nature portfolio

Peer Review File

Landscape of alternative splicing and polyadenylation during growth and development of muscles in pigs

Corresponding Author: Professor Xiuqin Yang

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This manuscript by Yuanlu Sun, Yu Pang, and colleagues describes a transcriptome study across five time points during pig muscle growth. The main innovation of the work lies in the technology used to profile the transcriptome, specifically the PacBio Iso-Seq. An extensive analysis of alternative splicing and polyadenylation is reported.

The topic is of interest despite its relatively limited scope (pig muscle development). The study is well-designed, and the analysis makes sense. The main results feature a much higher transcriptome complexity than reported in the reference annotation, which is expected, considering that pigs do not benefit from the same genome annotation effort as humans or mice. In livestock species, each transcriptome profiling study brings new genes and transcripts. An interesting observation about a 3'UTR lengthening over time is also noted.

There are still a few points that should be clarified and/or fixed. My main concerns are the following:

1) While the PacBio Iso-Seq data processing seems clear and correctly described (reads from all triplicates have been processed together for each time point separately, and then the resulting annotations have been merged using SQUANTI), this is not the case for the Illumina RNA-seq data. The Methods section lacks important information.

- See for instance line 396: "RNA-seq was used to correct the Iso-Seq data". How? This is not described. How much did the correction change the results?
- 399: "Reads were mapped [...] and assembled with StringTie". Were all the reads from all the 15 samples mapped and assembled together or was this done by time point like for the Iso-Seq? Or for each of the 15 replicates?
- Was a reference annotation used during the mapping and/or assembly step? Which one exactly? Please indicate the version and parameters of each step.
- Was a reference annotation used for the Iso-Seq data processing too? Which one exactly?
- Did this process result in a new gene annotation, or several new annotations? How did it compare with the Iso-Seq annotation?
- How many reads were generated, sequenced, mapped and filtered per library? An equivalent to Sup Table "Statistics of Iso-Seq data" would be appreciated for the Illumina data too.

2) According to the Methods (line 400), "the expression level of gene/transcript was measured with Fragments Per Kilobase of transcript

per Million (FPKM)," which should be provided by StringTie. Then "Counts Per Million (CPM)" are mentioned in the next sentence.

The issue is that none of these metrics should be used by DESeq2, as they both include an improper library-size normalization method. Therefore, the differential expression analysis might need to be redone using proper read count values and a suitable normalization method from the DESeq2 package (TMM, for instance, instead of FPKM or CPM).

3) Sequencing data is not available. No accession is provided for the NovaSeq reads, and while the Iso-Seq data seem to be registered, nothing is visible.

4) The number of AS events per time point (Fig 2d) approximately follows the number of distinct transcripts, the number of AS-genes and that of AS-transcripts. While this trend is reported, the reason is missing.

It seems to be simply due to the sequencing depth: according to the supplementary table "Statistics of Iso-seq data," samples B, C, and D have the highest numbers of FLNCs, while samples A and E have the lowest ones. There seems to be a correlation between the number of CCSs, FLNCs, genes, isoforms (Table 1), and AS events. The possibility that differences in data quantity might explain most of the results discrepancy between samples should be mentioned in the article. This is important because it shows that saturation was not achieved. Therefore, the transcriptome complexity has only been partially captured by the Iso-Seq experiments. The resulting important conclusion is that the reference annotation is even more incomplete than this study shows, and that more sequencing could improve it further. This is also supported by the fact that transcript isoforms often differ between time points, even within the same detected genes.

5) The resulting GTF or BED genome annotation(s) file(s) should be provided, as this is one of the most important outcomes of the study.

Minor points:

- Line 85: "After quality control, a total of 688,352, accounting for 99.97%, non-redundant high-quality consensus isoforms are obtained." Since this number is obtained by adding the number of non-redundant isoforms across samples, I do not think it can be "non-redundant."

- Line 88: "five samples." As there are 15 biological samples in total (3 animals across 5 time points), this "five samples" is a bit confusing. It should be mentioned in the main text that these are merged samples.

- Line 93: Table 1 is informative. The precise definition of "isoform" could be indicated a bit earlier in the main text instead of just in the Methods. Are two transcripts with the same introns but a different 5' extremity (one a bit longer than the other) considered two distinct isoforms? What about intronless transcripts?

- Line 95: Fig 1b: ENSEMBL, not Ensemble.

- Line 107: "none but novel transcripts." Typo.

- Line 111: "99.19% of annotated transcripts were predicted as protein-coding." How was the prediction performed?

- Line 117: "SQANTI3 was used to process the muscle transcriptome constructed." This should be earlier in the text, as Fig 1a already refers to known and novel genes and transcripts.

