The importance of incorporating transcript information in CLIP-seq data analysis

Michael Uhl
Albert-Ludwigs-Universitat Freiburg Technische Fakultat https://orcid.org/0000-0002-9563-4991
Dinh Van Tran
Albert-Ludwigs-Universitat Freiburg Technische Fakultat https://orcid.org/0000-0002-7357-4959
Rolf Backofen (backofen@informatik.uni-freiburg.de)
https://orcid.org/0000-0001-8231-3323

Research article

Keywords: CLIP-seq, eCLIP, Peak calling, RBP binding site prediction

Posted Date: March 31st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-18225/v1

RESEARCH

The importance of incorporating transcript information in CLIP-seq data analysis

Michael Uhl[1], Van Dinh Tran[1] and Rolf Backofen[1,2*]

*Correspondence:
backofen@informatik.uni-freiburg.de
[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany
Full list of author information is available at the end of the article

Abstract

Background: Current peak callers for identifying RNA-binding protein (RBP) binding sites from CLIP-seq data take into account genomic read profiles, but they ignore the underlying transcript information, that is information regarding splicing events. So far, there are no studies available that closer observe this issue.

Results: Here we show that current peak callers are susceptible to false peak calling near exon borders. We further quantify its extent in publicly available datasets, which turns out to be substantial. Finally, by providing a tool called CLIPcontext for automatic transcript and genomic context sequence extraction, we demonstrate that context choice also affects the performances of RBP binding site prediction tools.

Conclusions: Our results demonstrate the importance of incorporating transcript information in CLIP-seq data analysis. Taking advantage of the underlying transcript information should therefore become an integral part of future peak calling and downstream analysis tools.

Keywords: CLIP-seq; eCLIP; Peak calling; RBP binding site prediction

Background

Over the last decade, CLIP-seq (cross-linking and immunoprecipitation followed by next generation sequencing) [1] has become the state-of-the-art procedure to experimentally determine the precise transcriptome-wide binding locations of RNA-binding proteins (RBPs). Many variants have been introduced, out of which PAR-CLIP [2], iCLIP [3], and eCLIP [4] are currently the most widely used. Regardless of the variant, CLIP-seq is usually applied in vivo to a specific RBP, producing a library of reads bound by the RBP. Identification of binding sites is subsequently achieved by mapping the reads back to the corresponding reference genome and running a so called peak caller tool on the read profiles. A number of popular peak callers have emerged over the years, such as Piranha [5], CLIPper [6], PEAKachu [7], and PureCLIP [8].

While there exist various protocol-specific as well as more generic peak callers [9], none of the current tools takes into account the transcript information underlying the mapped reads. Instead, they extract binding regions directly from the genomic read profiles. This can be acceptable if the studied RBP binds intronic sequences or in general unspliced RNAs. However, if the RBP is actually predominantly binding to spliced RNAs, which should be true for most cytoplasmically active RBPs, ignoring transcript information potentially leads to false peak calling and the inclusion of non-authentic sequence context. This in turn can compromise the results

of downstream analysis tools like motif finders or binding site predictors, which usually take the genomic sequence context for extending the binding sites as well.

Here we show that current peak callers indeed have problems with correctly defining binding sites for RBPs binding predominantly to exonic regions. We further look at publicly available eCLIP datasets with binding sites identified by CLIPper and present comprehensive statistics regarding exonic binding frequencies. Focusing specifically on sites near exon borders, we report the extent of sites mostly affected by context sequence selection and false peak calling. To compare different sequence contexts, we implemented a tool called CLIPcontext for automatic transcript and genomic context sequence extraction. Finally, by applying CLIPcontext on sites near exon borders and use these for training three different binding site prediction tools, we demonstrate that sequence context choice can have a large impact on predictive performance.

