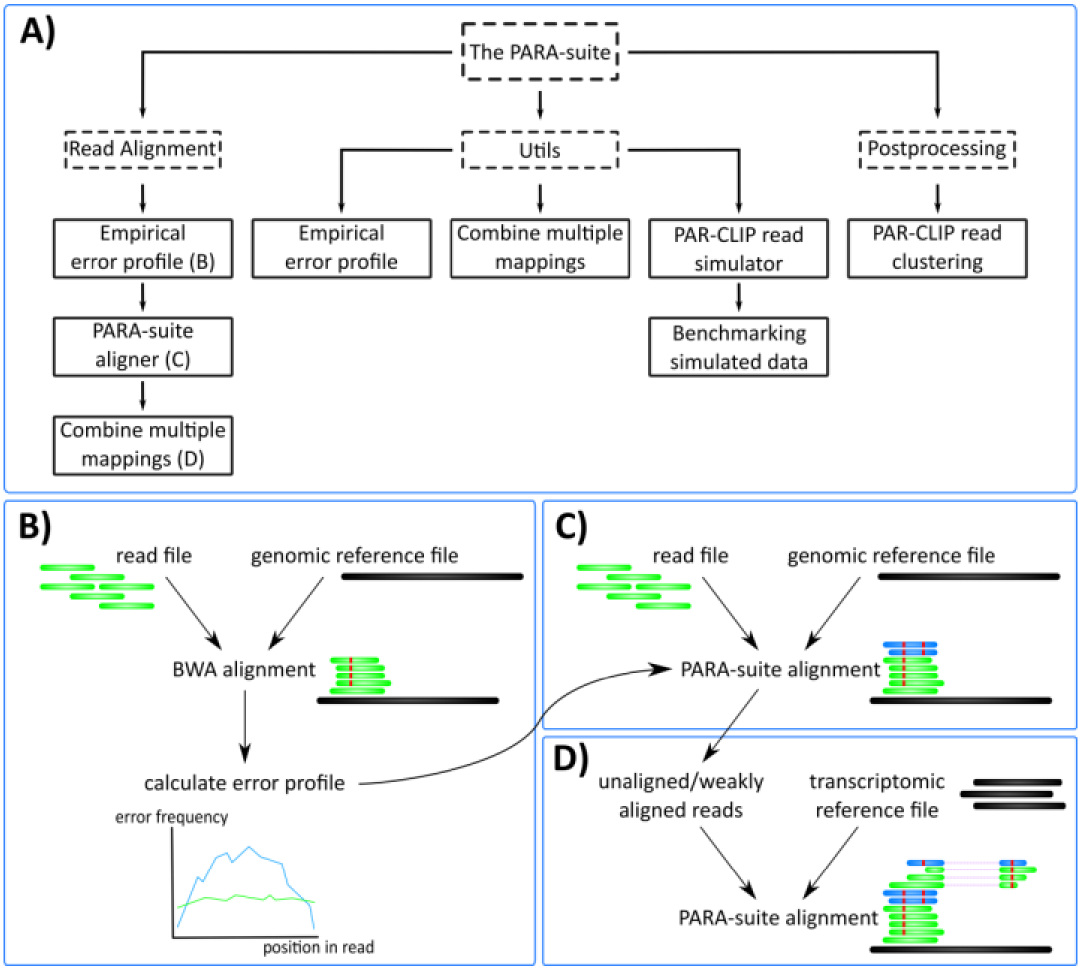


Figure 2: The PARA-suite. (A) The PARA-suite. Dashed boxes show packages while the other ones are executable programs. The Utils package includes tools for working with error-prone sequencing data and the postprocessing package contains a tool for clustering an aligned PAR-CLIP dataset to identify RBP-bound 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) The PARA-suite aligner is applied to the entire dataset to map error-prone reads, indicated here by the additional mapping of the two blue reads. (D) An optional alignment versus a transcriptome reference database can be executed using the PARA-suite aligner to identify previously unmapped reads.

2.4 Algorithm of the PARA-suite aligner

The general BWA algorithm uses a Burrows-Wheeler transform (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 (FM) index (Ferragina and Manzini 2000) - for a matching predecessor base of the bases of the read sequence 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).

