

Robust identification of Ptbp1-dependent splicing events by a junction-centric approach in Xenopus laevis

Maud Noiret, Agnès Méreau, Gaelle Angrand, Marion Bervas, Carole Gautier-Courteille, Vincent Legagneux, Stéphane Deschamps, Hubert Lerivray, Justine Viet, Serge Hardy, et al.

▶ **To cite this version:**

Maud Noiret, Agnès Méreau, Gaelle Angrand, Marion Bervas, Carole Gautier-Courteille, et al.. Robust identification of Ptbp1-dependent splicing events by a junction-centric approach in Xenopus laevis. Developmental Biology, 2017, 426 (2), pp.449-459. 10.1016/j.ydbio.2016.08.021 . hal-01533247

HAL Id: hal-01533247

https://univ-rennes.hal.science/hal-01533247

Submitted on 6 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ACCEPTED MANUSCRIPT

Robust identification of Ptbp1-dependent splicing events by a junction-centric approach in *Xenopus laevis*

Maud Noiret[a,b], Agnès Méreau[a,b], Gaëlle Angrand[a,b], Marion Bervas[a,b], Carole Gautier-Courteille[a,b], Vincent Legagneux[a,b], Stéphane Deschamps, Hubert Lerivray[a,b], Justine Viet[a,b], Serge Hardy[a,b], Luc Paillard[a,b], Yann Audic[a,b*]

[a]Université de Rennes 1, Université Européenne de Bretagne, Biosit, Rennes 35000, France

[b]Centre National de la Recherche Scientifique UMR 6290, Institut de Génétique et Développement de Rennes, Rennes 35000, France

*Corresponding author. Yann Audic. Tel.: +33 22323 4475; fax: +33 22323 4478. e-mail yann.audic@univ-rennes1.fr

Accepted manuscript

ACCEPTED MANUSCRIPT

Abstract

Regulation of alternative splicing is an important process for cell differentiation and development. Down-regulation of Ptbp1, a regulatory RNA-binding protein, leads to developmental skin defects in *Xenopus laevis*. To identify Ptbp1-dependent splicing events potentially related to the phenotype, we conducted RNAseq experiments following Ptbp1 depletion. We systematically compared exon-centric and junction-centric approaches to detect differential splicing events. We showed that the junction-centric approach performs far better than the exon-centric approach in *Xenopus laevis*. We carried out the same comparisons using simulated data in human, which led us to propose that the better performances of the junction-centric approach in *Xenopus laevis* essentially relies on an incomplete exonic annotation associated with a correct transcription unit annotation. We assessed the capacity of the exon-centric and junction-centric approaches to retrieve known and to discover new Ptbp1-dependent splicing events. Notably, the junction-centric approach identified Ptbp1-controlled exons in *agfg1*, *itga6*, *actn4*, and *tpm4* mRNAs, which were independently confirmed. We conclude that the junction-centric approach allows for a more complete and informative description of splicing events, and we propose that this finding might hold true for other species with incomplete annotations.

Keywords: Differential splicing, skin defects, DEXSeq, genome wide, allotetraploid

Accepted manuscript

ACCEPTED MANUSCRIPT

INTRODUCTION

Alternative splicing is critical in the production of the diversity of proteins that are encoded in the genome. Deep RNA sequencing revealed that almost all (94%) gene products in vertebrates are subject to alternative splicing, thereby dramatically expanding the potential repertoire of available proteins (Pan et al., 2008; Wang et al., 2008). Alternative splicing is controlled in time and space and allows for the tissue-specific production of mRNA isoforms with different coding potential and therefore for the production of proteins with different functions. Tissue-specific regulation of alternative splicing can be achieved either by the tissue-specific expression of regulatory RNA binding proteins (Jensen et al., 2000) or by minor changes in the relative levels of more ubiquitous splicing regulators (Singh and Valcárcel, 2005).

Among splicing regulators, Ptbp1 has been widely studied. However, its importance in animal development has generally precluded study of Ptbp1-dependent splicing events in whole animals. For example, in mice, the constitutive inactivation of *Ptbp1* leads to gastrulation defects with early lethality before stage 12 (Shibayama et al., 2009; Suckale et al., 2011), making the identification of altered splicing events difficult. This issue was partly resolved in mice with a conditional KO model. Conditionally inactivating *Ptbp1* in brain leads to specific phenotypes (Shibasaki et al., 2013), which are probably caused by misregulation of one or several targets of PTBP1 in neural cells. The aberrantly regulated RNAs could probably be identified to explain the brain phenotype of conditionally inactivated mice. However, these approaches remain heavy in mouse.

We use *Xenopus laevis* as a model organism to grasp Ptbp1-dependent regulations in a whole embryo. *ptbp1* is highly expressed in the developing Xenopus epidermis (Noiret et al., 2012). The epithelial specific RNA-binding protein Esrp1 directly up-regulates *ptbp1* expression, explaining the high level of Ptbp1 in epidermis (Méreau et al., 2015). Down-regulating *ptbp1* in Xenopus, either directly (by injecting morpholino antisense oligonucleotides against *ptbp1* mRNA) or indirectly (with morpholino antisense oligonucleotides against *esrp1*) leads to embryonic skin defects with the appearance of blisters developing along the dorsal fin of the embryos (Le Sommer et al., 2005; Méreau et al., 2015). We recently used deep RNA sequencing (RNAseq) to identify molecular events potentially responsible for the skin defects (Noiret et al., 2016). Initially, we had looked for RNAs with different abundances in control embryos and *ptbp1* morphants. Here, we ask if and how differential splicing analyses can be carried out from RNAseq data in *X. laevis*. Indeed, Dichman and colleagues used a combination of the *X.tropicalis* genome and of transcript reconstruction to identify Tra2b-dependent splicing events (Dichmann et al., 2015). We explore how

ACCEPTED MANUSCRIPT

the wealth of data brought by the recently published *X.laevis* genome supports a direct identification of differentially spliced RNAs using only *X.laevis* specific information and annotations.

Presently, deep RNA sequencing reads are in the hundred nucleotides range and are too short for a conclusive reconstruction of full-length mRNA isoforms. Many methods have been developed to characterize alternative splicing events from RNAseq data. The goal of all these methods is to quantify the relative usage of each exon, defined as the abundance of the exon normalized in some way by the abundance of the transcript including the exon. This normalization allows to focus on splicing patterns rather than on transcript levels. To do so, MISO integrates the number of reads aligning to the alternative exon with the number of junctional reads linking it to neighboring exons, with the numbers of junctional reads excluding it, and with the number of reads in the immediately neighboring exons (Katz et al., 2010). SpliceTrap generates an exon-trio database (all the possible 3 consecutive exons in the annotation), generates two isoforms for each trio (with or without the middle exon) and quantifies the relative abundances of the two isoforms to infer middle exon usage (Wu et al., 2011). DEXSeq normalizes each individual exon, or non-redundant exonic part, to all the other exons of the gene, and uses generalized linear models to model read counts (Anders et al., 2012). MATS counts the number of reads mapped to the junctions linking each exon to other exons, and the number of skipping junctions, and uses a Bayesian framework to identify differential splicing (Shen et al., 2012). rMATS is adapted for replicate RNAseq data (Shen et al., 2014).

All the above approaches end up with information about relative exon abundance, even if some of them rely on junctional reads. These "exon-centric" approaches are therefore highly dependent on prior exon identification and annotation, and their power may be weak for genomes that have not yet been as extensively studied as human or mouse. Furthermore, even in a completely sequenced and annotated model, it is still conceivable that a particular pathology or a specific genomic variant allows for the production of an as yet unannotated cryptic exon. Conversely, "junction-centric" approaches focus on the exon-exon junctions and compares junction usage, defined as the number of reads spanning each junction normalized by the abundance of the transcript, between two situations (Kakaradov et al., 2012; Li et al., 2015; Pervouchine et al., 2013).

Here, we used simulated datasets in human and experimental datasets in Xenopus laevis to compare exon-centric and junction-centric approaches. We conclude that the junction-centric approach is significantly more powerful with the current *X.laevis* genome annotation.

Accepted manuscript

ACCEPTED MANUSCRIPT

RESULTS AND DISCUSSION

Comparison of exon-centric and junction-centric approaches on simulated data

We envisioned different approaches to identify differentially spliced genes between two situations (Figure 1A). All these approaches start from sequencing reads in triplicate, a genome assembly and existing gene models (current annotation). We use the STAR mapper (Dobin et al., 2013) and we carry out the mapping in two successive passes as suggested (Kwon, 2015). From the first pass, we obtain a set of newly discovered exon-exon junctions used in at least one condition. We then re-map all the reads in the second pass using both the existing annotation and these newly discovered junctions. In the first approach, we use DEXSeq to identify differential junctions (Diff.junctions), although this package was initially described to infer statistically different exon usage from RNAseq data (Anders et al., 2012). In the second approach, we use the alignment files from STAR with HTSEQ (Anders et al., 2015) and a non-redundant ("flattened" according to HTSEQ terminology) annotation to obtain the number of reads in each exonic regions for each sample, and we again use DEXSeq to identify differential exonic regions (hereafter differential exons (Diff.exon), although they do not necessarily corresponding to *bona fide* exons due to the flattening). Finally, in the third approach, we generate a novel exonic region annotation from the splice junctions detected using a simple rule (see Materials and Methods), and we identify the differential junction-based (Diff. JB exon) exonic regions after counting reads with HTseq and DEXSeq (Figure 1A).

We used recently published sets of simulated data (Soneson et al., 2016) to compare the performances of the three approaches. They consist of six sets of human RNAseq data (two triplicates) where differential splicing has been introduced for a thousand of genes. Because *Xenopus laevis* annotation is not complete, we analyzed how a degraded annotation impacts the performances of the different approaches. Figure 1B illustrates the consequences of this degradation on one hypothetical gene. In the real situation, this gene consists of 5 exons: exons 2 and 3 are mutually exclusive and intron 3 has two alternative 5' splice sites, resulting in exon 3 being split into two non-redundant exonic regions (3a and 3b in Figure 1B, upper panel). The perfect annotation with 100% of the exons (A100 in Figure 1B) includes therefore 6 junctions and 6 exonic regions. If we suppose that depleting 20% of the exons in the original annotation (80% of the annotation is remaining, A80 in Figure 1B) results in loosing exon 1 of this particular gene, then the exonic approach solely based on the existing annotation fails to identify exon 1 while the junctional approach, which integrates newly discovered junctions, succeeds in identifying all the junctions.