Results and discussion

Ignoring transcript information compromises peak calling

To illustrate the issues current peak callers have with predominantly exon-binding RBPs, we chose one out of many eCLIP RBP cell type combinations (YBX3 K562) with large amounts of exonic binding regions (see Table S1 for eCLIP overlap statistics). In this paper, we call or count peak regions as overlapping or exon binding if they have an overlap of $\geq 90\%$ with exonic regions. 84.6% of YBX3 K562 merged peak sites overlap with exonic regions, out of which 51.0% are ≤ 50 nt away from exon borders. Fig. 1 shows the YBX3 K562 genomic reads profile visualized via IGV (Integrative Genomics Viewer) [10] over two genomic regions, with added peak regions from CLIPper, PEAKachu, and PureCLIP. Fig. 1a depicts a genomic region of 11 kb, containing the *PRDX6* gene. We can see that the read alignments clearly follow the exon annotations: most reads map to exons, including many intron-spanning ones (blue-gray lines), while only few reads map to introns. Not surprisingly, all three peak callers only report exonic peaks, often close or directly at exon borders. Given the alignment information, extending these peak regions with genomic context, as usually done prior to further analysis, is not correct. Instead, the transcript context of the spliced RNA should be used, which is where the actual RBP binding occurs. Zooming in on the matter, Fig. 1b shows a genomic region of 562 bp, comprising exon 6 and 7 of the *DDOST* gene. Again the mapped reads strongly suggest a spliced RNA context, given the many intron-spanning reads and almost no intron coverage. Keeping the intron therefore leads to an artificial split-up of peak regions spanning the exon border. Unaware of the split, peak callers might consequently call two peaks, whereas they should have treated the split peaks as one contiguous region.

In the Fig. 1b example, both CLIPper and PureCLIP call peaks at adjacent exon borders, while PEAKachu even calls a single peak over the entire intron. In general, PEAKachu and CLIPper define peak regions by fitting functions (Gaussian density versus splines) on the mapped reads. More precisely, CLIPper fits splines on the genomic read coverage profile counting each base of a read once, while PEAKachu replaces each read with a Gaussian, using the genomic mean of read start and end as the center of the Gaussian. Both methods thus have problems with split

reads, leading to PEAKachu calling peaks over introns in the presence of intron-spanning reads, and CLIPper calling peaks at exon ends with shared read coverage. In contrast, PureCLIP uses read starts to identify crosslink sites, which later can be merged into peak regions. This circumvents the described problems, as each read is considered only once at one genomic position. On the other hand, it still can be fooled since intron-spanning reads are treated no different to contiguously aligned reads. For the YBX3 dataset and with default settings, PEAKachu tends to call broader peaks than CLIPper, while PureCLIP peaks are much shorter (see Table S2 for peak statistics).

Exon binding is substantial in public CLIP-seq data

To quantify the extent of exon and near exon border binding in eCLIP data, CLIPper peak regions from 223 eCLIP datasets were overlapped with exon regions featuring strong experimental evidence (see methods section on data preparation). As shown in Fig. 2a, 61 datasets (27.4%) feature $\geq 50\%$ exonic sites, with 14 datasets (6.3%) even reaching $\geq 75\%$ (see Table S1 for full statistics on each dataset). Table S1 also lists the ratios of sites near exon borders and pair sites, i.e., two sites located at adjacent exon borders. Looking closer at the 61 datasets, 63.3% of exonic sites lie within ≤ 50 nt to exon borders, and 20.7% form pairs (< 10 nt distance of site ends to exon borders required for both sites of the pair). We thus have a substantial amount of sites susceptible to split peak calling and false sequence context choice. Since the selection procedure for splice isoforms (i.e., their exon regions) was quite strict, the actual percentages should be even higher. As the data features experiments from K562 and HepG2 cell lines, we also looked at the correlation of percentages for RBPs with experiments in both lines. Fig. 2b shows the correlation plot of exon site ratios, resulting in an R^2 score of 0.76. This suggests a general agreement in the amount of exon binding across cell lines. On the other hand, it also shows that classifying RBPs into spliced or unspliced binding oversimplifies actual binding patterns. Instead, the correct site context needs to be determined directly from the mapped data. One might wonder whether potentially problematic pair sites could easily be filtered out based on their assigned scores (i.e., log_2 fold changes) prior to data analysis. As shown in Fig. 2c, this is not the case, with an average score of 2.47 for pair sites and 2.17 for all exonic sites.