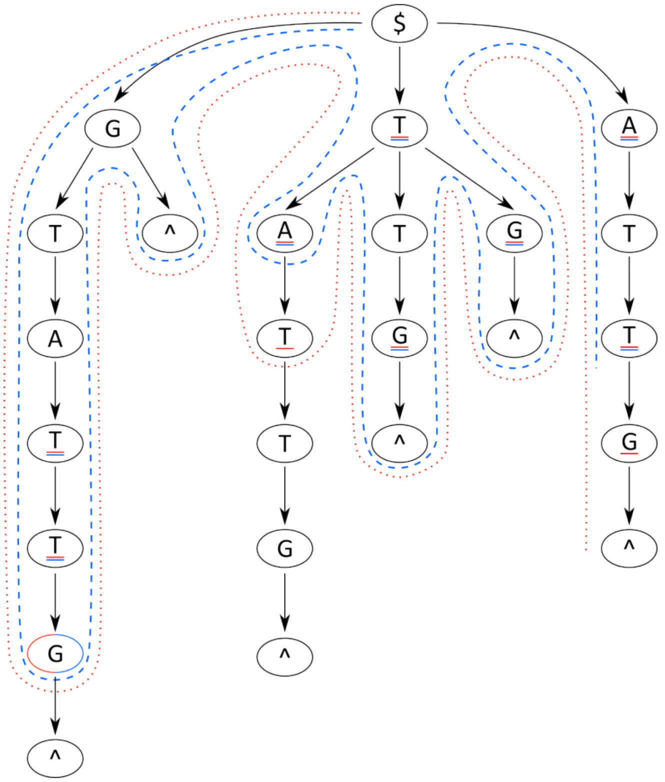


Figure 3: Suffix trie paths for BWA and PARA-suite. Paths of the algorithms through the suffix trie aligning the read sequence GCCATG$ against the reference sequence GTTATG$ (where $ means the end of a sequence). The red dotted line represents the algorithm of the BWA aligner, allowing for two mismatches, and the blue dashed line indicates the PARA-suite algorithm. The underlined bases represent positions where the respective aligner introduces a

mismatch. The example shows that the PARA-suite needs 14 comparisons while BWA needs 16 comparisons. Therefore, the PARA-suite is slightly faster than BWA at finding the same match represented by the red/blue circle (left path). Indels are not shown for simplicity.

The principle idea of the PARA-suite aligner is the introduction of a probability estimate for each comparison of the backward search. This enables weighting mismatches according to their probabilities they occur in the analyzed dataset. A sequencing run is initially characterized according to its underlying error probabilities. This allows to determine specific error-profiles for experimental techniques, such as the frequent T–C conversions in PAR-CLIP data, that 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 read bases {A, C, G, T} (Figure 4A). Indels are introduced during the alignment step separately, using estimated probabilities $I \in [0,1]$ for insertions and $D \in [0,1]$ for deletions.