ACCEPTED MANUSCRIPT

The JB exonic approach also fails to identify exon 1 because it is unable to set its 5' boundary. Finally, if a further depletion (20% of the annotation remaining, A20 in Figure 1B) additionally results in loosing exon 2 and the information about intron 3 alternative 5' splice sites, then junctions 2 to 6 can be identified (junction 1 is lost because its genomic coordinates fall outside the gene with this annotation). Only exons 3, 4 and 5 are retained by the exon-centric approach. The JB exon-centric approach should also re-discover the existence of alternative 5' splice sites in intron 3, but fail to identify exons 1 and 2 again due to the absence of 5' boundary.

We assessed the performances of the different approaches by taking them as binary classifiers, aimed at classifying each exon or junction as differential or non-differential. Evaluating the performance of a binary classifier relies on 4 different data, the numbers of true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN). Using simulated rather than experimental data allows to perfectly identify these sets of genes in different situations. We calculated the Matthews correlation coefficient (MCC) at $p<0.05$ for the exon-centric and the junction-centric approaches with the different annotation qualities (Figure 1C). The highest MCC (0.661) was achieved with the exonic approach and the perfect annotation (E_100), but the MCC of the junctional approach with the same annotation was in close proximity (0.641, J_100). As expected, degrading the annotation reduced the MCC of both approaches (Figure 1C). We also plotted ROC curves to confirm these findings (Figure 1D). In ROC curves, a completely random classifier would give a diagonal line, and the distance from the diagonal line measures the classifier's performance. Figure 1D shows that the performances of the junction-centric and exon-centric approaches are very similar with a low-quality annotation, and that better annotations improve the performances of both approaches with a markedly bigger effect on the exonic approach.

A caveat of the above analyses is that depleting the annotation of exons modifies not only the available exons, but also the definition of the transcriptional units (the gene coordinates GC). For the same hypothetical gene as above, making the correct gene coordinates available for the analysis allows an efficient discovery of junctions and exons even in the absence of previous annotation (Figure 1B, lower panel, A0+GC). Strikingly, the discriminative power of this approach was comparable to the J_100 or E_80 approaches (MCC=0.644, Figure 1C, overlapping ROC curves, Figure 1D). This suggests that a major parameter affecting the quality of the junction-centric analysis is the definition of the gene boundaries. We compared therefore directly the quality of the junction-centric approaches at different annotation qualities, using either the gene coordinates

Accepted manuscript

ACCEPTED MANUSCRIPT

derived from the exonic annotation as above, or using the best gene coordinates independently of the exonic annotation ("GC_J" panels in Figures 1B-F). Please note that by construction the GC_J_100 sample is strictly the same as the J_100 panel. Figures 1C and 1E clearly show that the junction-centric analysis is almost independent of the annotation used at the mapping step if the gene coordinates are good. This demonstrates the robustness of the junction-centric analysis to variations of the exonic annotation as far as correct gene boundaries are provided.

Finally, we compared the junction-centric analysis enriched with the gene coordinates with the junction-based exon-centric analysis (see Figures 1A-B). The JB exon-centric analysis is independent of the existing annotation when gene coordinates are provided both looking at MCCs (0.655 and 0.650 with complete or null exonic annotation, respectively, Figure 1C) and ROC curves (Figure 1F). It is also very similar to the junction-centric analysis. Altogether, these data show that the best results are achieved with an exon-centric approach for perfectly annotated genomes. However, the junction-centric approach only requests correct gene starts and ends and is expected to be more suited to model organisms with incomplete genome annotation like *Xenopus laevis*. Finally, the junction-based exonic approach behaves the same as the junction-centric approach. Because it is slightly more complicated as it requests exonic regions to be reconstructed from the junctional reads, we won't consider this approach in the following analyses.

RNAseq of *ptbp1* morphants and control *Xenopus* embryos

We have described the construction and deep sequencing of 6 libraries from pools of non-injected *X.laevis* embryos and *ptbp1* morphants (Noiret et al., 2016) (Figure 2A). We have obtained on average 56 millions of reads per condition (Figure 2B). We mapped them is two successive passes using the STAR mapper (Dobin et al., 2013) and the most recent genome assembly (v9.1) as shown in Figure 1A. With this procedure, 93 % of the reads on average were mapped (Figure 2B), to be compared with 85 % attained with TopHat2 and the v7.1 genome assembly (Noiret et al., 2016). We identified differential junctions and exons as above. Furthermore, to better understand to what extent the quality of annotation impacts the results with experimental rather than simulated data, we generated a third set of data termed annotation-supported (An-S) junctions, where only the junctions linking two annotated exons were retained. Among them, the differential An-S junctions were identified with DEXSeq as above. It is expected that differences between the exon set and the An-S junction set arise from the different counting schemes (counting exons or junctions), while differences between the An-S junction set and the all-junction set arise from annotation

Accepted manuscript

ACCEPTED MANUSCRIPT

improvement during the first pass of mapping.

Alternative splicing events identified by the exon-centric and the junction-centric approaches

We identified 493 genes with at least one differential exon (p<0.05, adjusted for multiple tests) (Table ST1), 1275 genes with at least one differential junction (Table ST2), and 920 genes with at least one annotation-supported differential junction (Table ST3). The Venn diagram in Figure 2C shows the overlap between these sets of genes. As expected, the An-S junction dataset is essentially included in the all-junction dataset. Only a very small number of genes (4+2) are present in the An-S junction dataset but not the all-junction dataset, owing to the slightly different size factor and dispersion estimate in DEXSeq. Three-hundred-and-four (284+20) genes are at the overlap between the exon group and the all-junctions group, revealing that almost two-thirds (304/493) of the genes with at least one differential exon also have at least one differential junction. Most of these genes (93%, 284/304) have at least one differential An-S junction, revealing that when a gene includes differential exons and junctions, the differential junctions are generally annotation-supported. Starting from the 284 genes identified in all three approaches, the An-S junction approach enriches the repertoire of potential splicing events controlled by Ptbp1 by 224% (636 (630+2+4)/284) while the exon-centric approach enriches it by only 73% (209 (185+20+4)/284). Hence, with experimental data in *Xenopus laevis*, counting junctions rather than exons increases the number of detected events. Compared with the An-S junction approach, using newly discovered junctions next allows to identify 361 new splicing events, which is an additional 39% (361 (341+20)/920) increase.

We next asked, for the genes with both differential exons and junctions, if the differential exons and junctions are topologically linked to the same splicing event. Because more than one differential event can be detected in each gene, these 304 genes with at least one differential exon and one differential junction correspond to 562 exons and 836 junctions (Figure 2D). Three fourths (421/562) of the differential exons are topologically associated with a differential junction, where "topologically associated" means an overlap of the genomic coordinates of the exons and junctions. Similarly, almost 81% (676/836) of the differential junctions are "topologically associated" with a differential exon (Figure 2D). We conclude that, for the genes with both differential junctions and differential exons, the exon-centric and junction-centric approaches are largely in agreement to identify and characterize the same differential events.

Accepted manuscript

ACCEPTED MANUSCRIPT

We finally conducted a GO-term enrichment analysis for the genes with at least one differential junction or exon (Figure 2E). Most terms are shared, with similar *p*-values for enrichment. Furthermore, these common terms are generally linked to general processes like cell shape or differentiation, cell adhesion, signaling pathways or transport. Finding these processes enriched is not surprising given the phenotype of *ptbp1* morphants, with the appearance of dorsal blisters revealing defective structure or adhesion of the epidermis cells to their substrate (Noiret et al., 2016). Taken together, these data show that a large part of the differential splicing events revealed by the exon-centric approach are also identified by the junction-centric approaches, but that the junction-centric approaches identify numerous additional splicing events.

Compared performances of the exon-centric and the junction-centric approaches in *Xenopus laevis*

We used the same indicators as above (MCC at $p<0.05$ and ROC curves) to assess the performances of the exon-centric and the junction-centric approaches. A major issue with experimental data, compared with simulated data, is that we do not know *a priori* which exons or junctions are differentially used, making it impossible to sort between "true" and "false" positives or negatives. To overcome this difficulty, we hypothesized a strong conservation of the Ptbp1-mediated post-transcriptional networks in vertebrates. Llorian *et al* identified splicing events in 210 genes controlled by PTBP1 in HeLa cells (Llorian et al., 2010). We could identify the *Xenopus* orthologues of 114 out of these 210 genes based on gene names (Figure 3A, Table ST4). We classified these 114 genes as TP if they have at least one differential exon or junction in our analyses in Xenopus, and as FN otherwise. Similarly, we randomly picked 114 Xenopus genes whose human orthologues are not controlled by PTBP1, and we classified them as FP if they have at least one differential exon or junction in our analyses, and as TN otherwise. The numbers of FP and TN that are given below are the means of 100 replicates (Figure 3A). Importantly, the conservation of post-transcriptional networks between human and Xenopus is probably not complete and the gene expression program is also different between HeLa cells and Xenopus embryos. These differences certainly result in an under-estimation of the absolute performances of the exon-centric and junction-centric approaches, but still allows relative comparisons between the different approaches to be made.

We first counted the numbers of TP, FN, FP and TN setting the *p*-value thresholds at 0.05 (Figure 3B). Counting junctions rather than exons more than doubles the number of TP (27 instead of 12).

ACCEPTED MANUSCRIPT

Allowing for junction reannotation again improves TP identification (35 TP with the all-junctions method). However, using the junction-centric approaches also increases the numbers of FP. The MCC takes into account true and false positives and negatives. The MCC of the junction-centric approaches are above that of the exon-centric approach, with a marginal advantage of the all-junction approach over the An-S-junction approach (Figure 3B). The ROC curves shown in Figure 3C confirm that the performances of the An-S junctions and the all-junctions approaches are very similar, and better than the performances of the exon approach. Together, these data show that counting junctions rather than exons increases the number of detected events with better performance, and that taking into account newly discovered junctions further expands the repertoire of detected splicing events without any performance loss.