Sequence context influences binding site prediction performances

Based on the considerable amount of sites near exon borders, we further investigated whether different sequence contexts could also influence the performances of binding site prediction tools. For this we constructed different context datasets for 5 RBPs, by focusing on RBPs with high amounts of exonic sites (≥ 80 %) and choosing 5 RBPs randomly within this range (see methods section for details). We kept only sites ≤ 50 nt from exon borders and extended the sites 80 nt up- and downstream with both genomic and transcript context (see Table S3 for dataset details). Note that this also includes sites at transcript ends, where full extension is only possible in the genomic context case. To assess any effects, three different prediction tools (DeepBind [11], GraphProt [12], and GraphProt2 [13]) were run on both context sets, using 10-fold cross validation and no additional features (i.e., only sequence

information). Figure 2d shows the performance results as average accuracies over the 5 datasets, for both genomic and transcript context sets (see Table S4 for detailed results). As we can see, using the more authentic transcript context considerably improves accuracies for all three tools, showcasing that context sequence choice can have a large influence on predictive performance and thus on what is learned. One could argue that including large amounts of context sequence bears the risk of learning binding site-unspecific patterns. We acknowledge that this can influence predictions. Some bias from the negative set is also possible, although we tried to minimize this by random sampling from the whole gene sequence and no overlap with positive sites. On the other hand, intronic context near exon borders also harbors various recognizable regions, like the polypyrimidine tract, or splice donor and acceptor sites, which can lead to wrong conclusions for spliced RNA binding RBPs. Moreover, learning the transcript context for RBPs binding to spliced RNA can also be advantageous, especially when predicting on gene sequences that contain introns. Here we conveniently extracted the transcript context from a set of well annotated splice isoforms, ignoring the context information contained in the CLIP-seq data. Future binding site prediction tools should therefore focus on identifying the actual sequence context directly from the CLIP-seq data, ideally by employing a peak caller that implements this functionality.

Conclusions

In this paper we raised the issue of ignoring transcript information in the process of peak calling and beyond. We showed that current peak callers by design are prone to false peak calling near exon borders, and that peak regions near exon borders are frequent in publicly available datasets. We also saw that sequence context choice has a profound effect on predicting sites near exon borders. Taken together, incorporating transcript information leads to more authentic results and thus should become an integral feature of future peak calling and downstream analysis methods.

Methods

Data preparation and exon overlap statistics

eCLIP datasets out of two cell lines (HepG2, K562) were downloaded from the ENCODE project website [14] (November 2018 release). Altogether the data covers 150 RBPs, divided into 103 HepG2 and 120 K562 sets, resulting in 223 datasets. We directly used the genomic binding regions (genome assembly GRCh38) determined by CLIPper, available in BED format for each replicate (2 replicates per dataset). For each RBP cell type combination, replicate binding sites were merged by keeping only the sites with the highest log_2 fold change (FC) in case of overlapping sites. After filtering sites by FC ≥ 1, sites were overlapped with exon regions of the most prominent transcripts using intersectBed (bedtools 2.29.0 [15]) and a required exon overlap $\geq 90\%$ for a region to be counted as exon overlapping. We defined the most prominent isoform of a gene based on the information Ensembl (Ensembl Genes 97, GRCh38.p12) provides for each transcript through hierarchical filtering: APPRIS annotation [16] (highest priority, labels principal1-5), and transcript support level (TSL, labels 1-5). We considered only genes with isoforms featuring these labels and transcripts that belong to the GENCODE basic gene set, resulting in 29,798

isoforms and 238,271 exon regions. Exon overlap statistics for the 223 datasets are stored in Table S1.

Peak caller setup

To illustrate potential peak caller problems, we chose an RBP cell type combination with a relatively high amount of exonic binding regions (YBX3 K562, 84.6%), out of which 51.0% are close to exon borders (region ends $\leq$ 50 nt from exon borders, see Table S1 for statistics). We collected peak regions identified by three peak callers: CLIPper, PEAKachu, and PureCLIP. For CLIPper, we downloaded the called peak regions from ENCODE (https://www.encodeproject.org, ENCODE ID ENCFF210TQC), selected the peaks called on replicate 1, and filtered them by a minimum log_2 fold change of 1. For PEAKachu and PureCLIP, we downloaded the mapped reads in BAM format (replicate 1 and mock input control, ENCODE IDs ENCFF184OIG and ENCFF409LAD) and used the R2 reads (second pair reads) as experiment and control libraries. PEAKachu was run on Galaxy [17] (https://usegalaxy.eu, Galaxy tool version 0.1.0.2) with default settings and a fold threshold of 2. PureCLIP (version 1.3.1) was installed locally and run with default parameters, setting $-dm$ 8 for merging called crosslink sites into peak regions.