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

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

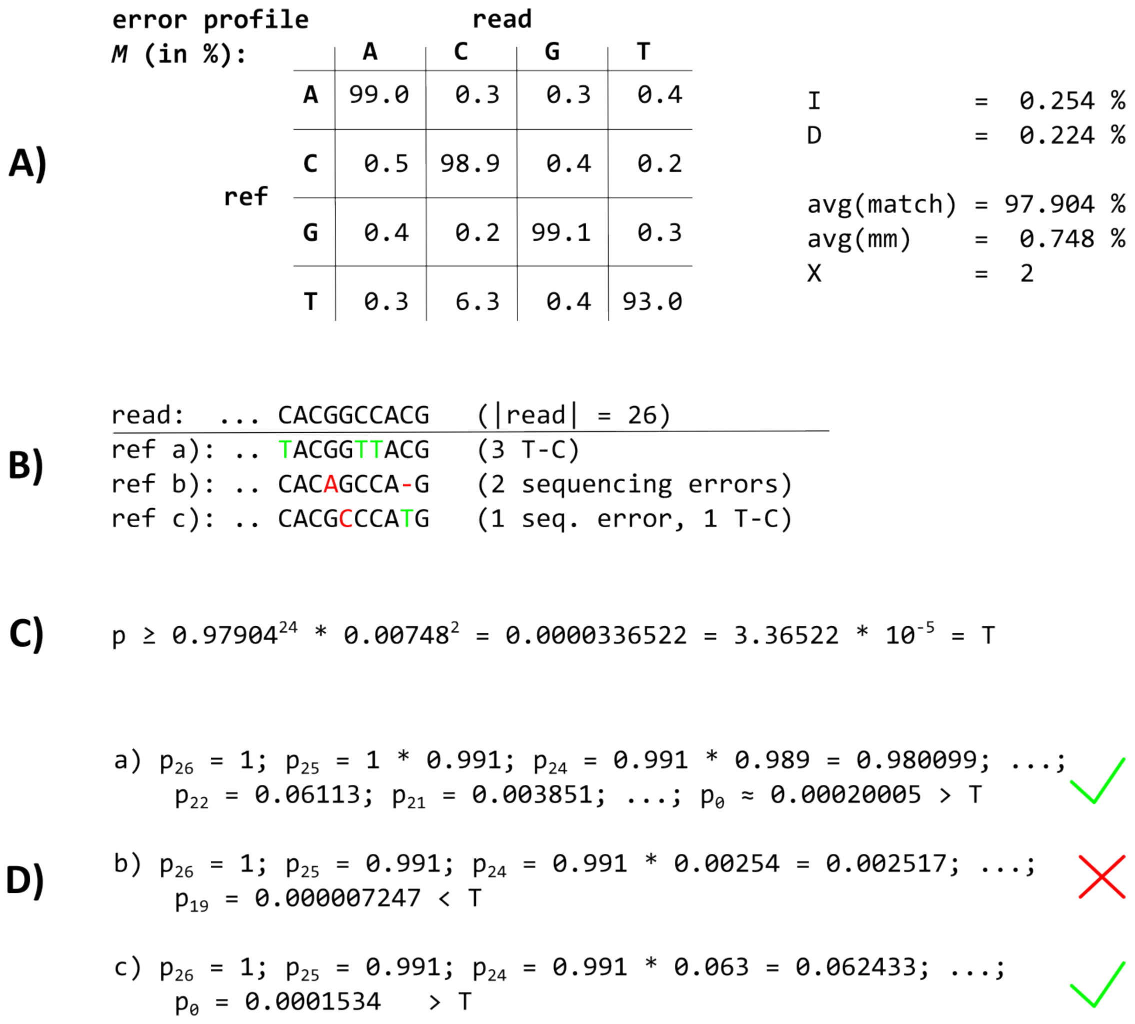
with $p_{|\text{read}|} = 1$, starting with $i = |\text{read}| - 1$ and decreasing *i* at each step, except in the case of a deletion, where $i \geq 0$.

Before the alignment of a particular read starts, a minimal threshold *T* for the probability *p* is necessary, to decide whether a reads is accepted as aligned or rejected. The calculation for *T* is dependent 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 as a default is 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*, while frequent errors, such as T–C errors in PAR-CLIP reads, are the most favored substitutions in the alignment process (Figure 4B–D).