Analysing the genes found in the exon-centric approach only

While the above data globally demonstrate better performances of the junction-centric approaches compared with the exon-centric approach, 185 genes have differential exons without any detected differential junction (Figure 2C). We visually inspected these genes, which contain 208 differential exons (Table ST5). We sorted them in 5 classes (RI, retained intron; 3CPA, 3' Cleavage and PolyAdenylation; SJS, Supported by a Junction in Sashimi plot; VWE, Very Weakly Expressed; NSJS, Non-Supported by a Junction in Sashimi), and their distribution between the classes is shown in Figure 4A. The 3CPA class corresponds to the distal region of tandem cleavage/polyadenylation sites (Figure 4B). Any modification in the ratio of proximal to distal cleavage site usage results in modifying the number of reads in the distal region. This can be detected by the exon-centric approach, but is undetectable in the junction-centric in the absence of quantifiable junction. We can therefore suppose that 3CPA class corresponds to real differential exons that intrinsically fall out of reach of junction-centric detection. The SJS class also probably groups a majority of truly differential exons. In this class, there exist junctions skipping the exon identified as differential, but these junctions were not identified as differential themselves (Figure 4C). Conversely, in the NSJS class, no skipping junction was detected in any of the 6 analyzed samples (Figure 4D). This implies that the exons identified as differential are very probably constitutive exons. We assume that the NSJS class results from exon-centric approach background, while the SJS class corresponds to splicing patterns really differing between controls and morphants. The last class of exons (VWE) corresponds to weakly expressed genes (Figure 4E). Since the exonic counts are spread on whole exons, the average number of reads assigned to exons is higher than the average number of reads assigned to junction. This might allow the difference between the two conditions to attain statistical significance with the exon-centric, but not the junction-centric approach. Indeed, the number of

Accepted manuscript

ACCEPTED MANUSCRIPT

reads per gene (a proxy for gene expression level) is lower for genes with only one or more differential exons than for genes with both differential exons and junctions (Figure 4F, p=1.1x10^{-5}, Student's t-test). This is consistent with the exon-centric approach being more sensitive than the junction-centric approaches for weakly expressed genes. Hence, we can conservatively assume that the VWE,, 3CPA and SJS classes correspond to real differential splicing patterns that are missed in the junction-centric approach. However, a large majority (the NSJS class, 145/208) of the events identified only by the exon-centric approach are not reliable.

Experimental validation of the splicing events identified by the exon-centric and junction-centric approaches

While the above data show the superiority of the junction-centric over the exon-centric approach, we wanted to validate these results by other experimental approaches on a limited number of Ptbp1-controlled splicing events. We therefore examined some particular genes previously known to be controlled by Ptbp1 in *Xenopus* embryos (Figure 5), and we tested some of the newly discovered Ptbp1-controlled splicing events (Figure 6). Owing to alternative maturation, *tpm1* mRNA has three alternative 3' terminations (Figure 5A). Constitutive exon 8 is either spliced to exon 9A9' in muscular cells, or to exon 9D in non-muscle cells ("O5" isoform). Exon 9A9' behaves either as a terminal exon ("a7" isoform) or an internal exon spliced to exon 9B ("a2" isoform). Ptbp1 represses exon 9A9' usage and favours the non-muscular isoform O5 at the expense of the two other isoforms (Hamon et al., 2004; Le Sommer et al., 2005). Figure 5A, right panel, gives the *tpm1a* exons or junctions identified as differential in RNAseq experiments between controls and *ptbp1* morphants. Since *Xenopus laevis* is allotetraploid, each gene exists as two homeologs or pseudo-alleles that have the same structure and function and are located on homeologous chromosomes. We give the results for both pseudo-alleles of *tpm1* (*tmp1l* and *tpm1s*). Exons 9D, and 9A9' plus 9B, are missing in *tmp1l* and *tmp1s* annotations, respectively. The exon-centric and An-S junction-centric approaches are fully consistent with each other and with the existing annotation. We detected the stimulation of exon 9B and of junction 9A-9B in *tmp1l* and the repression of exon 9D and of junction 9A-9D in *tpm1s*. Taking into account newly discovered junctions in the "all-junctions" approach completed this picture, with the detection of both differential junctions in both pseudo-alleles (Figure 5A). We think that the junction-centric approach failed to reveal the stimulation of 8-9A9' because it is already the predominant event in control embryos. Nevertheless, these results indicate the all-junction-centric approach describes the changes to *tmp1* RNA maturation in Ptbp1-depleted embryos more precisely than the other approaches and in both pseudo-alleles.

Accepted manuscript

ACCEPTED MANUSCRIPT

We carried out the same comparison with three other mRNAs. Ptbp1 controls the maturation of its own pre-mRNA by repressing exon 11 inclusion, in human cells (Wollerton et al., 2004) and *Xenopus* embryos (Méreau et al., 2015). This represents the basis of a negative feedback loop controlling *ptbp1* expression, since the isoform devoid of exon 11 is targeted to rapid degradation by nonsense mediated decay. Despite correct annotation for both pseudo-alleles, the exon-centric approach detected the stimulation of exon 11 in *ptbp1l* only, while the junction-centric approach revealed a repression of the 10-12 junction for both pseudo-alleles (Figure 5B). Ptbp1 similarly controls the inclusion of *ptbp2* exon 10 (Méreau et al., 2015; Spellman et al., 2007). Despite the correct annotation of both pseudo allele in this region, we detected a change to *ptbp2* mRNA splicing in *ptbp1* morphants only using the junction-centric approach (Figure 5C). Finally, *actn1* pre-mRNA includes two exons named NM and SM. It was shown previously that the depletion of Ptbp1 favoured the SM exon at the expense of the NM exon, and that some maturation products with both exons were also detected in the absence of Ptbp1 (Le Sommer et al., 2005). The exon-centric approach failed to detect any change in splicing pattern of *actn1*, and the An-S junction approach only detected reduced NM-EF2 in *actn1s*, at least in part due to the missing annotation of exon SM. Conversely, the all-junction-centric approach revealed the stimulation of the junctions that include exon SM in both pseudo-alleles (Figure 5D). This analysis of 8 genes (4 pairs of pseudo-alleles) previously known to be regulated by Ptbp1 in *Xenopus* embryos is consistent with the hierarchy set by the above performance comparisons: the exon-centric approach only retrieved 3 out of 8 genes, while the An-S junction approach retrieved 7 and the all-junction approach retrieved them all.

The all-junction approach revealed one (both *ptbp1* and *ptbp2* pseudo-alleles), two (both *tpm1* pseudo-alleles, *actn1l*), or three (*actn1s*) differential junctions in each gene (Figure 5A-D). We systematically counted the numbers of differential junctions retrieved within the genes containing at least one differential junction. Figure 5E (left part) shows that about two-thirds (819/1275 genes) of them contain only one differential junction. Forty-five genes contained at least 5 differential junctions, the top gene being Xelaev18037026m.g_kif20b-like.S with 19 differential junctions. When looking at genes harboring at least 2 differential junctions (456 genes and 1336 differential junctions), about 2/3 (866/1339) of the differential junctions are supported by at least one other differential junction (on the basis of overlapping junction coordinate) Figure 5E (right part).

We next performed RT-PCR experiments on some of the newly discovered Ptbp1-controlled splicing events. We injected *Xenopus* embyos with control (ctrl) or *ptbp1* morpholinos as in Figure

ACCEPTED MANUSCRIPT

1A, and we allowed them to develop before RNA extraction. We first analysed *agfg1* (Figure 6A) and *itga6* (Figure 6B) as examples of cassette exon-containing mRNAs. The junctions involving the exon labeled B in Figure 6A were found to be differentially used between controls and *ptbp1* morphants. Specifically, the usages of the junctions between exons A and B, and between exons B and C, were increased, whereas the usage of the junction between exons A and C was reduced. This indicates a stimulation of exon B inclusion in *ptbp1* morphants, but the exon-centric approach failed to detect exon B up-regulation despite correct annotation (Figure 6A, upper panel). We carried out RT-PCR experiments to test if exon B is more frequently included in *ptbp1* morphants. The isoform containing exon B was much more abundant in *ptbp1* MO-injected embryos than in control embryos (Figure 6A, lower panel, compare lane 3 to 1-2). Co-injecting a *ptbp1* mRNA resistant to the Morpholino inhibition together with the MO decreased the amount of exon B-containing isoform (lane 4), while the mRNA alone had no effect (lane 5). This rescue experiment confirms the specificity of the *ptbp1* MO. Both the exon-centric and the junction-centric approaches revealed that *itga6* exon 6A was skipped in *ptbp1* morphants (Figure 6B, upper panel). RT-PCR experiments confirmed the specific repression of exon 6A in *ptbp1* morphants (lower panel) and its partial restoration by the co-injection of an immune *ptbp1* mRNA together with the MO. Hence, the junction-centric analysis discovered novel cassette exons controlled by Ptbp1 in *agfg1* and *itga6* RNAs, which were confirmed by independent RT-PCR experiments.