- Line 124: "there are very a few fusion." Typo.

- Line 145: "both of the Spearman correlations are increased." Typo.

- Line 152: "We obtain most AS events from 30-d-old pigs." See my comment about sequencing depth above.

- Line 167: "RNA-seq reveals numerous differential alternative splicing events." In terms of nomenclature, "RNA-seq" alone is a bit ambiguous because Iso-Seq is an RNA-seq technology too. Please use something like "Illumina RNA-seq" or "NovaSeq" at least in the title section to limit possible confusion.

- Line 176: "3063 DEGs" should be "the 3063 DEGs."

- Line 187: "A total of 7399 unique DAS events." DAS is not defined earlier in the main text.

- Line 209: "AMT is extensively occurred." Typo.

- Line 303: "In particular, The ES is a" should be "In particular, ES is a."

- Fig 7a: Distributions should not be represented with a pie chart.

Reviewer #2

(Remarks to the Author)
In this manuscript, the authors analyzed the transcriptome of pig muscles at five developmental stages using polyadenylation selected long-read isoform sequencing. The analysis of alternative splicing (AS) revealed that most genes underwent alternative splicing. The top 100 genes with the most isoforms were associated with muscle growth and development. Most AS-transcripts were expressed only at specific periods. Genes exhibited more exon skipping in 210-day-old pigs, which may be related to the decreased expression of spliceosome components. The major transcript (MT) was altered, potentially affecting the sequence and function of the protein. Alternative polyadenylation (APA) analysis revealed that 1723 genes had multiple polyadenylation sites (PAS) across all five periods. The proportion of distal PAS was higher in each period. The number of major PAS genes altered at different developmental stages was less than those undergoing alternative splicing. This dataset significantly enriches our understanding of the muscle transcriptome.

The importance of AS and PAS in muscle growth and development needs further emphasis, as it is not very clear. For instance, analyzing the proportion of MT transcripts versus other isoform transcripts could highlight the role of these other isoforms. This would further elucidate the role of AS.

Beyond the positive correlation between the number of transcript isoforms and gene length or exon number, the relationship between the number of transcript isoforms and gene expression remains unclear. Could the abundance of transcript isoforms be by-products of high gene expression and inaccurate regulation? This question can also relate to the proportion of other isoforms in the first question.

Can AS and PAS be analyzed jointly to determine if their changes are coupled?

A motif analysis of specific polyadenylation sites at different periods can be conducted to identify transcription factors in muscle that may regulate variable polyadenylation sites. This analysis may reveal insights more significant than the PAS phenomenon alone.

Were Mgenes and Pgenes transformed at different times? How does this transformation impact gene expression?

Reviewer #3

(Remarks to the Author)
The manuscript presents a comprehensive study on alternative splicing (AS) and alternative polyadenylation (APA) during growth and development of porcine muscles. The research demonstrated that transcript diversity plays an important role in muscle growth and development, and identified major transcript and major polyadenylation site of isoformic genes in different development stages. Furthermore, it revealed that, with aging, long isoforms tend to be prevalent underlying which the mechanisms are explored. The manuscript shows a new approach to alternative polyadenylation analysis based on long-read isoform sequencing. It is well structured and presents a huge quantity of data.

Minor comments

1、In general, Materials and Methods and Results should be written in past tense. Please check carefully to correct use of verb tenses;
2、Table 1, in the table, A, B, C, D, E represent for ?, it should be given;
3、Lines 44-45, the abbreviation for microRNA, long noncoding RNA, and circular RNA should be deleted as they present only one time in the manuscript;
4、Line 105, “novel” should be “novels”;
5、Line 116, “Fig. 1d, e” should be “Figs. 1d, e”, review the full manuscript to avoid the similar issues;
6、Line 166, the end of the line, “Fig. 2G” should be “Fig. 2g”;
7、Lines 173-174, the full name of the genes should be given, please review all the paper to avoid similar issues;
8、Line 183, “Regulation of lipolysis in adipocyte” is repetitive;
9、Line 296, “multiple PAS” should be “multiple PASs”;
10、Line 346, “All these support” should be “All these supports”;
11、Line 372, “the full-length cDNA were synthesized” should be “the full-length cDNA was synthesized”;
12、Line 388, “high quality” should be “high-quality”;
13、Line 386, “FL non-chimeric (FLNC)” should be FLNC, because it has been defined in Line 381;

14、Lines 396-411, the reference or website for some programs missed, such as HISAT2 and DESeq2, should be supplemented;
15、Line 410, the full name of PPI should be given.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Landscape of alternative splicing and polyadenylation during growth and development of muscles in pigs

1) I would like to thank the authors for providing more information about Illumina RNA-seq data production and analysis.