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

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

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

Figure 4: The PARA-suite aligner approach. (A) The error profile probability matrix *M* and indel probabilities *I* and *D*, which are used as input for the PARA-suite alignment algorithm, as well as some results of the intermediate calculations of the PARA-suite alignment algorithm. In *M*, only T–C conversions have a higher probability (6.3%) compared to sequencing error and indel probabilities. (B) The last characters of a particular read and three example mapping positions within a reference, called ref a–c. (C) The calculation of a maximum threshold *T* for the mapping probability *p* (see formula in main text, and values from A in this image). (D) The mapping probability calculation of the read when mapping to the references a–c. The read fails to map against ref b with two sequencing errors, while reference a and 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* (*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 consist only of these outermost bases, they exhibited a 2–3 times 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 later on 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 first two properties (a) and (b) arise from the RNAse T1 treatment of the cells or the lysate during the PAR-CLIP experimental protocol. As only the short RNA fragments which are not digested by the endonuclease (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 reads had a length usually shorter than 30 bases. As the transcript regions outside of 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) arise from the incorporation of 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 the analysis of 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 utilized real datasets. The respective sequencing error and T–C conversion profiles were generated based on alignments of all three datasets against the human reference genome sequence version 38 (GRCh38) (Lander, Linton et al. 2001). The error profile and additional estimated distributions are similar to the ones from PAR-CLIP data on the two RBPs *HuR* and *MOV10*, indicating that these profiles represent 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 stacking into clusters bound by the RBP, to 65%. Such true RBP binding sites show 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 reflect RNA contaminations which 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 detected cluster of 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, except for 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 real PAR-CLIP data, presumably because of a small percentage of multiple mutations occurring at individual sites. Apart from this, the simulated datasets appear to be representative for 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 analyzed aligners, BWA and Bowtie, have often been employed 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 a 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 later referred to as those of the "PARA-suite pipeline", while the results of the genomic alignment step only using the PARA-suite are referred to as those of the "PARA-suite aligner". For the PARA-suite aligner, the sequencing error and T–C conversion profiles for the simulated datasets were obtained based on the BWA alignments allowing for two mismatches (BWA 2MMs) for each of the simulated datasets separately. For a performance overview, we estimated the average of the recall, precision and accuracy for every aligner over the 10 simulated datasets (calculation described in Supplementary Methods). Unfortunately, BMix does not report negative clusters (contaminations), thus we were not able to calculate the recall nor the accuracy, but only the precision.

In terms of overall performance, the PARA-suite alignment performed best, with an accuracy of 69.74% for the PARA-suite aligner 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 1, 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 comparing the PARA-suite alignment pipeline with these two aligners, the PARA-suite pipeline showed an increase of 16.95% and 4.97% in the overall accuracy, respectively. Notably, an average of 1.56% reads aligned by the PARA-suite pipeline are spanning an exon–exon junction, but were not identified by the genomic reference mapping step only, but required alignment versus 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 1, Figure S5). Here, the PARA-suite aligner and pipeline was ranked on places 1 and 3 regarding recall, and places 1 and 2 regarding precision, respectively, out of eight analyzed alignment scenarios. Hence, the PARA-suite aligner and pipeline offer a notable performance increase regarding all relevant performance measures in comparison to commonly used computational analysis setups.

We then tested the accuracy of the binding site detection algorithms BMix, PARalyzer and the hierarchical clustering of the PARA-suite using read alignment of the PARA-suite (Table S4). The hierarchical clustering identified most correct binding sites; 3.26% more correct sites than BMix and 5.54% more correct binding sites than PARalyzer. However, BMix identified fewer false binding sites in comparison to the hierarchical clustering (20.30% less), and compared to PARa lyzer (69.85% less). Furthermore, we investigated whether the PARA-suite alignment increased the number of detected binding sites, irrespective of the used detection algorithm. In conjunction with BMix, BWA 2MM (second best aligner) identified 7.17% less correct binding sites than the PARA-suite aligner. For PARalyzer, BWA 2MM identified 2.97% less than by the PARA-suite aligner. Finally, the hierarchical clustering identified 7.52% more correct binding sites for the PARA-suite than for BWA 2MM. Overall, the combination of BMix with the PARA-suite alignment provided the most accurate results on our simulated data.

Table 1: Alignment accuracy on simulated PAR-CLIP data. Most accurate alignment results for different parameter settings for every read aligner on 10 simulated PAR-CLIP datasets. The results are averaged per read aligner over all 10 datasets and are sorted by the accuracy.

Aligner	Accuracy (in %)	Recall (in %)	Precision (in %)	Mapped overall	Mapped correctly	Real time (s)	Memory (GB)
PARA-suite pipeline	73.14	84.49	71.85	1,024,792	969,948	396.8	6.27
PARA-suite	69.74	82.16	68.24	975,672	924,802	153.7	4.42
BWA 2MMs	68.17	82.31	64.98	959,171	904,034	359.2	4.42
Bowtie 2MMs	63.38	77.91	60.93	886,512	840,540	120.6	4.46
BWA PSSM	59.80	74.04	58.72	818,895	793,007	25.4	2.26
TopHat	59.69	76.10	55.35	844,902	791,549	282.9	-
Bowtie2	56.22	73.23	51.43	763,893	745,531	13.4	3.32
STAR	50.74	69.57	43.02	826,871	672,920	248.6	28.39
MOSAIK	47.61	66.12	39.24	1,294,747	632,656	9,481.4	194.80
Subjunc	35.42	50.61	26.09	597,400	469,751	64.2	6.65