We next analyzed other splicing events found to be potentially regulated by Ptbp1 in the junction-centric approach. *actn4* pre-mRNA contains a set of two mutually exclusive exons labeled C and D in Figure 6C, upper panel. The junction-centric approach identified 5 differential junctions: the B-C and C-E junctions were reduced, whereas the B-D, C-D and D-E junctions were increased (Figure 6C, upper panel). However the C-D junction was supported by about 30 times less reads than the others and only in the *actn4l* pseudo allele, indicating a minor splicing events. These results suggest that exon D usage is stimulated in *ptbp1* morphants at the expense of exon C, but we only detected exon C repression in the exon-centric approach (Figure 6C, upper panel). We carried out RT-PCR experiments with primers in exons A and E. Because exons C and D have the same size, and a *Sac1* restriction site lies within exon C, we cut the amplimeres with *Sac1* before gel loading. We found that the *actn4* amplimeres obtained from *ptbp1* morphants were predominantly *Sac1*-resistant (lane 3), revealing exon D inclusion, whereas the *actn4* amplimeres obtained from non-injected, control MO-injected embryos (lanes 1 and 2), or rescue RNA alone injected embryos (lane 5) were almost fully cleaved by *Sac1*, revealing exon C inclusion. The situation was intermediate when the embryos were co-injected with *ptbp1* MO and RNA (lane 4). Hence, in accordance with the

Accepted manuscript

ACCEPTED MANUSCRIPT

junction-centric analysis, Ptbp1 favors the exon C-containing isoform. We finally chose *tpm4* mRNA as an example of alternative terminal exon-containing mRNA. This mRNA contains two alternative terminal exons (E1 and E2), and the junction-centric analysis suggested an increase of terminal exon E1 usage and a decrease of terminal exon E2 usage in *ptbp1* morphants (Figure 6D, upper panel). The exon-centric approach only revealed the down-regulation of exon E2. In RT-PCR, the amount of mRNA containing terminal exon E2 was low in *ptbp1* morphants (Figure 6D, lower left panel, lane 3) compared with the other conditions. Conversely, the mRNA containing terminal exon E1 was undetectable except in *ptbp1* morphants and to a lesser extent in embryos co-injected with *ptbp1* MO and RNA. The total amount of *tpm4* mRNA was apparently low in *ptbp1* morphants, which was confirmed by the amplification of constitutive exons A to D (lower right panel). This suggests that terminal exons E1 and E2 confer different stabilities to their respective mRNAs isoforms. Together, these data show that the junction-centric approach allows identifying at least 3 types of alternative splicing events (cassette exon, mutually exclusive exons and alternative terminal exons) with a great accuracy. These events were not or were only partly detected with the exon-centric approach.

Conclusions

We show here that, to detect differential splicing events in *Xenopus laevis*, a junction-centric approach taking into account newly discovered junctions performs much better than the widely used exon-centric approach. It takes advantages of available R statistical packages and softwares, and it has the potential to discover new (pathologic or cryptic) splicing events. Using simulated data, we also have observed that the exon-centric approach was the best one with a perfect annotation, but that its performances decreased rapidly when exon annotation was partially depleted. By contrast, the junction-centric approach was virtually insensitive to any degradation of the annotation as far as correct gene coordinates were provided. These observations reveal that, while the current exon annotation of *Xenopus laevis* genome is far from complete, gene annotation is good enough for a high-performance junction-centric approach. Annotation of transcriptional units can rely on specific methodologies such as CAGE to map transcription start sites (Carninci et al., 1996) , or RNA-PET dedicated to the identification of gene boundaries (Peters and Velculescu, 2005). In addition, expressed sequence tags (EST) provide some hundreds of nucleotides of sequences originating from the 5' or 3' end of expressed mRNAs.Many *X. laevis* and *tropicalis* ESTs are represented in the databases (677911 and 1271480 in deEST release 130101,respectively),which probably contributes to the good quality of transcription unit annotation, hence the superior performances of the junction-

Accepted manuscript

ACCEPTED MANUSCRIPT

centric approach. Therefore, while we think that our finding that a junction-centric approach performs better than an exon-centric approach can be generalized to any model organism with an incomplete annotation, we also think that the *Xenopus laevis* model may be particularly suited for the junction-centric approach.

Our initial aim when we undertook this piece of work was to understand the molecular reasons of the specific phenotype of *ptbp1* morphants, namely the appearance of blisters on the dorsal fin revealing epidermis instability (Noiret et al., 2016). It is therefore highly encouraging to find that many Ptbp1-controlled genes that we discover are linked to cell shape or adhesion. Specifically, *actn4* and *tpm4* encode actinin alpha 4 (Murphy and Young, 2015) and tropomyosin 4 (Gunning et al., 2015), two cytoskeletal proteins that control the actin network. *ITGA6* encodes integrin alpha 6, which plays a critical structural role in the hemidesmosome by its dimerisation with ITGB4, and human *ITGA6* is a causal gene in epidermolysis bullosa with pyloric atresia (Schumann et al., 2013). The inclusion of *ITGA6* exon 6A, which we find here to be Ptbp1-controlled in Xenopus, characterizes the epithelial isoform of ITGA6 (Goel et al., 2014), and the depletion of Ptbp1 in Xenopus switches *itga6* splicing to a mesenchymal isoform. Elucidating whether or not the defective splicing of these mRNAs in *ptbp1* morphants contributes to defective epidermis stability will require further experiments.

MATERIALS AND METHODS

Embryos and Morpholino injection

Sexually mature *Xenopus laevis* females were induced to lay eggs by injection of 500 U hCG (Chorulon, Intervet). Eggs were collected and *in vitro* fertilized with testis lysate as described (Noiret et al., 2016). After dejelling with 2 % cysteine (pH 7.9), we transfered eggs in 1X F1 with 4 % Ficoll. We injected two-cell embryos in both blastomeres with 30 ng of *ptbp1* or control morpholino per blastomere. The development was conducted in 0.1X F1 at 20°C.

Library preparation, mapping and differential analysis of *Xenopus laevis* data

Library preparation and sequencing have been described (Noiret et al., 2016). Briefly, we extracted total RNA from stage 26 embryos, and we prepared an unstranded library with the TRUSEQ mRNA library preparation kit (Illumina) from 1 µg of total RNA for each sample as described in (Noiret et al., 2016). The librairies were sequenced on a Hiseq 2000 for 2x101 bp by the Genoscope (Evry).

Accepted manuscript

ACCEPTED MANUSCRIPT

Reads are deposited at the European Nucleotide Archive (PRJEB8711).

RNAseq data were trimmed and filtered to remove adapters. Mapping was realised in two passes as recommended using STAR v2.4.0 (Dobin et al., 2013) on the *X.laevis* genome (v9.1), and taking advantage of the virtual machine environment developed by the Genouest platform. In the first pass the *X.laevis* annotation (Xlaevisv1.8.Named.gene.gff3) was included after minor modifications (replacement of CDS, 5'UTR, 3'UTR by "exon") and the Star mapper was instructed to allow the discovery of new junctions for all the mapped samples. Mapping results were discarded and new junctions were added as annotation along the *X.laevis* GFF3 to allow for a second pass mapping where uniquely mapped reads were collected. The visualization of the mapped data was conducted using IGV (Thorvaldsdóttir et al., 2013) using the *X.laevis* v9.1 genome as reference along with annotation files.

Differential splicing analysis was conducted in R (R Core Team, 2013). using DEXSeq (Anders et al., 2012). The R scripts for differential splicing and follow-up analyses are available on the authors' lab website (https://igdr.univ-rennes1.fr/en/research/research-groups/luc-paillard-group/gene-expression-and-development-group-publications) For the exon-centric analysis, the Xlaevisv1.8.Named.gene.gff3 annotation was made non-redundant ("flattened") using the python script available with DEXSeq (dexseq_prepare_annotation.py) to generate a gtf file composed of non redundant exonic-parts (DEXSEQ.Xlaevisv1.8.Named.gene.exon_reannotated.gtf). Counts per exonic-parts were generated with HTSEQ (Anders et al., 2015) on this flattened annotation. For the junction-centric analysis the counts per junction directly available from Star mapper was annotated to attribute each junction to a gene based on the Xlaevisv1.8.Named.gene.gff3 annotation in a strand-specific manner. For further analysis, only non-ambiguous junctions that could be attributed to only one gene were conserved. The annotation-supported junctions (An-S-junction) are a subset of the non-ambiguous junctions selected by keeping only the junctions for which both the start and the end of the junctions are framed and contiguous to 2 different exonic part as defined above.

For both exon-centric and junction-centric approaches, the differential analysis comprised the following steps: estimation of size factor, estimation of dispersion, testing for differential exon usage. Junctions or exons were considered differential when the *p*-value (Benjamini-Hoschberg adjusted for multiple testing) was below 0.05. Bed files listing differentially used junction and exon were generated for visualization in IGV. To analyze the performance of our approaches we chose an available positive dataset of genes with splicing events regulated by PTBP1 and experimentally

ACCEPTED MANUSCRIPT

identified in Human HeLa cells (Llorian et al., 2010). Among the 210 human genes with PTBP1-regulated splicing, 114 have identified orthologues in the Xenopus genome. These 114 genes are considered as the positive set of PTBP1 regulated genes. To generate a negative dataset we sampled 114 genes from the non-PTBP1 dependent genes that we considered as our negative dataset. We used 100 different subsampling and computed for the combination of positive and negative sets the True Positive (TP), False positive (FP), True Negative (TN) and False Negative (FN) numbers for each of the differential analysis (Exon, All_junctions, An_S junctions). Based on these results ROC curves were constructed and the Matthews correlation coefficient calculated (MCC) for $p < 0.05$. All computation were performed with R.

The GO term enrichment analysis was conducted on the human GO annotation using the TopGO (v.2.20.0) package (Alexa and Rahnenfuhrer, 2010) after converting the Xenopus gene name to human gene name.

Analysis on simulated data

Soneson and colleagues (Soneson et al., 2016) generated simulated fastq files corresponding to 2 x 3 samples where a thousand genes are differentially spliced between two conditions (array express repository E-MTAB-3766). These 1000 protein-coding genes (ENSG) correspond to the positive sets, and we used them to assess the discriminative power of the different approaches with several rates of annotation degradation. The initial annotation (GRCh37.71) was limited to protein-coding genes. The annotation was downsampled by excluding 20%, 80% or all of the exonic annotation present initially. The depletion was conducted on the negative genes on the one hand and on the positive genes on the other hand to have a similar depletion in both groups. Depletion of the annotation was visually assessed in IGV.