However, my question about how the RNA-seq data was used to improve the Iso-Seq annotation has not been adressed. Changin "correct" by "optimize" in the manuscript does not give much information, and the added lines 421-423 make little sense to me: "The structure of 5' end of known transcripts was optimized with the reference genome (S. scrofa 11.1_release109), while the novel ones was optimized with RNA-Seq data."

Again: how?

2) OK for me.

3) According to https://ngdc.cncb.ac.cn/gsa/s/a6TPc6SR, the Iso-seq data seems to be only available as processed and mapped reads (bam files) but I could not find the original raw sequencing data. Shouldn't it be provided too?

Also, the data release date is set to "2025-11-18". This has to be changed before the article gets published, I am assuming this will be fixed.

4) OK for me.

5) If I understand correctly, there should be 6 annotations: one per stage (5) and one final/merged/total annotation mentioned in line 95 ("All downstream analyses were based on the subset of Cupcake-filtered transcripts").

However, the authors provided two annotation files:
- Supplementary Data 1: The genome annotation file of each stage of Iso-Seq
- Supplementary Data 2: The genome annotation file of the sum of five stages of Iso-Seq

If I am not mistaken, both provided files contain one annotation each:
- In SupData 1, I count 22691 genes (22040 ENSEMBL IDs and 651 PacBio IDs) and 73145 transcripts (45893 ENSEMBL IDs and 27252 PacBio IDs).
- In SupData 2, I count 22681 genes (22040 ENSEMBL IDs and 641 PacBio IDs) and 73427 transcripts (45893 ENSEMBL IDs and 27534 PacBio IDs).
So I guess that SupData 2 corresponds to the "Total" column of Table 1, with 641 novel genes and 27534 novel transcripts, therefore to the merged annotation used in the study.

Could the authors then clarify in the manuscript what SupData 1 is exactly, and what is the difference between these two files?

New typos:
84: "filtered"
87: "GTT"
89: "aligned"

Reviewer #2

(Remarks to the Author)
The author answered all my questions and I recommend accepting the manuscript.

Rebuttal letter

Reviewers' comments:

Reviewer #1:

1) While the PacBio Iso-Seq data processing seems clear and correctly described (reads from all triplicates have been processed together for each time point separately, and then the resulting annotations have been merged using SQUANTI), this is not the case for the Illumina RNA-seq data. The Methods section lacks important information.
RE: We sincerely thank the reviewer for careful reading. The information lacked had been supplemented in the manuscript, please check in Lines 427-436.

- See for instance line 396: "RNA-seq was used to correct the Iso-Seq data". How? This is not described. How much did the correction change the results?
RE: We are sorry for our carelessness. The RNA-Seq was used for structural optimization of the ends of Iso-Seq data in which the genes were novel. We did not describe clearly in the original manuscript, now we have supplemented the detailed information. Please check in Lines 421-423.

- 399: "Reads were mapped [...] and assembled with StringTie". Were all the reads from all the 15 samples mapped and assembled together or was this done by time point like for the Iso-Seq? Or for each of the 15 replicates?
RE: Thanks to the experts' review. Each of the 15 replicates were mapped and assembled individually. We did not describe clearly in the original manuscript. Now the detailed information was added. Please check in Lines 438-441.

- Was a reference annotation used during the mapping and/or assembly step? Which one exactly? Please indicate the version and parameters of each step.
RE: Thank the reviewer for careful reading. Yes, we used reference annotation (*S. scrofa* 11.1_release109) during the mapping and assembly step. Now the information was supplemented (Lines 438-441). Both of mapping and assembly step used the same reference annotation. The version and main parameters of softwares used for mapping and assembly step have been added in Methods section (Lines 438-441).

- Was a reference annotation used for the Iso-Seq data processing too? Which one exactly?
RE: Thank the reviewer for careful reading. Both of NovaSeq and Iso-Seq used the same reference annotation. The reference annotation version was *S. scrofa* 11.1_release109. The information of reference annotation has been added in Line 417, Line 422 and Line 439.

- Did this process result in a new gene annotation, or several new annotations? How did it compare with the Iso-Seq annotation?

RE: Thank the reviewer for careful reading. RNA-seq produced new gene annotation, but we did not analyze the annotation, and did not compare it with the Iso-Seq annotation.

- How many reads were generated, sequenced, mapped and filtered per library? An equivalent to Sup Table "Statistics of Iso-Seq data" would be appreciated for the Illumina data too.