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 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 BWA 2MMs, BWA PSSM and the PARA-suite aligner (without the transcriptome mapping step to achieve comparable results). We measured the fraction of aligned reads for all the aligners on the three datasets (Table S5). The PARA-suite aligner generated the largest fraction of aligned reads over all three datasets in comparison to BWA 2MM and BWA PSSM. Next, we stacked (clustered) all aligned reads using BMix and the hierarchical clustering tool of the PARA-suite (Table 2). BWA 2MM identified fewer binding sites compared to BWA PSSM or the PARA-suite, for read alignment prior to either BMix or hierarchical clustering. Using the hierarchical clustering, the PARA-suite reported the largest number of binding sites for two out of the three datasets. BWA PSSM identified 6.90% more clusters than the PARA-suite aligner for the *FUS* dataset, while the PARA-suite aligner identified 3.98% more clusters for the *EWSR1* dataset and 19.21% more clusters for the *TAF15* dataset compared to BWA PSSM. In comparison to the numbers reported in the original publication, the use of the PARA-suite aligner and hierarchical clustering increased the number of binding sites by 33.71% for *EWSR1*, 16.77% for *FUS* and decreased them by 12.56% for *TAF15*. After extracting distinct genes from all binding sites identified by the PARA-suite (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 identified genes, with 2702 genes targeted by all three RBPs (Figure S6).

Table 2: Detected binding sites for the *FET* protein family. Number of binding sites for the *FET* protein family identified by the three aligners BWA PSSM, BWA 2MMs and the PARA-suite in combination with the hierarchical clustering of the PARA-suite. Filters were applied according to Section 2.2.

	BWA 2MM / BMix	**BWA 2MM / Clustering**	**BWA PSSM / BMix**	**BWA PSSM / Clustering**	**PARA-suite / BMix**	**PARA-suite / Clustering**
EWSR1	20,703	22,760	24,639	27,550	25,478	28,692
FUS	14,768	36,861	19,628	51,606	19,006	48,042
TAF15	5,086	5,810	5,238	6,130	5,862	7,588

3.5 Analysis of PAR-CLIP data on *HuR*

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

We used the read aligners Bowtie2, Bowtie 2MM, BWA 2MM, BWA PSSM and the PARA-suite to align the pre-processed dataset against the human genome reference GRCh38. Then, we applied BMix and the hierarchical clustering of the PARA-suite to determine the binding sites of *HuR* derived using the different read aligners. BWA PSSM in conjunction with BMix identified most RBP binding sites within the genome – 3.69% more than the PARA-suite (Table S6). When comparing the detected binding sites of BMix and the PARA-suite hierarchical clustering for alignments created by the PARA-suite aligner (binding site positions overlapping by at least 13 bases), the difference was only marginal, with an overlap of more than 98.25% for the two methods. A recent study of this dataset reported binding sites using Bowtie 2MM for the alignment step and PARalyzer for the binding site detection. We found the use of any BWA PSSM or the PARA-suite alignment in conjunction with either BMix or hierarchical clustering to increase the number of detected binding sites by 2.87% – 7.84%.

We searched for the exact binding motifs of *HuR* (ATTTA, ATTTTA and ATTTTTA) within the BMix detected binding sites within 3’ UTRs or introns for all tested read aligners. We found that all aligners performed comparably, with motifs present in 42% to 44% of all detected binding sites. The largest fraction was achieved using read alignments with BWA PSSM (44.33%), while the PARA-suite aligner in combination with BMix found 42.53% most likely correct binding sites. 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 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 of the *HuR* dataset (Mukherjee, Corcoran et al. 2011) only reports 12 out of these 15 genes with a confirmed binding site.