Construction of sequence context sets

For comparing the effects of different sequence contexts on predictive performance, we randomly chose 5 eCLIP sets featuring high percentages of exonic binding sites (from 75.69 to 91.94%, see Table S1). All exonic sites within $\leq$ 50 nt of exon borders were selected, filtered by FC to obtain $\sim$ up to 10,000 binding regions, and mapped to the transcriptome using our tool CLIPcontext (minimum exon overlap set to 90%), taking the most prominent transcript isoform for each gene as described. Centered exon mapping sites were extended to 61 nt, and overlapping sites at exon borders were merged by keeping the one with higher FC. To include more context, we further extended the sites up- and downstream by 80 nt (new maximum length 221 nt, see CLIPcontext parameters). CLIPcontext outputs both the transcript context set and its corresponding genomic context set. Finally, to generate one negative set for both genome and transcript context sets, we randomly selected genomic sites based on two criteria: 1) their location on genes covered by eCLIP peak regions and 2) no overlap with any eCLIP peak regions from the experiment. Sequence context set statistics are stored in Table S3.

Tool setup for context predictions

Three RBP binding site prediction tools (DeepBind, GraphProt, and GraphProt2) were trained on the described context sets (see previous methods section). DeepBind models were trained using the DeepRAM [18] framework, including hyperparameter optimization. For GraphProt, model hyperparameters were optimized using a separate optimization set of 500 positive and 500 negative examples. GraphProt2 models were trained using default parameters and no hyperparameter optimization. Both GraphProt and GraphProt2 models were trained using only sequence features. The accuracy measure, i.e., the proportion of correctly classified instances, was used in combination with 10-fold cross validation to measure model performances over

5 datasets. Accuracies are reported in Table S4, together with standard deviations from cross validation (except for GraphProt, since it does not output single accuracies during cross validation).

CLIPcontext availability and usage instructions

CLIPcontext is available together with a comprehensive documentation on GitHub (https://github.com/BackofenLab/CLIPcontext), as well as on Bioconda (https://anaconda.org/bioconda/clipcontext). Besides mapping of input sites to the transcriptome or genome, CLIPcontext also offers modes for the extraction of: sites near exon borders, a list of most prominent transcripts, intronic sites, or exon and intron regions for a given set of transcripts.

Abbreviations

CLIP-Seq: cross-linking and immunoprecipitation followed by next generation sequencing; eCLIP: enhanced CLIP; iCLIP: individual-nucleotide CLIP; IGV: Integrative Genomics Viewer; PAR-CLIP: photoactivatable-ribonucleoside-enhanced CLIP; RBP: RNA-binding protein

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
CLIPcontext is available on GitHub (https://github.com/BackofenLab/CLIPcontext) and Bioconda (https://anaconda.org/bioconda/clipcontext). Supplementary data is also stored in the GitHub repository.

Competing interests
The authors declare that they have no competing interests.

Funding
MU was funded by Deutsche Forschungsgemeinschaft (DFG) grant BA 2168/11-1 SPP 1738 and BA2168/11-2 SPP 1738. VDT was funded by DFG grant BA 2168/3-3.

Author's contributions
RB and MU conceived the study. VDT and MU performed the binding site predictions. MU performed the remaining data analysis, wrote the draft, and implemented the software. RB, VDT, and MU contributed to and approved the final manuscript.

Acknowledgements
We thank Martin Raden and Gabriel Pratt for their invaluable suggestions on the topic.

Author details

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany. [2]Signalling Research Centres BIOSS and CIBSS, University of Freiburg, Schaenzlestr. 18, 79104 Freiburg, Germany.