RE: We agree with the reviewer. Based on the comment, the original “Supplementary Table 2 Statistics of Iso-Seq” has been changed to "Supplementary Table 2 Statistics of Iso-Seq and RNA-Seq data".

2) According to the Methods (line 400), "the expression level of gene/transcript was measured with Fragments Per Kilobase of transcript per Million (FPKM)," which should be provided by StringTie. Then "Counts Per Million (CPM)" are mentioned in the next sentence. The issue is that none of these metrics should be used by DESeq2, as they both include an improper library-size normalization method. Therefore, the differential expression analysis might need to be redone using proper read count values and a suitable normalization method from the DESeq2 package (TMM, for instance, instead of FPKM or CPM).

RE: Thanks very much for your careful review. In this study, read count values were used by DESeq2 rather than FPKM or CPM to analyze differentially expressed genes (DEGs). FPKM was just a method that we used to provide information on gene expression levels. CPM was just used to filter out non--expressed genes during DEG identification. In the original manuscript we did not describe clearly, and now detailed information was given. Please check in Lines 442-447. We are so sorry for our carelessness and mistakes. Thanks very much for your expert's reminder.

3) Sequencing data is not available. No accession is provided for the NovaSeq reads, and while the Iso-Seq data seem to be registered, nothing is visible.

RE: Thank the reviewer for careful reading. We feel sorry for our carelessness. Both of the NovaSeq reads and the Iso-Seq data had been registered. The new accession (https://ngdc.cncb.ac.cn/gsa/s/a6TPc6SR) has been provided in the manuscript (Line 628).

4) The number of AS events per time point (Fig 2d) approximately follows the number of distinct transcripts, the number of AS-genes and that of AS-transcripts. While this trend is reported, the reason is missing.

It seems to be simply due to the sequencing depth: according to the supplementary table "Statistics of Iso-seq data," samples B, C, and D have the highest numbers of FLNCs, while samples A and E have the lowest ones. There seems to be a correlation between the number of CCSs, FLNCs, genes, isoforms (Table 1), and AS events. The possibility that differences in data quantity might explain most of the results discrepancy between samples should be mentioned in the article. This is important because it shows that saturation was not achieved. Therefore, the transcriptome complexity has only been partially captured by the Iso-Seq experiments. The resulting important conclusion is that the reference annotation is even more incomplete than this study shows, and that more sequencing could

improve it further. This is also supported by the fact that transcript isoforms often differ between time points, even within the same detected genes.
RE: Yes, you are right. Thanks very much for your careful review and the professional opinion. The number of AS events per time point might correlate with data quantity. So, the description “We obtained most AS events from 30-d-old pigs that stage obtained the most of CCS, FLNC and high-quality isoforms (n = 3395, associated with 1643 genes, 38.8% of all AS-genes).” is not accurate. The result might be caused by unparalleled data quantity. We have deleted it from the revised manuscript and checked the full text to avoid the similar mistakes. Please check in Line 155. The corresponding discussion and conclusion have been added in the revised manuscript as suggested, please check in Lines 312-318, and Line 379. Thanks again.

5) The resulting GTF or BED genome annotation(s) file(s) should be provided, as this is one of the most important outcomes of the study.
RE: Thanks for your constructive suggestion. The new gene annotation has been referred in revised manuscript, please check in Lines 86-87.

Minor points:

- Line 85: "After quality control, a total of 688,352, accounting for 99.97%, non-redundant high-quality consensus isoforms are obtained." Since this number is obtained by adding the number of non-redundant isoforms across samples, I do not think it can be "non-redundant."
RE: Yes, you are right. It is our mistake. Based on the comments, we have deleted the word “non-redundant” within the whole manuscript.

- Line 88: "five samples." As there are 15 biological samples in total (3 animals across 5 time points), this "five samples" is a bit confusing. It should be mentioned in the main text that these are merged samples.
RE: Thank the reviewer for careful reading. As suggested by reviewer, we have corrected the “five samples” into “five stages” (Line 88) and checked the full text to avoid the similar mistakes.

- Line 93: Table 1 is informative. The precise definition of "isoform" could be indicated a bit earlier in the main text instead of just in the Methods.
Are two transcripts with the same introns but a different 5' extremity (one a bit longer than the other) considered two distinct isoforms? What about intronless transcripts?
RE: Thanks to the review of the expert. Based on the comments, we have indicated “Isoforms” in the Introduction section (Line 49). Owing to the intrinsic characteristics of Iso-Seq, the 5' extremity might be incomplete. Here, we considered the sequences with a different 5' extremity as one isoform. We did not describe clearly in original manuscript. Now detailed information was supplemented in Methods sections. Please check in Line 419-421.