We carried out mapping and identification of differential exons and junctions as described above for *Xenopus laevis* data with the following modifications. Firstly, when indicated, we combined the degraded annotation with the full gene coordinate annotation. Secondly, we produced alternative exonic annotations based on the junctions identified following mapping and analysis. The junction-based (JB) exonic regions are defined by taking for each 5' splice site of a junction the closest upstream (gene strand wise) 3' splice and defining this interval as an exonic region. For each 3' splice site the closest 5' splice downstream is selected and this interval also defines an exonic region. Genes boundaries are treated as 3' splice sites for the start of the gene and 5' splice sites for

Accepted manuscript

ACCEPTED MANUSCRIPT

the end of the gene. The JB exonic region annotation is then made non-redundant (flattened) as described above.

Isolation of RNA and analysis by RT-PCR

Total RNA was extracted from embryos as above, and reverse transcriptions were carried out using SuperscriptII reverse transcriptase (Invitrogen). Briefly, 2µg of RNA were used for the RT reactions with random primers. We carried out semi-quantitative PCR from one-twentieth of the RT reactions with exon-specific primers to obtain alternatively spliced products. The forward primer were radiolabeled before the PCR with γ-^{32}P ATP and T4 PNK. DNA was amplified by 25 cycles (94°C for 30s, 55°C for 30s and 72°C for 60s). The amplimeres were resolved on 8 % polyacrylamide gels, and the gels were dried and autoradiographed (Phosphorimager). The primer sequences are *agfg1*, CTCACAATTCTGCCCA and ACTTGGGAAAATTGTCAAACTGAGC; *itga6*, GGTGTACCTTGGTGGATTAT and TACAGCGTGGTATCGTG; *actn4*, CTTTCAATGCCCTTATCCATAGACA and ATCACTAGCCAGCTTTTCATAGTCC; tpm4Afwd, AGAGGAGCGTGCAGAGGTGTC; tpm4Drev CTGCAAATTCAGCCCGGGTTTCAG; tpm4Dfwd GCTGAAACCCGGGCTGAATTT; tpm4E1rev CTACAAGGAGGTCATGTCATTG; tpm4E2rev TGGAACACAGTACAACATGTG; *eef1a1*, GAGAGGGAAGCTGCTGAGATGG and CCACAGGGAGATGTCAATGGTA.

Acknowledgments

The authors wish to thank the many laboratory members who contributed to the genomic data and annotations that have been made available to the community through Xenbase (http://www.xenbase.org). Mapping and initial data treatments were realized on the computing infrastructure from the Genouest platform (http://www.genouest.org/). Maud Noiret was supported by a Ph.D fellowship from the Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche. Sequencing was made possible by a grant from the Genoscope. Thanks to Yann Le Cunff (IGDR) for helpful discussions regarding ROC curves.

References

Accepted manuscript

ACCEPTED MANUSCRIPT

Alexa, A., and Rahnenfuhrer, J. (2010). topGO: Enrichment analysis for Gene Ontology. R package version 2.20.0

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res. *22*, 2008–2017.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., et al. (1996). High-Efficiency Full-Length cDNA Cloning by Biotinylated CAP Trapper. Genomics *37*, 327–336.

Dichmann, D.S., Walentek, P., and Harland, R.M. (2015). The alternative splicing regulator Tra2b is required for somitogenesis and regulates splicing of an inhibitory Wnt11b isoform. Cell Rep *10*, 527–536.

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*, 15–21.

Goel, H.L., Gritsko, T., Pursell, B., Chang, C., Shultz, L.D., Greiner, D.L., Norum, J.H., Toftgard, R., Shaw, L.M., and Mercurio, A.M. (2014). Regulated splicing of the α6 integrin cytoplasmic domain determines the fate of breast cancer stem cells. Cell Rep *7*, 747–761.

Gunning, P.W., Hardeman, E.C., Lappalainen, P., and Mulvihill, D.P. (2015). Tropomyosin - master regulator of actin filament function in the cytoskeleton. J. Cell. Sci. *128*, 2965–2974.

Hamon, S., Le Sommer, C., Mereau, A., Allo, M.-R., and Hardy, S. (2004). Polypyrimidine tract-binding protein is involved in vivo in repression of a composite internal/3' -terminal exon of the Xenopus alpha-tropomyosin Pre-mRNA. J. Biol. Chem. *279*, 22166–22175.

Jensen, K.B., Dredge, B.K., Stefani, G., Zhong, R., Buckanovich, R.J., Okano, H.J., Yang, Y.Y., and Darnell, R.B. (2000). Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. Neuron *25*, 359–371.

Kakaradov, B., Xiong, H.Y., Lee, L.J., Jojic, N., and Frey, B.J. (2012). Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data. BMC Bioinformatics *13 Suppl 6*, S11.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods *7*, 1009–1015.

Kwon, T. (2015). Benchmarking Transcriptome Quantification Methods for Duplicated Genes in Xenopus laevis. Cytogenet. Genome Res. *145*, 253–264.

Le Sommer, C., Lesimple, M., Mereau, A., Menoret, S., Allo, M.-R., and Hardy, S. (2005). PTB regulates the processing of a 3'-terminal exon by repressing both splicing and polyadenylation. Mol. Cell. Biol. *25*, 9595–9607.

Li, Y., Rao, X., Mattox, W.W., Amos, C.I., and Liu, B. (2015). RNA-Seq Analysis of Differential Splice Junction Usage and Intron Retentions by DEXSeq. PLoS ONE *10*, e0136653.

Accepted manuscript

ACCEPTED MANUSCRIPT

Llorian, M., Schwartz, S., Clark, T.A., Hollander, D., Tan, L.-Y., Spellman, R., Gordon, A., Schweitzer, A.C., de la Grange, P., Ast, G., et al. (2010). Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. Nat. Struct. Mol. Biol. *17*, 1114–1123.

Méreau, A., Anquetil, V., Lerivray, H., Viet, J., Schirmer, C., Audic, Y., Legagneux, V., Hardy, S., and Paillard, L. (2015). A posttranscriptional mechanism that controls Ptbp1 abundance in the Xenopus epidermis. Mol. Cell. Biol. *35*, 758–768.

Murphy, A.C.H., and Young, P.W. (2015). The actinin family of actin cross-linking proteins - a genetic perspective. Cell Biosci *5*, 49.

Noiret, M., Audic, Y., and Hardy, S. (2012). Expression analysis of the polypyrimidine tract binding protein (PTBP1) and its paralogs PTBP2 and PTBP3 during Xenopus tropicalis embryogenesis. Int. J. Dev. Biol. *56*, 747–753.

Noiret, M., Mottier, S., Angrand, G., Gautier-Courteille, C., Lerivray, H., Viet, J., Paillard, L., Mereau, A., Hardy, S., and Audic, Y. (2016). Ptbp1 and Exosc9 knockdowns trigger skin stability defects through different pathways. Dev. Biol. *409*, 489–501.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Pervouchine, D.D., Knowles, D.G., and Guigó, R. (2013). Intron-centric estimation of alternative splicing from RNA-seq data. Bioinformatics *29*, 273–274.

Peters, B.A., and Velculescu, V.E. (2005). Transcriptome PETs: A genome's best friends. Nature Methods *2*, 93–94.

R Core Team (2013). R: A language and environment for statistical computing. (Vienna, Austria: the R Foundation for Statistical Computing).

Schumann, H., Kiritsi, D., Pigors, M., Hausser, I., Kohlhase, J., Peters, J., Ott, H., Hyla-Klekot, L., Gacka, E., Sieron, A.L., et al. (2013). Phenotypic spectrum of epidermolysis bullosa associated with α6β4 integrin mutations. Br. J. Dermatol. *169*, 115–124.

Shen, S., Park, J.W., Huang, J., Dittmar, K.A., Lu, Z., Zhou, Q., Carstens, R.P., and Xing, Y. (2012). MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. Nucleic Acids Res. *40*, e61.

Shen, S., Park, J.W., Lu, Z., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U.S.A. *111*, E5593–E5601.

Shibasaki, T., Tokunaga, A., Sakamoto, R., Sagara, H., Noguchi, S., Sasaoka, T., and Yoshida, N. (2013). PTB deficiency causes the loss of adherens junctions in the dorsal telencephalon and leads to lethal hydrocephalus. Cereb. Cortex *23*, 1824–1835.

Shibayama, M., Ohno, S., Osaka, T., Sakamoto, R., Tokunaga, A., Nakatake, Y., Sato, M., and Yoshida, N. (2009). Polypyrimidine tract-binding protein is essential for early mouse development and embryonic stem cell proliferation. FEBS J. *276*, 6658–6668.

Accepted manuscript

ACCEPTED MANUSCRIPT

Singh, R., and Valcárcel, J. (2005). Building specificity with nonspecific RNA-binding proteins. Nat. Struct. Mol. Biol. *12*, 645–653.

Soneson, C., Matthes, K.L., Nowicka, M., Law, C.W., and Robinson, M.D. (2016). Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. Genome Biol. *17*, 12.

Spellman, R., Llorian, M., and Smith, C.W.J. (2007). Crossregulation and Functional Redundancy between the Splicing Regulator PTB and Its Paralogs nPTB and ROD1. Molecular Cell *27*, 420–434.

Suckale, J., Wendling, O., Masjkur, J., Jäger, M., Münster, C., Anastassiadis, K., Stewart, A.F., and Solimena, M. (2011). PTBP1 is required for embryonic development before gastrulation. PLoS ONE *6*, e16992.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinformatics *14*, 178–192.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wollerton, M.C., Gooding, C., Wagner, E.J., Garcia-Blanco, M.A., and Smith, C.W.J. (2004). Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. Mol. Cell *13*, 91–100.

Wu, J., Akerman, M., Sun, S., McCombie, W.R., Krainer, A.R., and Zhang, M.Q. (2011). SpliceTrap: a method to quantify alternative splicing under single cellular conditions. Bioinformatics *27*, 3010–3016.

Accepted manuscript

ACCEPTED MANUSCRIPT

SUPPLEMENTARY MATERIAL

Supplemental Table 1 (.csv).

A table expanding the data shown in Figure 2C. For each differential exon within the 493 genes with at least one differential exon, we show the gene identification number, the exon basemean (mean of the counts across all the samples), the *p*-value (adjusted for multiple testing), the $\log_2$(fold change), the genomic coordinates, and the gene name (may be similar to the gene identification number).