Discussion

We here describe a read simulator to mimic PAR-CLIP datasets with error profiles drawn from real PAR-CLIP datasets and the PARA-suite pipeline for error-aware read alignment and processing. Furthermore, we provide 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, to identify the most accurate read aligner and parameter settings on PAR-CLIP reads. Common read simulators such as ART or GemSim do not allow simulating PAR-CLIP reads with their specific error profiles. We characterized some of the unique properties of PAR-CLIP sequence datasets that have, to our knowledge, so far not been analyzed, such as 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 in comparison to 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 into the simulation process.

Based on the simulated PAR-CLIP datasets, we determined parameter settings delivering 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 analyses showed the read alignment to be crucial for RBP binding site detection from 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 contrary assumptions, such as a similar read coverage across all exons or long reads, to achieve high confidence k-mer spectra. Accordingly, these three aligners were outperformed by other methods (Table S3–4). In addition, recent studies have shown that BWT based aligners have less sensitivity in regions with genomic variation (Gontarz, Berger et al. 2013). Interestingly, MOSAIK, an error-aware aligner based on hash queries that was shown to be more robust on RNA-Seq reads than BWT based aligners (Lee, Stromberg et al. 2014), was also outperformed by most other tested methods. Although it is robust on longer

RNA-Seq reads, it seems to struggle with the very short PAR-CLIP reads. The PARA-suite alignment pipeline allowed to increase the fraction of aligned reads in comparison to other aligners, including 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 applied downstream binding site detection algorithm. Different from 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 with the BWA algorithm. Another difference to the BWA PSSM algorithm is that the latter introduces mismatches under consideration of the base calling quality scores and a probabilistic background model for matching bases in addition to the input error profile. In contrast, the generic error profile estimation of the PARA-suite is not limited to any specific input profile. Further applications of our software could thus be the analysis of other types of error-prone sequencing data such as bisulphite sequencing data, which introduces a high amount of C–T mutations (Frommer, McDonald et al. 1992) or data from low-quality ancient DNA samples (Briggs, Stenzel et al. 2007).

Our analysis of combinations of read aligners and binding site detection algorithms on simulated and real datasets indicate that no single software performed best in terms of binding site detection on the available PAR-CLIP datasets. This observation was recently also made on further datasets (Kassuhn, Ohler et al. 2016). Our analysis of the *HuR* and *FUS* datasets revealed that U-rich binding sites tended to show higher rates of T–C conversions per read and were best aligned by BWA PSSM. RBPs with a more heterogeneous nucleotide distribution (e.g. *EWSR1* and *TAF15*) within the binding site are better assessed by the PARA-suite aligner. This is supported by an analysis of uridylate-rich sequences from our simulated data aligned by BWA PSSM and the PARA-suite (Supplementary Results and Supplementary Table 7). Thus, a preliminary analysis of the error profile using the PARA-suite error profiler could allow determining the best approach for analyzing 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. *CA: Digital Equipment Corporation* **Technical Report 124 Palo Alto**.

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., Wilder, S. P., Zadissa, A., Aken, B. L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S. M., Spudich, G., Trevanion, S. J., Yates, A., Zerbino, D. R. and Flicek, P. (2015). Ensembl 2015. *Nucleic Acids Res.* **43**(Database issue): D662-669.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1): 15-21.

Ferragina, P. and Manzini, G. (2000). Opportunistic data structures with applications. *Proceedings of the 41st Symposium on Foundations of Computer Science*: 390-398.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L. and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U S A* **89**(5): 1827-1831.

Glisovic, T., Bachorik, J. L., Yong, J. and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582**(14): 1977-1986.

Golumbeanu, M., Mohammadi, P. and Beerenwinkel, N. (2015). BMix: probabilistic modeling of occurring substitutions in PAR-CLIP data. *Bioinformatics*: btv520.

Gontarz, P. M., Berger, J. and Wong, C. F. (2013). SRmapper: a fast and sensitive genome-hashing alignment tool. *Bioinformatics* **29**(3): 316-321.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M. and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**(1): 129-141.

Hieronymus, H. and Silver, P. A. (2004). A systems view of mRNP biology. *Genes Dev.* **18**(23): 2845-2860.