- Line 95: Fig 1b: ENSEMBL, not Ensemble.

RE: We feel sorry for our carelessness. In our resubmitted manuscript, the typo has been revised. Thanks for your correction. Please check in Fig. 1b.

- Line 107: "none but novel transcripts." Typo.

RE: Thanks for your careful checks. We have corrected "none but novel transcripts." into “only novel transcripts”. Please check in Line107.

- Line 111: "99.19% of annotated transcripts were predicted as protein-coding." How was the prediction performed?

RE: Thanks to the review of the expert. We predicted protein-coding potential with SQANTI3 (v5.1.1) by the default parameters. We have added the content in Methods section (line 426).

- Line 117: "SQANTI3 was used to process the muscle transcriptome constructed." This should be earlier in the text, as Fig 1a already refers to known and novel genes and transcripts.

RE: Yes, you are right. We corrected the contents as suggested. Please check in Lines 84-85 and Line 89.

- Line 124: "there are very a few fusion." Typo.

RE: Thanks for your careful checks. We have corrected " there are very a few fusion." into “there were a small number of fusions”. Please check in Line 124.

- Line 145: "both of the Spearman correlations are increased." Typo.

RE: Thanks for your careful checks. We have corrected " both of the Spearman correlations are increased." into “both the Spearman correlations increased.” Please check in Lines 145-146.

- Line 152: "We obtain most AS events from 30-d-old pigs. " See my comment about sequencing depth above.

RE: We sincerely thank the reviewer for careful reading. We have deleted the sentence from the revised manuscript. Please check in Line 155.

- Line 167: "RNA-seq reveals numerous differential alternative splicing events." In terms of nomenclature, "RNA-seq" alone is a bit ambiguous because Iso-Seq is an RNA-seq technology too. Please use something like "Illumina RNA-seq" or "NovaSeq" at least in the title section to limit possible confusion.

RE: Thanks for your careful checks. We have corrected "RNA-seq" into “Illumina RNA-seq” as suggested. Please check in Line 170.

- Line 176: "3063 DEGs" should be "the 3063 DEGs."

RE: Thanks for your correction. The typo is revised in the manuscript. Please check in Line

183.

- Line 187: "A total of 7399 unique DAS events." DAS is not defined earlier in the main text.

RE: We feel sorry for our carelessness. We have defined the DAS earlier as suggested. Please check in Line 195 and Line 457. We have checked the full text to avoid the similar mistakes.

- Line 209: "AMT is extensively occurred." Typo.

RE: Thanks for your correction. We have revised "AMT is extensively occurred" into “AMT occurs extensively”. Please check in Line 217.

- Line 303: "In particular, The ES is a" should be "In particular, ES is a."

RE: Thanks for your correction. The typo has been revised in the manuscript. Please check in Line 330.

- Fig 7a: Distributions should not be represented with a pie chart.

RE: Thanks for your careful checks. We have represented distributions with a bar graph instead of a pie chart. Please check in Fig. 7a.

Reviewer #2:

Beyond the positive correlation between the number of transcript isoforms and gene length or exon number, the relationship between the number of transcript isoforms and gene expression remains unclear. Could the abundance of transcript isoforms be by-products of high gene expression and inaccurate regulation? This question can also relate to the proportion of other isoforms in the first question.

RE: We sincerely thank the reviewer for careful reading. According to Spearman correlation analysis, the relationship between the number of transcript isoforms and gene expression were found that there was a positive correlation (r_s=0.547) between them (Fig. 2d). You are clever. There might be by-products of high gene expression and inaccurate regulation although RT-PCR verification of AS transcripts/events indicated the reliablity of Iso-Seq data (Figs. 2h and 4d in the resubmitted manuscript). Thanks for the professional opinion, we have added the content in Results section (Lines 148-149) and Discussion section (Lines 318-322).

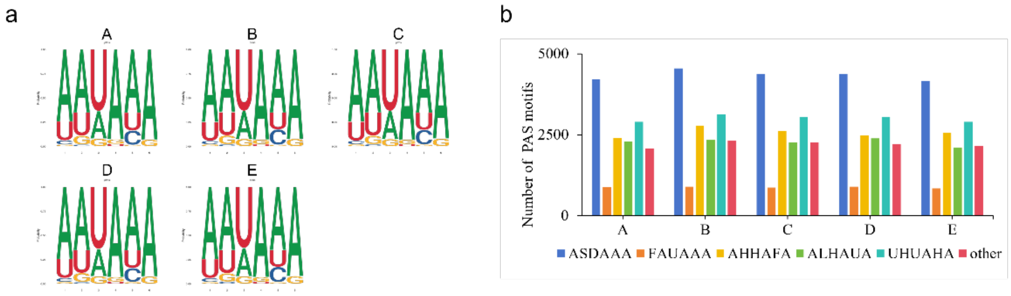
Can AS and PAS be analyzed jointly to determine if their changes are coupled?