Supplemental Table 2 (.csv).

A table expanding the data shown in Figure 2C. For each differential junction within the 1275 genes with at least one differential junction, we show the unique identifier of the junction, the gene identification number, the junction basemean (mean of the junction counts across all the samples), the *p*-value (adjusted for multiple testing), the $\log_2$(fold change), the genomic coordinates, the gene name (may be similar to the gene identification number), the strand and the number of other differential junctions supporting each differential junction.

Supplemental Table 3 (.csv).

A table expanding the data shown in Figure 2C. For each differential junction within the 920 genes with at least one differential annotation-supported junction, we show the unique identifier of the junction, the gene identification number, the junction basemean (mean of the junction counts across all the samples), the *p*-value (adjusted for multiple testing), the $\log_2$(fold change), the genomic coordinates, the gene name (may be similar to the gene identification number) and the strand.

Supplemental Table 4 (.csv).

A table expanding the data shown in Figure 3A. For the 210 genes with PTBP1-controlled splicing events in HeLa cells (Llorian et al., 2010), we indicate if one orthologue exists in Xenopus laevis annotation (based on identical gene name).

Supplemental Table 5 (.csv).

A table expanding the data shown in Figure 4A. It lists the 185 genes identified in the exon-centric approach, but not the junction-centric approach. The classification (RI, 3CPA, SJS, VWE, NSJS) of each of the 208 differential exons is given. Because some genes have more than one differential exon, the differential exons are numbered from 5' to 3'.

Accepted manuscript

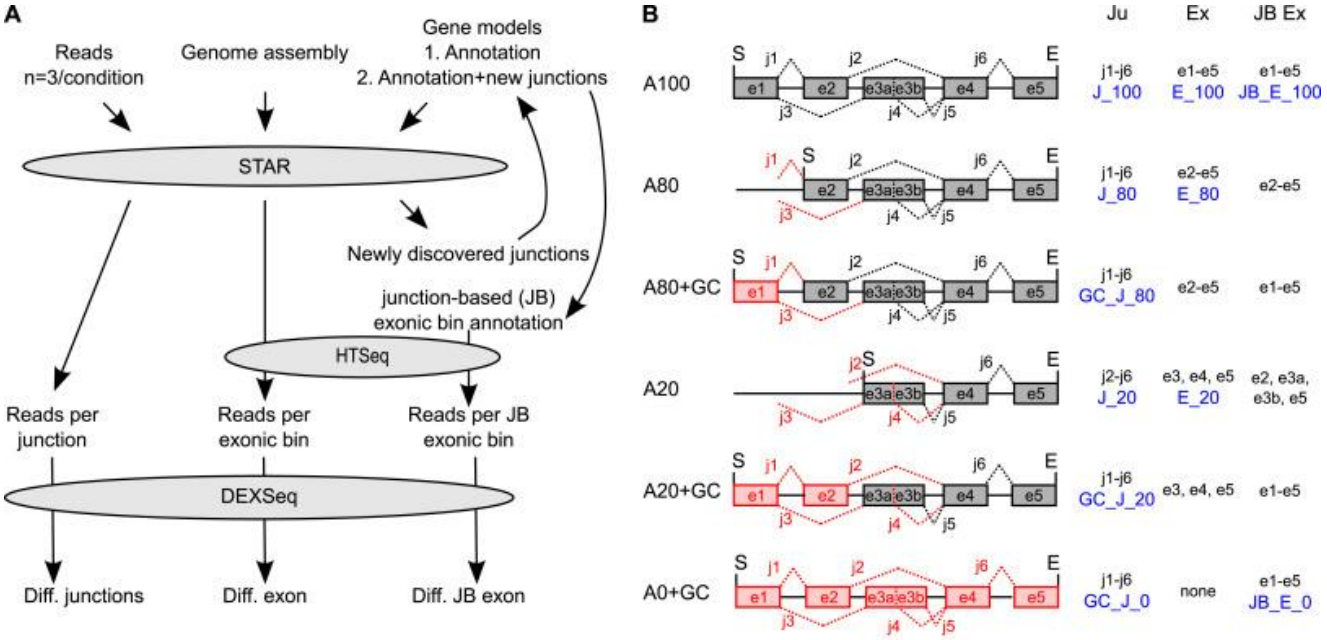
ACCEPTED MANUSCRIPT

Supplemental Table 6 (.csv).

A table expanding the data shown in Figure 6. The exons were given arbitrary names (A to E) in Figure 6, and this table gives the genomic coordinates for each. Exons presented with an * are defined based on the RNAseq data.

Accepted manuscript

ACCEPTED MANUSCRIPT



C

	J_100	J_80	J_20	E_100	E_80	E_20	GC_J_80	GC_J_20	GC_J_0	JB_E_100	JB_E_0
TP	616	490	436	727	622	405	615	612	612	630	627
FN	384	510	564	273	378	595	385	388	388	370	373
FP	261	195	185	420	347	114	254	247	246	253	260
TN	19149	19215	19225	18990	19063	19296	19156	19163	19164	19157	19150
MCC	0.6413	0.5752	0.5359	0.6611	0.6132	0.5472	0.6435	0.6442	0.6446	0.6545	0.6496

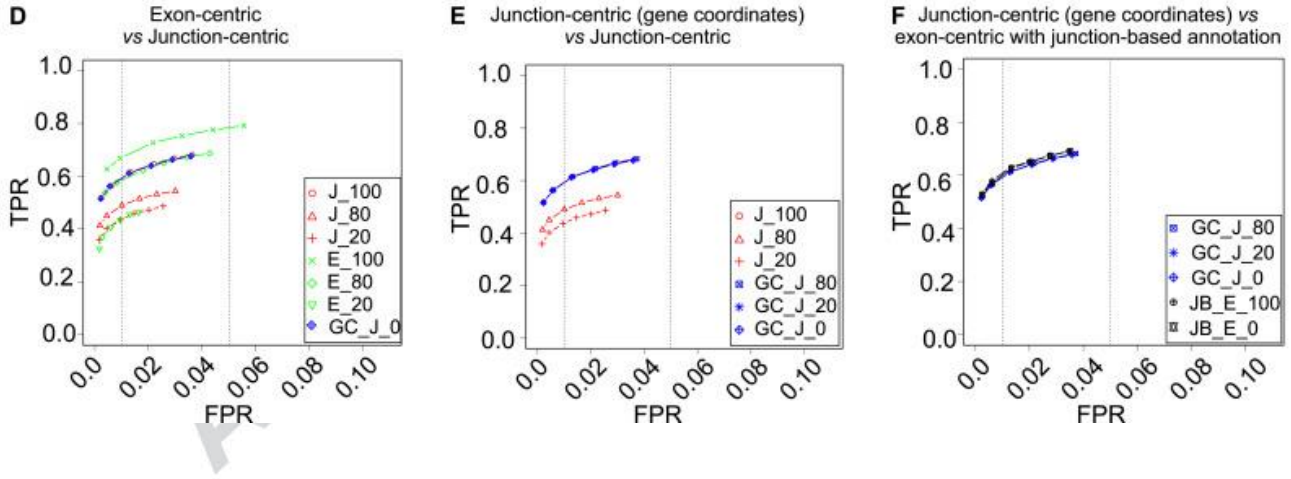
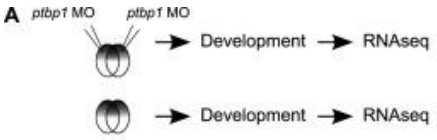
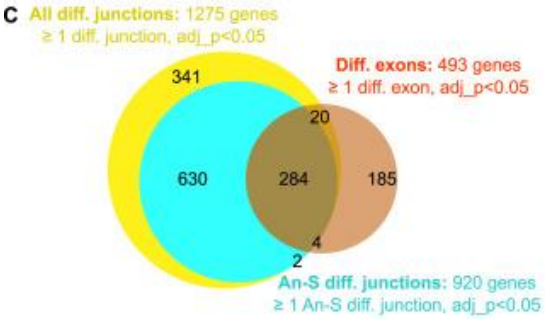


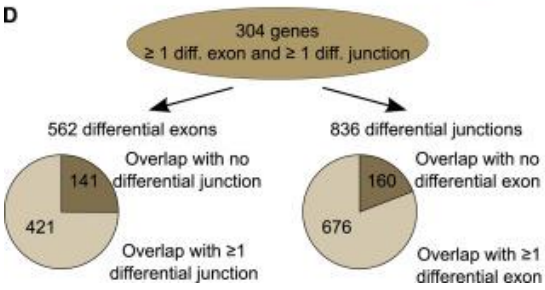
Figure 1. Comparison of exon-centric and junction-centric approaches with simulated data. **A**, Overview of the analysis protocols for the RNAseq data. Starting from 6 RNAseq datasets (2 triplicates), we used the mapper STAR in two successive passes to obtain the numbers of reads per junction for each sample. We also used the sequencing data to reconstruct gene models and generate junction-based (JB) exonic regions annotation. We next used HTSeq to count the number of reads in each exonic region (solely based on the preexisting annotation of using the novel JB annotation). We analysed statistically the data with DEXSeq to identify differential junctions, differential exons, and differential JB exons. **B**, Cartoon illustrating the impact of degrading the annotation on the identification of exons and junctions of a hypothetical gene. A100 corresponds to the perfect annotation and A80, A20 and A0 to a situation where 80%, 20% and 0% of the annotated exons, respectively, are retained. With GC, the correct gene coordinates (start S and end E) are provided in the analysis irrespective of the degree of exonic annotation degradation. The exons and junctions present in the preexisting annotation are in black and those inferred from the sequencing data are in red. **C**, Number of TP, FN, FP and TN, and Matthews correlation coefficient (MCC) in the situations shown in blue in **B**, with an adjusted *p*-value threshold set at 0.05. **D-F**. ROC curves obtained by plotting the true positive rate [TP/(TP+FN)] against the false positive rate [FP/(FP+TN)] for different *p*-value thresholds (0.001, 0.01, 0.05, 0.10, 0.15, 0.20).