References

1. Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., *et al.*: Hits-clip yields genome-wide insights into brain alternative rna processing. Nature **456**(7221), 464 (2008)
2. Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano Jr, M., Jungkamp, A.-C., Munschauer, M., *et al.*: Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. Cell **141**(1), 129–141 (2010)
3. König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., Ule, J.: iclip reveals the function of hnrnp particles in splicing at individual nucleotide resolution. Nature structural & molecular biology **17**(7), 909 (2010)
4. Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., *et al.*: Robust transcriptome-wide discovery of rna-binding protein binding sites with enhanced clip (eclip). Nature methods **13**(6), 508 (2016)

5. Uren, P.J., Bahrami-Samani, E., Burns, S.C., Qiao, M., Karginov, F.V., Hodges, E., Hannon, G.J., Sanford, J.R., Penalva, L.O., Smith, A.D.: Site identification in high-throughput rna–protein interaction data. Bioinformatics **28**(23), 3013–3020 (2012)
6. Lovci, M.T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T.Y., Stark, T.J., Gehman, L.T., Hoon, S., *et al.*: Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nature structural & molecular biology **20**, 1434 (2013)
7. Bischler, T., Maticzka, D., Förstner, K.U., Wright, P.R.: PEAKachu. https://github.com/tbischler/PEAKachu
8. Krakau, S., Richard, H., Marsico, A.: PureCLIP: capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data. Genome biology **18**(1), 240 (2017)
9. Uhl, M., Houwaart, T., Corrado, G., Wright, P.R., Backofen, R.: Computational analysis of clip-seq data. Methods **118**, 60–72 (2017)
10. Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P.: Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. Briefings in bioinformatics **14**(2), 178–192 (2013)
11. Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J.: Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. Nature biotechnology **33**(8), 831 (2015)
12. Maticzka, D., Lange, S.J., Costa, F., Backofen, R.: Graphprot: modeling binding preferences of rna-binding proteins. Genome biology **15**(1), 17 (2014)
13. Uhl, M., Heyl, F., Backofen, R., et al.: Graphprot2: A novel deep learning-based method for predicting binding sites of rna-binding proteins. bioRxiv, 850024 (2019)
14. Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., *et al.*: Encode data at the encode portal. Nucleic acids research **44**(D1), 726–732 (2015)
15. Quinlan, A.R., Hall, I.M.: Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010)
16. Rodriguez, J.M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.-J., Lopez, G., Valencia, A., Tress, M.L.: Appris: annotation of principal and alternative splice isoforms. Nucleic acids research **41**(D1), 110–117 (2012)
17. Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., *et al.*: The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic acids research **46**(W1), 537–544 (2018)
18. Trabelsi, A., Chaabane, M., Ben-Hur, A.: Comprehensive evaluation of deep learning architectures for prediction of dna/rna sequence binding specificities. Bioinformatics **35**(14), 269–277 (2019)

Figures

Figure 1 IGV snapshot of two genomic regions with mapped YBX3 K562 eCLIP data. 1: read profile (coverage), 2: read alignments, 3: gene annotations (thick blue regions are exons, thin blue regions introns), 4: CLIPper peaks, 5: PEAKachu peaks, 6: PureCLIP peaks. For clarity only gene strand reads from replicate 1 are displayed. **a** *PRDX6* whole gene region (length 11 kb, maximum read coverage 1141). **b** *DDOST* gene exons 6 and 7 region (length 562 bp, maximum read coverage 167).

Figure 2 Exon binding statistics of eCLIP datasets and prediction results for different sequence contexts. **a** Distribution of exonic site ratios for 223 eCLIP datasets over four percentage ranges. For each range, the percentage (number) of sets with ratios falling into this range is given. **b** Correlation plot of exonic site ratios for RBPs present in two cell lines (HepG2 and K562). **c** Site score distributions for all exonic sites and exonic sites that form pairs by being located at adjacent exon borders. log_2 fold change values of the sites determined by CLIPper were taken as site scores. Only pair sites with a distance of < 10 nt to exon borders were considered. **d** Average classification accuracies over 5 eCLIP datasets for 3 RBP binding site prediction methods, comparing genome and transcript context.

Additional Files

Additional file 1: Table S1: Exon overlap statistics of ENCODE eCLIP datasets (.xlsx)
Additional file 2: Supplementary tables S2-S4. (.pdf)