RE: Thanks for your good suggestion. We made a correlation analysis between the number of AS and PAS, and found that AS and APA are relatively independent events. The results have been added in the resubmitted manuscript. Please check in Line241-243.

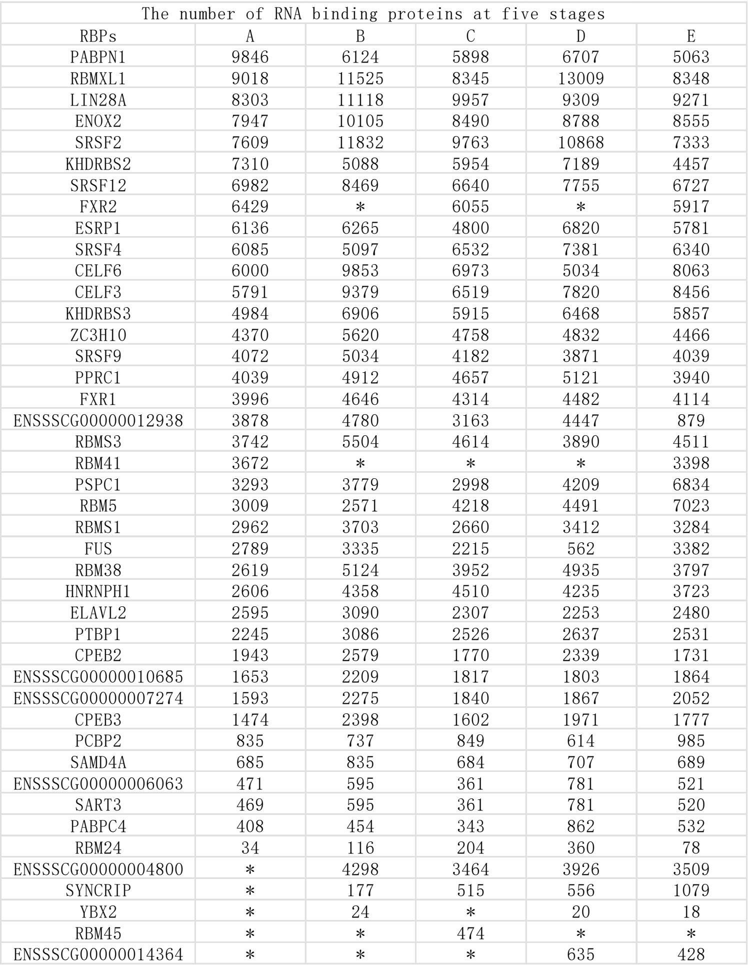
A motif analysis of specific polyadenylation sites at different periods can be conducted to identify transcription factors in muscle that may regulate variable polyadenylation sites. This analysis may reveal insights more significant than the PAS phenomenon alone.

RE: Thanks to the review of the expert. We calculated the base frequency occurred in

PASs and the main kinds of hexameric motifs of PASs among five stages of development. We found that the most frequently used bases at the PAS sites were A, A, U, A, A, and A. Both the base frequency and the number of six enriched hexameric motifs of PASs among the five stages were similar. We also analyzed the distribution of RBPs on the region around the PAS (±100 nt) at five stages using Analysis of Motif Enrichment (AME, v5.5.7) method of the MEME suite and obtained the motifs of RBPs from CisBP-RNA database (http://cisbp-rna.ccbr.utoronto.ca/bulk.php). The type of RBPs at five stages were similar too. Please check in the following figures. So, we speculated that little differences existed in transcription factors among five stages. The results were not given in the revised manuscript.



The motif analysis of polyadenylation sites. a, The base frequency of PAS motif at five stages. **b,** The number of PAS motif at five stages. Six enriched hexameric motifs were identified with ASDAAA (S=A, U, G or C, D= U or G), FAUAAA (F=G or C), AHHAFA (H=A or U), UHUAHA, and Others (AAAAAG, AAUGAA, AGAUAA).