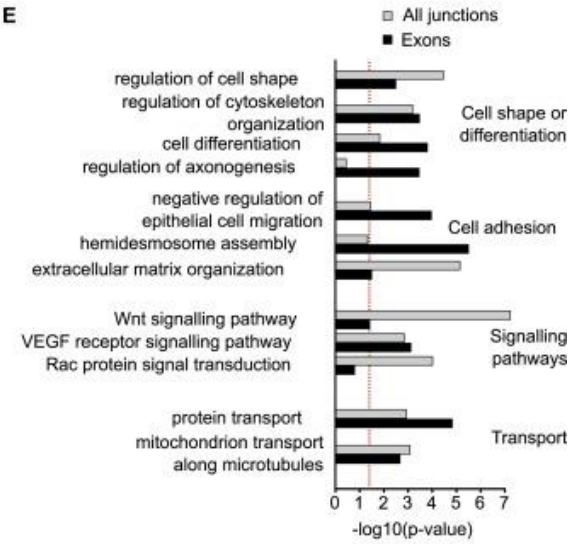


B

Dataset	NI 1	NI 2	NI 3	*ptbp1* MO 1	*ptbp1* MO 2	*ptbp1* MO 3
Reads	47,776,009	59,501,795	55,832,172	50,471,304	59,596,114	60,627,942
% mapped	93.0	93.2	92.6	93.4	93.0	92.6



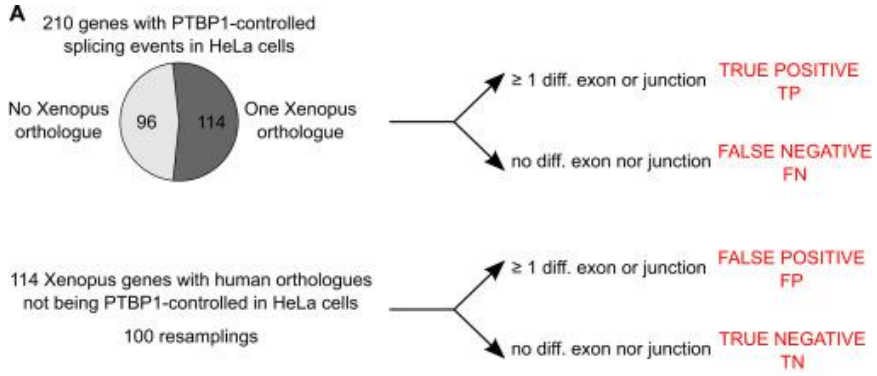




ACCEPTED MANUSCRIPT

Figure 2. Comparison of the exon-centric and junction-centric approaches to identify differential splicing patterns in *Xenopus laevis*. A, RNAseq was carried out from 3 independent pools of stage 28 embryos injected with the morpholino against *ptbp1* mRNA (*ptbp1* MO), or left uninjected. **B**, Table summarizing RNAseq data. "% mapped" is to the percentage of read pairs uniquely mapped to the Xenopus genome (v9.1). **C**, Venn diagram showing the overlap of the genes with at least one differential exon, the genes with at least one annotation-supported differential junction, and the genes with at least one differential junction. **D**, For the 304 (284+20) genes with at least one differential exon and one differential junction, we indicate the numbers of differential exons and junctions. The pie charts show the percentage of differential exons associated with one differential junction (left), and the percentage of differential junctions associated with one differential exon (right). **E**, Comparison of enriched GO terms in differentially spliced genes identified by the junction-centric and the exon-centric approaches. The dotted red line indicates p=0.05. The GO terms are on the left, and we clustered the enriched GO terms in 4 main classes (right).

Accepted manuscript



B $p<0.05$

Approach	Exon	All junctions	An_S junctions
TP	12	35	27
FN	102	79	87
FP	2.7	7.4	5.0
TN	111.3	106.6	109.0
MCC	0.17	0.31	0.28

C ROC curves

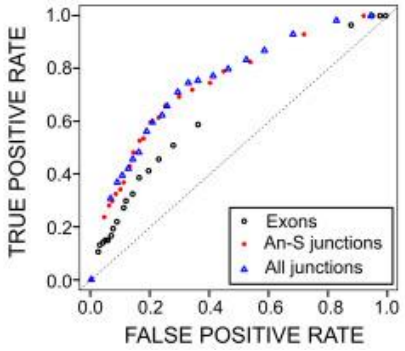


Figure 3. Assessment of exon-centric and junction-centric performances. A, Pie chart showing the number of identified Xenopus orthologues of human genes with splicing events controlled by PTBP1 (Llorian et al., 2010). We classified the 114 Xenopus genes as True Positive (TP) when at least one (Llorian et al., 2010). We classified the 114 Xenopus genes as True Positive (TP) when at least one differential exon or junction was identified in our experiments, and as False Negative (FN) when no differential exon or junction was identified. Conversely, we sampled 100 times 114 Xenopus genes the human orthologues of which are not regulated by PTBP1, and we classified them as False Positive (FP) when we retrieved them in our experiments, and as True Negative (TN) otherwise. The numbers of FP and TN in the following panels are means of the repeated samplings. **B**, Number of TP, FN, FP and TN, and Matthews correlation coefficient (MCC) in the exon-centric, the annotation-supported junctions approach, and the all junctions approach, with an adjusted *p*-value threshold set at 0.05. **C**. ROC curves for different *p*-value thresholds (0.05 steps).

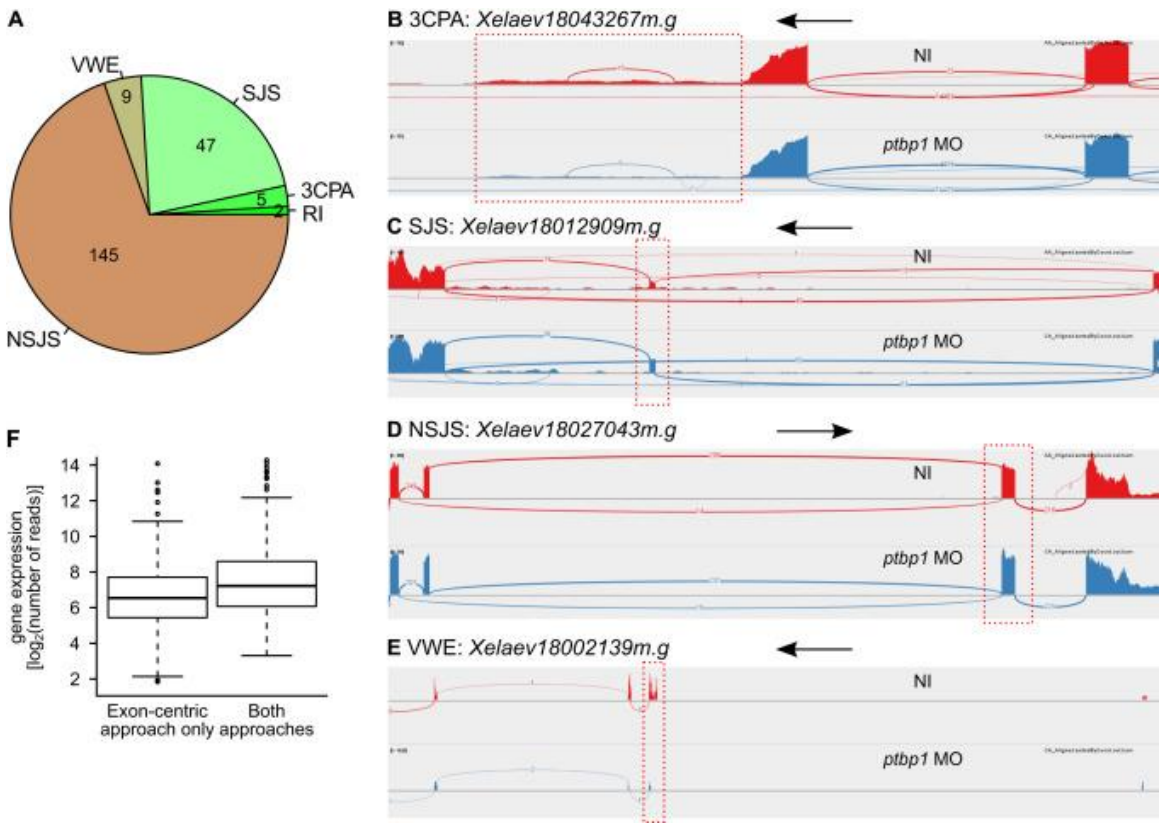
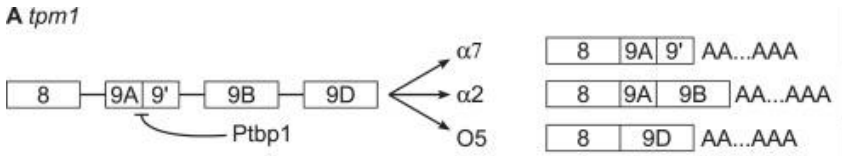
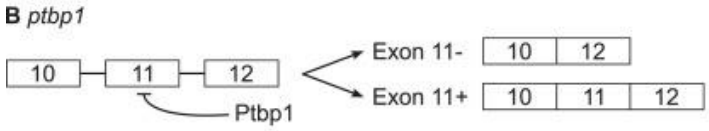

A
VWE
9
SJS
47
3CPA
5
RI
2
145
NSJS
B 3CPA: Xelaev18043267m.g
NI
ptbp1 MO
C SJS: Xelaev18012909m.g
NI
ptbp1 MO
D NSJS: Xelaev18027043m.g
NI
ptbp1 MO
E VWE: Xelaev18002139m.g
NI
ptbp1 MO
F
gene expression
[log2(number of reads)]
14
12
10
8
6
4
2
Exon-centric approach only
Both approaches

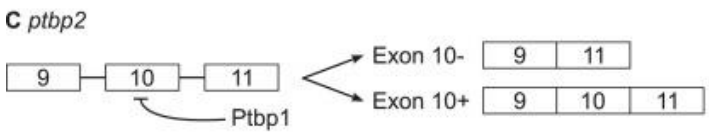
Figure 4. Analysis of the genes only identified by the exon-centric approach. A, We manually classified the 189 genes with one or more differential exon, but without any differential junction, into 5 classes: RI (retained intron), 3CPA (differential exon 3' to a cleavage and polyadenylation site), SJS (supported by a junction in Sashimi plots), VWE (very weakly expressed) and NSJS (non-supported by a junction in Sashimi plots). The pie chart shows the distribution of the genes between these 5 classes. **B-E**, Sashimi plots of representative genes within each of 4 classes (no gene belonging to the RI class is shown since this class only contains two members). The orientations of the gene are given by the arrows. The genomic regions identified as a differential exon are circled dotted red. **F**, Boxplot of the number of reads per gene, for genes identified with the exon-centric approach only and genes identified with both the exon-centric and the all-junctions-centric approaches.