Figures

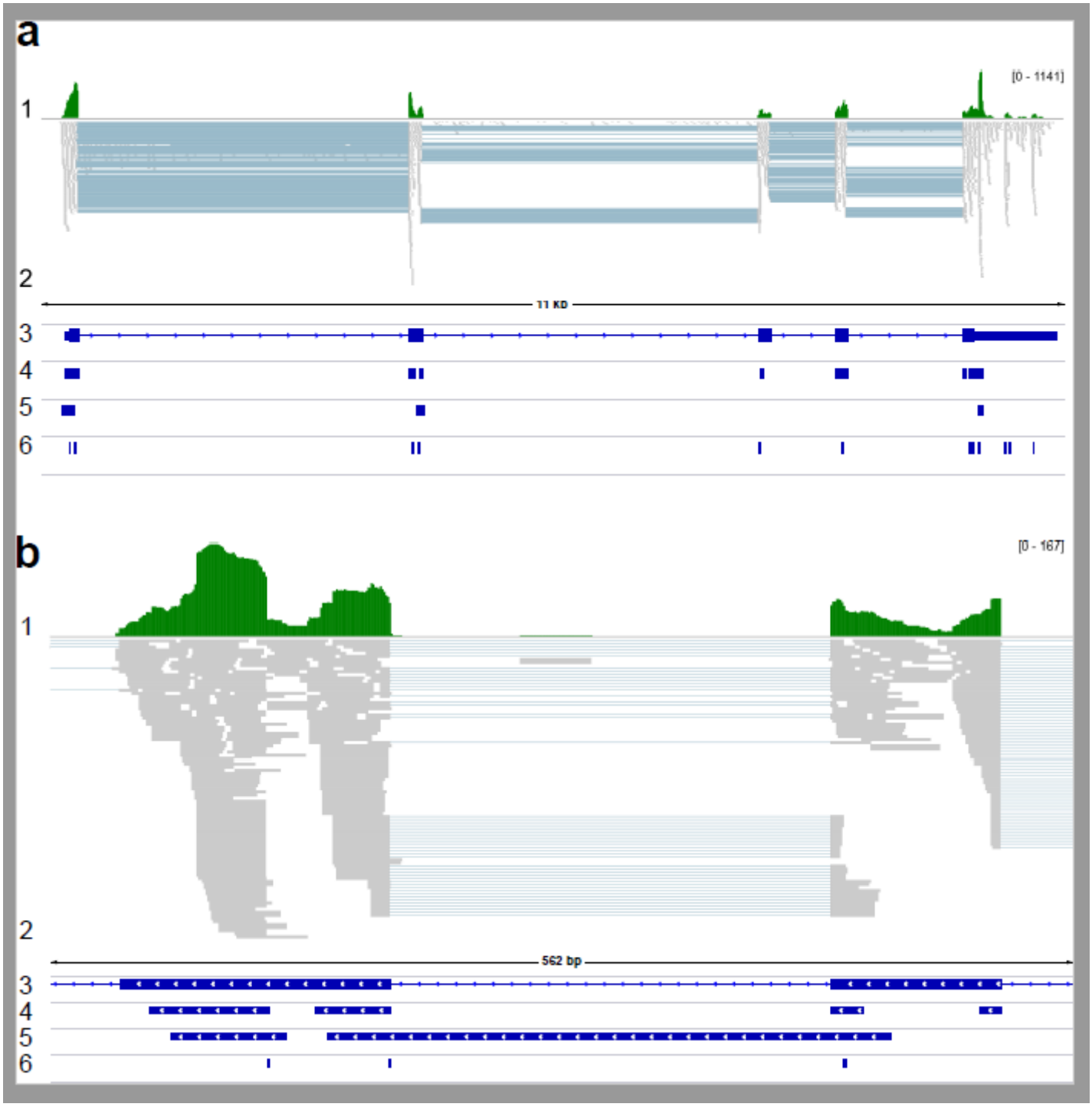

Figure 1

IGV snapshot of two genomic regions with mapped YBX3 K562 eCLIP data. 1: read profile (coverage), 2: read alignments, 3: gene annotations (thick blue regions are exons, thin blue regions introns), 4: CLIPper peaks, 5: PEAKachu peaks, 6: PureCLIP peaks. For clarity only gene strand reads from replicate 1 are displayed. a PRDX6 whole gene region (length 11 kb, maximum read coverage 1141). b DDOST gene exons 6 and 7 region (length 562 bp, maximum read coverage 167).