The number of RNA binding proteins at five stages					
RBPs	A	B	C	D	E
PABPN1	9846	6124	5898	6707	5063
RBMXL1	9018	11525	8345	13009	8348
LIN28A	8303	11118	9957	9309	9271
ENOX2	7947	10105	8490	8788	8555
SRSF2	7609	11832	9763	10868	7333
KHDRBS2	7310	5088	5954	7189	4457
SRSF12	6982	8469	6640	7755	6727
FXR2	6429	*	6055	*	5917
ESRP1	6136	6265	4800	6820	5781
SRSF4	6085	5097	6532	7381	6340
CELF6	6000	9853	6973	5034	8063
CELF3	5791	9379	6519	7820	8456
KHDRBS3	4984	6906	5915	6468	5857
ZC3H10	4370	5620	4758	4832	4466
SRSF9	4072	5034	4182	3871	4039
PPRC1	4039	4912	4657	5121	3940
FXR1	3996	4646	4314	4482	4114
ENSSSCG00000012938	3878	4780	3163	4447	879
RBMS3	3742	5504	4614	3890	4511
RBM41	3672	*	*	*	3398
PSPC1	3293	3779	2998	4209	6834
RBM5	3009	2571	4218	4491	7023
RBMS1	2962	3703	2660	3412	3284
FUS	2789	3335	2215	562	3382
RBM38	2619	5124	3952	4935	3797
HNRNPH1	2606	4358	4510	4235	3723
ELAVL2	2595	3090	2307	2253	2480
PTBP1	2245	3086	2526	2637	2531
CPEB2	1943	2579	1770	2339	1731
ENSSSCG00000010685	1653	2209	1817	1803	1864
ENSSSCG00000007274	1593	2275	1840	1867	2052
CPEB3	1474	2398	1602	1971	1777
PCBP2	835	737	849	614	985
SAMD4A	685	835	684	707	689
ENSSSCG00000006063	471	595	361	781	521
SART3	469	595	361	781	520
PABPC4	408	454	343	862	532
RBM24	34	116	204	360	78
ENSSSCG00000004800	*	4298	3464	3926	3509
SYNCRIP	*	177	515	556	1079
YBX2	*	24	*	20	18
RBM45	*	*	474	*	*
ENSSSCG00000014364	*	*	*	635	428

Were Mgenes and Pgenes transformed at different times? How does this transformation impact gene expression?

RE: Thanks to the experts' review. Yes, you are right. Mgenes and Pgenes transformed at different times, and with time increasing, the number of Mgenes increase, i.e. the expression level of isoforms with distal PAS increases in general. The effect of this transformation on gene expression was analyzed as suggested, please check in Lines 265-271 and Figs. 6f, g.

Reviewer #3:

Minor comments

1、In general, Materials and Methods and Results should be written in past tense. Please check carefully to correct use of verb tenses;

RE: We sincerely thank the reviewer for careful reading. We are sorry for our carelessness. We have checked and corrected the use of verb tenses within the whole manuscript.

2、Table 1, in the table, A, B, C, D, E represent for ?, it should be given;

RE: We are sorry for our carelessness. We have given the table legend under the Table 1. Please check in Table 1.

3、Lines 44-45, the abbreviation for microRNA, long noncoding RNA, and circular RNA should be deleted as they present only one time in the manuscript;

RE: Thanks to the review of the expert. We have deleted the abbreviation. Please check in Lines 43-45.

4、Line 105, “novel” should be “novels”;

RE: We feel sorry for our carelessness. In our resubmitted manuscript, the typo is revised. Thanks for your correction. Please check in Line 107.

5、Line 116, “Fig. 1d, e” should be “Figs. 1d, e”, review the full manuscript to avoid the similar issues;

RE: Thanks for your correction. Please check in Line 116. We have checked and corrected the typo within the whole manuscript.

6、Line 166, the end of the line, “Fig. 2G” should be “Fig. 2g”;

RE: Thanks for your correction. The typo is revised.

7、Lines 173-174, the full name of the genes should be given, please review all the paper to avoid similar issues;

RE: Thanks for your careful checks. Based on your comments, we have wrote the full name of the genes, and have reviewed the whole manuscript. Please check in Lines 177-181.

8、Line 183, “Regulation of lipolysis in adipocyte” is repetitive;

RE: Thanks for your careful checks. We have deleted the extra word. Please check in Line 190.

9、Line 296, “multiple PAS” should be “multiple PASs”;

RE: Thanks for your correction. The typo is revised. Please check in Line 324.

10、Line 346, “All these support” should be “All these supports”;

RE: Thanks for your correction. The typo is revised. Please check in Line 372.

11、Line 372, “the full-length cDNA were synthesized” should be “the full-length cDNA was synthesized”;

RE: Thanks for your correction. The typo is revised. Please check in Line 400.

12、Line 388, “high quality” should be “high-quality”;

RE: Thanks for your correction. The typo is revised. Please check in Line 416. We have checked the full text to avoid the similar mistakes.