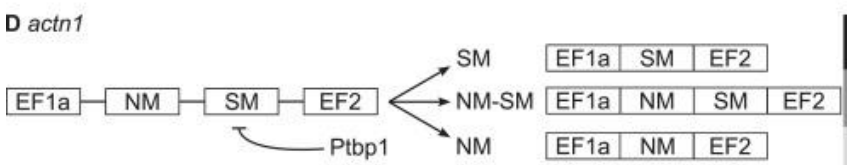
	Exon	All junctions	An_S junctions
tpm1l	9B UP	9A-9B UP 8-9D DOWN	9A-9B UP
tpm1s	9D DOWN	9A-9B UP 8-9D DOWN	8-9D DOWN



	Exon	All junctions	An_S junctions
ptbp1l	11 UP	10-12 DOWN	10-12 DOWN
ptbp1s		10-12 DOWN	10-12 DOWN

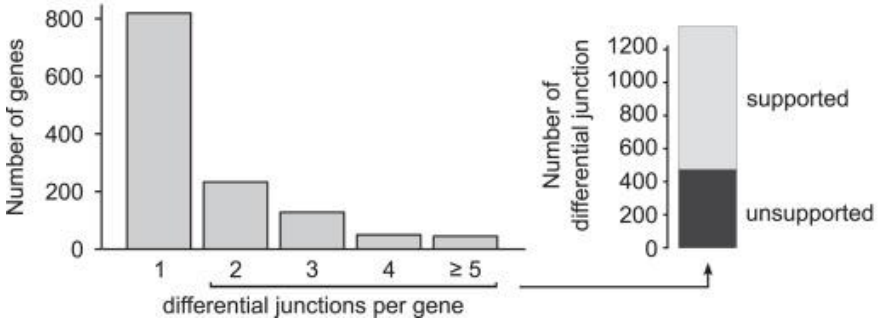


	Exon	All junctions	An_S junctions
ptbp2l		9-11 DOWN	9-11 DOWN
ptbp2s		9-11 DOWN	9-11 DOWN



	Exon	All junctions	An_S junctions
actn1l		NM-SM UP SM-EF2 UP	
actn1s		NM-SM UP SM-EF2 UP NM-EF2 DOWN	NM-EF2 DOWN

E Distribution of the genes per number of differential junctions



ACCEPTED MANUSCRIPT

Figure 5. Analysis of genes known to be regulated by Ptbp1 in Xenopus. A, *tpm1* (Hamon et al., 2004), **B**, *ptbp1* (Méreau et al., 2015), **C**, *ptbp2* (Méreau et al., 2015), **D**, *actn1* (Le Sommer et al., 2005). Left, for each gene, the Ptbp1-repressed exon and the genomic region encompassing it, as well as the different RNA processing patterns, are diagrammed. Introns are represented as horizontal lines and exons as boxes. Right, tables summarizing the exons and junctions identified as differential in *ptbp1* morphants by either of the three approaches. Due to Xenopus laevis tetraploidization, each gene is present as two pseudo-alleles indicated "l" and "s".**E**, Number of genes with the indicated numbers of differential junctions retrieved by the all-junction approach (right part). Number of differential junctions, among the genes harboring at least 2 differential junctions, supported or not by another differential junction.

Accepted manuscript

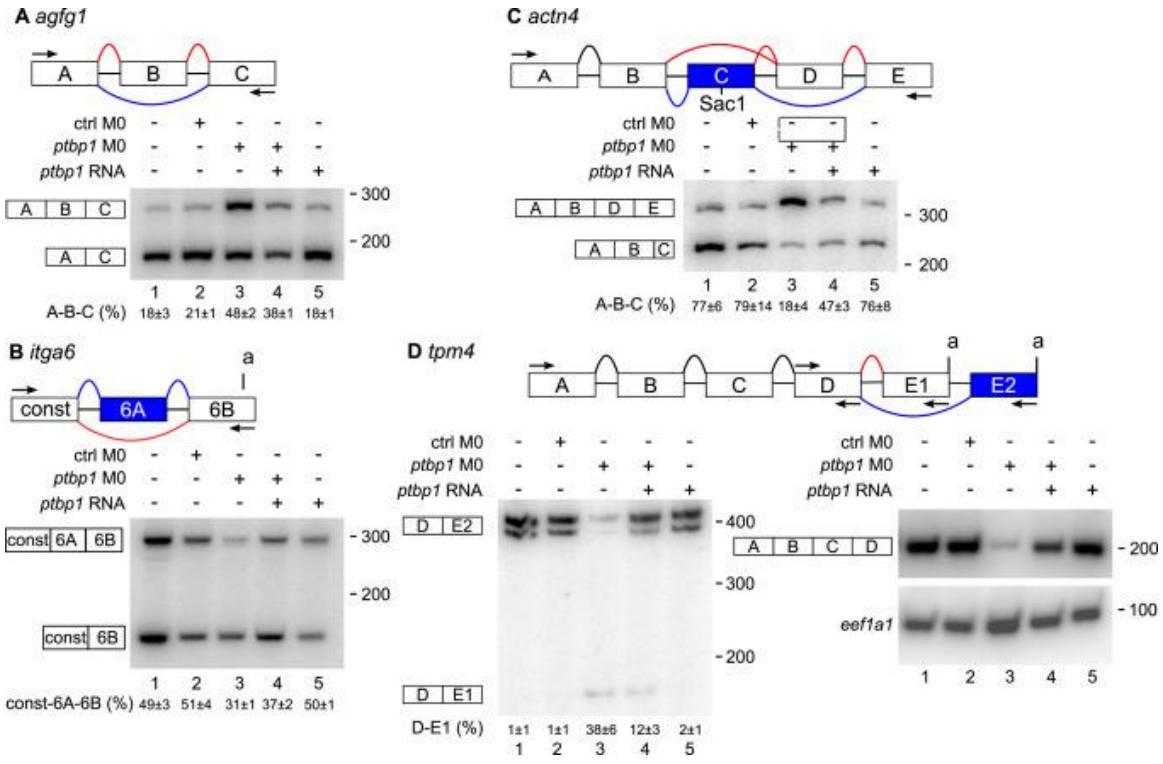

A agfg1
A
B
C
ctrl M0
ptbp1 M0
ptbp1 RNA
A B C
A C
300
200
1 2 3 4 5
A-B-C (%) 18±3 21±1 48±2 38±1 18±1
B itga6
a
l
const
6A
6B
const 6A 6B
const 6B
const-6A-6B (%) 49±3 51±4 31±1 37±2 50±1
C actn4
D
E
Sac1
A B D E
A B C
A-B-C (%) 77±6 79±14 18±4 47±3 76±8
D tpm4
a
a
E1
E2
D E2
D E1
400
D-E1 (%) 1±1 1±1 38±6 12±3 2±1
A B C D
eef1a1
100

Figure 6. Confirmation of newly discovered Ptbp1-controlled splicing events. A, *agfg1*, **B**, *itga6*, **C**, *actn4*, **D**, *tpm4*. **Upper panels**, genomic region encompassing the Ptbp1-regulated exons. Introns are represented as horizontal lines and exons as boxes. Except for *itga6*, the exons are given arbitrary names (A to E), and their genomic coordinates are given in Table ST6. "A" stands for polyadenylation signal. The splice junctions are positioned along the gene structure. The junctions shown in black were not detected as differently used in control embryos and *ptbp1* morphants, while the junctions shown in red and blue were detected as significantly (adjusted $p<0.05$) increased and decreased, respectively, in *ptbp1* morphants. The exons in blue were detected as significantly (adjusted $p<0.05$) decreased in *ptbp1* morphants. **Lower panels**, autoradiograms of representative RT-PCR experiments carried out with RNAs extracted from stage 28 embryos previously injected with the indicated molecules, and using the primers indicated by arrows in the upper panels. In all the experiments, the primers are designed to amplify both pseudo-alleles, and the forward primer is radiolabeled (*). The quantifications below the gels are means ± s.d. of 3 independent experiments. **A**, *agfg1* pre-mRNA contains a cassette exon (B), and its splicing was analysed with primers in flanking exons A and C. **B**, *itga6* pre-mRNA contains a cassette exon (6A), and its splicing was analysed with primers in the constitutive flanking exons "const" and 6B. **C,** *actn4* pre-mRNA contains a set of mutually exclusive exons (C and D), and its splicing was analysed with primers in flanking exons A and E, with *Sac1* digestion before gel loading. **D**, *tpm4* pre-mRNA contains two alternative terminal exons (E1 and E2), and its splicing was analysed with one forward primer in exons D and reverse primers in exons E1 and E2. Exons E2 of *tpm4l* and *tpm4s* differ by an indel located in the 3'UTR, explaining that the D-E2 amplimere is a doublet. The total amount of *tpm4* mRNA was appraised from RT-PCR with primers in exons A and D, and *eef1a1* (EF1a) confirmed similar RNA extractions.

Highlights

ACCEPTED MANUSCRIPT

- The *Xenopus laevis* genome supports differential splicing analysis.
- A junction-centric analysis of splicing allows for efficient detection of splicing events.
- Junction-centric analysis of splicing is robust to annotation changes.
- Ptbp1 regulation of splicing is partially conserved between *Xenopus* and human.

Accepted manuscript