Hoell, J. I., Hafner, M., Landthaler, M., Ascano, M., Farazi, T. A., Wardle, G., Nusbaum, J., Cekan, P., Khorshid, M. and Burger, L. (2014). Transcriptome-Wide Identification of Protein Binding Sites on RNA by PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation). *Handbook of RNA Biochemistry: Second, Completely Revised and Enlarged Edition*. A. B. R.K. Hartmann, A. Schön, and E. Westhof. Weinheim, Wiley-VCH Verlag GmbH & Co. KGaA. **II:** 877-898.

Hoell, J. I., Larsson, E., Runge, S., Nusbaum, J. D., Duggimpudi, S., Farazi, T. A., Hafner, M., Borkhardt, A., Sander, C. and Tuschl, T. (2011). RNA targets of wild-type and mutant FET family proteins. *Nat. Struct. Mol. Biol.* **18**(12): 1428-1431.

Huang, W., Li, L., Myers, J. R. and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics* **28**(4): 593-594.

Kassuhn, W., Ohler, U. and Drewe, P. (2016). *Cseq-simulator: a data simulator for CLIP-Seq experiments*. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing.

Kerpedjiev, P., Frellsen, J., Lindgreen, S. and Krogh, A. (2014). Adaptable probabilistic mapping of short reads using position specific scoring matrices. *BMC Bioinformatics* **15**(1): 100.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M. and Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods* **8**(7): 559-564.

Kloetgen, A., Münch, P. C., Borkhardt, A., Hoell, J. I. and McHardy, A. C. (2015). Biochemical and bioinformatic methods for elucidating the role of RNA-protein interactions in posttranscriptional regulation. *Brief. Funct. Genomics* **14**(2): 102-114.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M. and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **17**(7): 909-915.

Laehnemann, D., Borkhardt, A. and McHardy, A. C. (2015). Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Brief. Bioinform.*: bbv029.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M. and FitzHugh, W. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**(4): 357-359.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**(3): R25.

Lebedeva, S., Jens, M., Theil, K., Schwanhäusser, B., Selbach, M., Landthaler, M. and Rajewsky, N. (2011). Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol. Cell* **43**(3): 340-352.

Lee, W.-P., Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P. and Marth, G. T. (2014). MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS One* **9**(3): e90581.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Liao, Y., Smyth, G. K. and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**(10): e108.

Lukong, K. E., Chang, K.-w., Khandjian, E. W. and Richard, S. (2008). RNA-binding proteins in human genetic disease. *Trends Genet.* **24**(8): 416-425.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**(1): 10-12.

McElroy, K. E., Luciani, F. and Thomas, T. (2012). GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics* **13**(1): 74.

Mukherjee, N., Corcoran, D. L., Nusbaum, J. D., Reid, D. W., Georgiev, S., Hafner, M., Ascano, M., Tuschl, T., Ohler, U. and Keene, J. D. (2011). Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell* **43**(3): 327-339.

Mukherjee, N., Jacobs, N. C., Hafner, M., Kennington, E. A., Nusbaum, J. D., Tuschl, T., Blackshear, P. J. and Ohler, U. (2014). Global target mRNA specification and regulation by the RNA-binding protein ZFP36. *Genome Biol.* **15**(1): R12.

Nabors, L. B., Suswam, E., Huang, Y., Yang, X., Johnson, M. J. and King, P. H. (2003). Tumor Necrosis Factor α Induces Angiogenic Factor Up-Regulation in Malignant Glioma Cells A Role for RNA Stabilization and HuR. *Cancer Res.* **63**(14): 4181-4187.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T. and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.*: gku1341.

SEQC/MAQC-III-Consortium (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**(9): 903-914.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**(1): 308-311.

Sievers, C., Schlumpf, T., Sawarkar, R., Comoglio, F. and Paro, R. (2012). Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Res.* **40**(20): e160.

Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M. and Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol.* **13**(8): R67.

Tan, A. Y. and Manley, J. L. (2009). The TET family of proteins: functions and roles in disease. *J. Mol. Cell. Biol.* **1**(2): 82-92.

Trapnell, C., Pachter, L. and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9): 1105-1111.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet.* **30**(9): 418-426.

Zhang, C. and Darnell, R. B. (2011). Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol.* **29**(7): 607-614.