13、Line 386, “FL non-chimeric (FLNC)” should be FLNC, because it has been defined in Line 381;

RE: Thanks for your correction. The typo is revised. Please check in Line 409. We have checked the full text to avoid the similar mistakes.

14、Lines 396-411, the reference or website for some programs missed, such as HISAT2 and DESeq2, should be supplemented;

RE: Thanks for your careful checks. We have supplemented the information of version or website in the manuscript. Please check in Lines 438-444.

15、Line 410, the full name of PPI should be given.

RE: Thanks for your careful checks. We have given the full name of PPI in Results section of the manuscript. Please check in Lines 232-233.

Rebuttal letter

Reviewers' comments:

Reviewer #1:

1) I would like to thank the authors for providing more information about Illumina RNA-seq data production and analysis.

However, my question about how the RNA-seq data was used to improve the Iso-Seq annotation has not been adressed. Changin "correct" by "optimize" in the manuscript does not give much information, and the added lines 421-423 make little sense to me: "The structure of 5' end of known transcripts was optimized with the reference genome (S. scrofa 11.1_release109), while the novel ones was optimized with RNA-Seq data."

Again: how?

RE: Thanks very much for your careful reviewing. We are sorry for our carelessness. Now the detailed information has been added into the revised manuscript. Please check in Lines 421-430.
Sorry again.

2) OK for the reviewer.

3) According to https://ngdc.cncb.ac.cn/gsa/s/a6TPc6SR, the Iso-seq data seems to be only available as processed and mapped reads (bam files) but I could not find the original raw sequencing data. Shouldn't it be provided too?

Also, the data release date is set to "2025-11-18". This has to be changed before the article gets published, I am assuming this will be fixed.

RE: Iso-Seq was performed in the Pacbio Sequel II, and the raw data was output as bam files. So, the data uploaded were orginal and not processed and mapped. We have already released the data as suggested. Please check in https://bigd.big.ac.cn/gsa/browse/CRA013591.

4) OK for the reviewer.

5) If I understand correctly, there should be 6 annotations: one per stage (5) and one final/merged/total annotation mentioned in line 95 ("All downstream analyses were based on the subset of Cupcake-filtered transcripts").

However, the authors provided two annotation files:

- Supplementary Data 1: The genome annotation file of each stage of Iso-Seq
- Supplementary Data 2: The genome annotation file of the sum of five stages of Iso-Seq

If I am not mistaken, both provided files contain one annotation each:
- In SupData 1, I count 22691 genes (22040 ENSEMBL IDs and 651 PacBio IDs) and 73145 transcripts (45893 ENSEMBL IDs and 27252 PacBio IDs).
- In SupData 2, I count 22681 genes (22040 ENSEMBL IDs and 641 PacBio IDs) and 73427 transcripts (45893 ENSEMBL IDs and 27534 PacBio IDs).

So I guess that SupData 2 corresponds to the "Total" column of Table 1, with 641 novel genes and 27534 novel transcripts, therefore to the merged annotation used in the study.

Could the authors then clarify in the manuscript what SupData 1 is exactly, and what is the difference between these two files?

RE: Yes, you are very right. There are should 6 annotation files including five individual stages and one merged annotation. We made a mistake in uploading Sup Data1 last time. We are so sorry for our carelessness and mistakes, and thanks very much for your careful review. Now we withdraw the original Sup Data1, the original Sup Data2 (renamed Supplementary Data6 in revised manuscript) was used in this study.
Now we uploaded the annotation files of five stages separately de novo, named Supplementary Data1-5 in revised manuscript.
There are differences in processing the data presented in the original Sup Data1 and Sup Data2. For original Sup Data1, the raw data of five sample was first filtered with cDNA_Cupcake (v28.0.0) individually, followed by SQANTI (v5.1.1) filtering, and then the data of five sample were merged into one, finally redundant sequences were removed.
For original Sup Data2, the raw data of five sample was first filtered with cDNA_Cupcake (v28.0.0) individually, and the results were merged into one integrative data. Next, the redundant sequences caused by merging were removed with cDNA_Cupcake (v28.0.0). Then SQANTI (v5.1.1) filtering was used.
Thanks very much for your expert's reminder.

New typos:
84: "filtered"
RE: Thanks for your correction. We have revised "filtered" into “filtering”. Please check in Line 85.

87: "GTT"
RE: Thanks for your correction. We have deleted "GTT". Please check in Line 87.

89: "aligned"
RE: Thanks for your correction. We have revised "aligned" into “aligning”. Please check in Line 89.