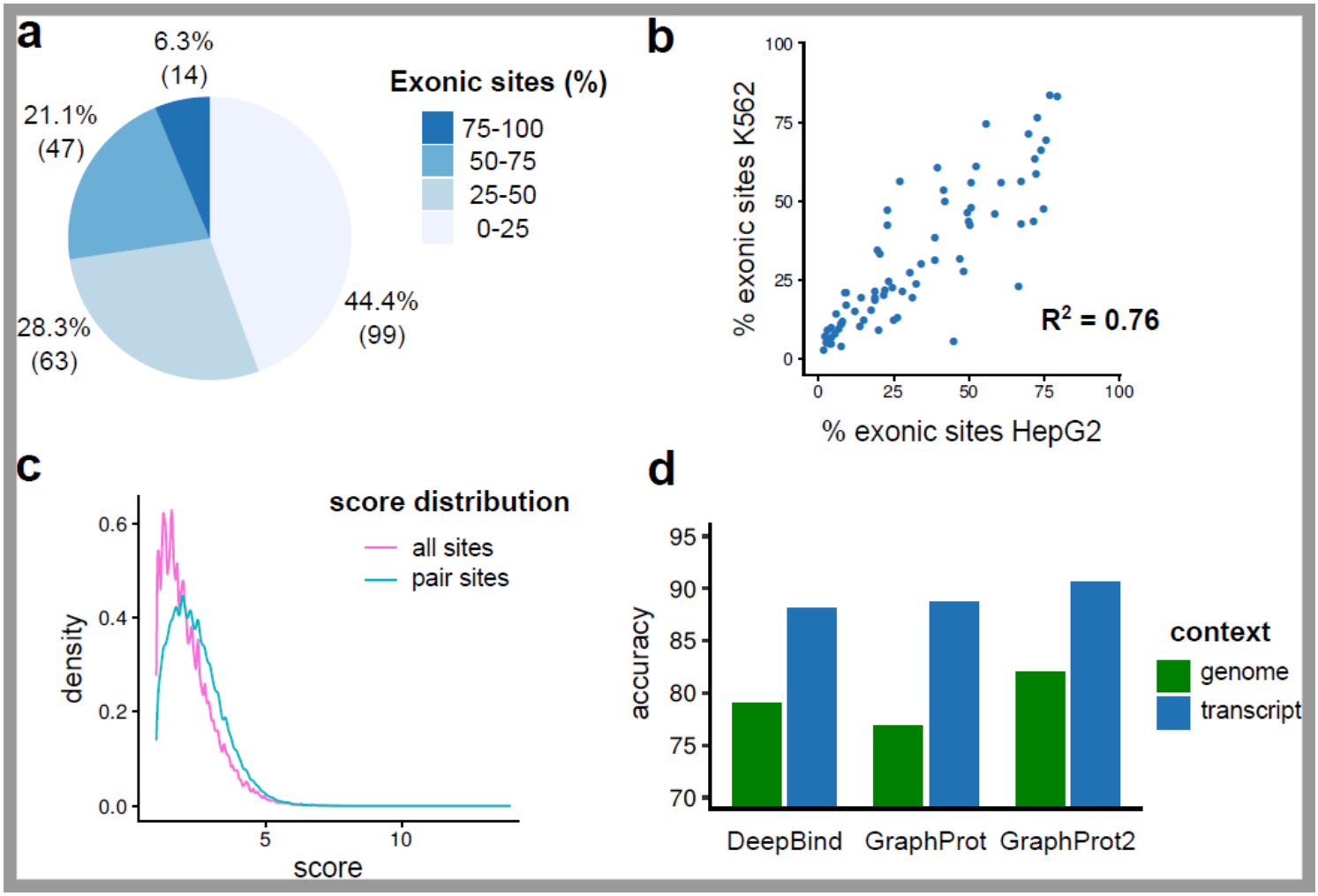

Figure 2

Exon binding statistics of eCLIP datasets and prediction results for different sequence contexts. a Distribution of exonic site ratios for 223 eCLIP datasets over four percentage ranges. For each range, the percentage (number) of sets with ratios falling into this range is given. b Correlation plot of exonic site ratios for RBPs present in two cell lines (HepG2 and K562). c Site score distributions for all exonic sites and exonic sites that form pairs by being located at adjacent exon borders. log2 fold change values of the sites determined by CLIPper were taken as site scores. Only pair sites with a distance of < 10 nt to exon borders were considered. d Average classification accuracies over 5 eCLIP datasets for 3 RBP binding site prediction methods, comparing genome and transcript context.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementarytableS1.xlsx
- SupplementarytablesS2S4.pdf