Cell Reports, Volume *26*

Supplemental Information

Regulation of Intronic Polyadenylation by PCF11 Impacts mRNA Expression of Long Genes

Ruijia Wang, Dinghai Zheng, Lu Wei, Qingbao Ding, and Bin Tian

SUPPLEMENTARY MATERIALS

Supplementary Tables

Table S1. Sequencing data generated in this study, Related to STAR Methods.

GEO sample ID	Sequencing method	Sample description
GSM3190450	3'READS	Total RNA, proliferation sample, C2C12 cells, replicate 1
GSM3190451	3'READS	Total RNA, proliferation sample, C2C12 cells, replicate 2
GSM3190452	3'READS	Total RNA, differentiation, C2C12 cells, replicate 1
GSM3190453	3'READS	Total RNA, differentiation, C2C12 cells, replicate 2
GSM3171721	3'READS	4sU labeled RNA, control sample, C2C12 cells, replicate 1
GSM3171722	3'READS	4sU labeled RNA, control sample, C2C12 cells, replicate 2
GSM3171723	3'READS	Total RNA, control sample, C2C12 cells, replicate 1
GSM3171724	3'READS	Total RNA, control sample, C2C12 cells, replicate 2
GSM3171725	3'READS	4sU labeled RNA, siPcf11 sample, C2C12 cells, replicate 1
GSM3171726	3'READS	4sU labeled RNA, siPcf11 sample, C2C12 cells, replicate 2
GSM3171727	3'READS	Total RNA, siPcf11 sample, C2C12 cells, replicate 1
GSM3171728	3'READS	Total RNA, siPcf11 sample, C2C12 cells, replicate 2
GSM3171729	3'READS	Total RNA, 4T1 WT cells, replicate 1
GSM3171730	3'READS	Total RNA, 4T1 WT cells, replicate 2
GSM3171731	3'READS	Total RNA, 4T1 IPA^{Pcf11}-KO cells, replicate 1
GSM3171732	3'READS	Total RNA, 4T1 IPA^{Pcf11}-KO cells, replicate 2
GSM3506180	3'READS	Total RNA, control sample, NIH3T3 cells, replicate 1
GSM3506181	3'READS	Total RNA, control sample, NIH3T3 cells, replicate 2
GSM3506182	3'READS	Total RNA, siPcf11 sample, NIH3T3 cells, replicate 1
GSM3506183	3'READS	Total RNA, siPcf11 sample, NIH3T3 cells, replicate 2
GSM3171746	RNA-seq	Total RNA, control sample, 3T3-L1 cells, replicate 1
GSM3171747	RNA-seq	Total RNA, siPcf11 sample, 3T3-L1 cells, replicate 1

Table S2. Real-time PCR primers used in this study, Related to STAR Methods.

Gene Name (target region)	Purpose	Sequence
Pcf11 (exon 1 and intron 1)	IPA isoform expression	Forward: 5'-GCTGACCATTCTAGCCGAGGAGAA Reverse: 5'-GAAGAATAGGAGGCTGCGGG
Pcf11 (exon 1 – exon2)	Splicing of intron 1	Forward: 5'-GGAAGAGAATATCTCACTGCCTT Reverse: 5'-TGGAAGCTTCTCTGAGGAAGGA
Pcf11 (exon 2 – exon3)	Gene expression	Forward: 5'-GGAAGAGAATATCTCACTGCCTT Reverse: 5'-GCAGAGGTTTAATAGGCCAAGC
Pcf11 (exon 12 – exon14)	Gene expression	Forward: 5'- GCAAAACAGAACCGAGAAAGA Reverse: 5'- TGTTCTTGACAGATTTCACAACTC
Pcf11(exon 15 – exon16)	Gene expression	Forward: 5'-ACCATCCATCATGTTATGAAGATTATCA Reverse: 5'-TGCAATTCGTTTTTGACAATGTT
Fgf2	IPA isoform expression	Forward: 5'-AAACAGGAACCGGAAGTGCAT Reverse: 5'-ATACCCCATCACTGTCCCTTG
Fgf2	TPA isoform expression	Forward: 5'-CTTCACGGAACTCAGCTGCTA Reverse: 5'-TAGGGTAGCATACCTTGGCG
Cstf3	TPA isoform expression	Forward: 5'-ACAAGTGGATGAGCTGATGGAA Reverse: 5'-CTGAATCCTCGTTGGGCCTT
Cstf3	IPA isoform expression	Forward: 5'-ATAGACAAAGCACGGAAGACT Reverse: 3'-GTGTAAGCTGTAATTGCCATC
CYPH	Gene expression	Forward: 5'-ATGGTCAACCCCACCGTGT Reverse: 5'-TTCCTGCTGTCTTTGGAACTTTGTC
GAPDH	Gene expression	Forward: 5'-TCACCACCATGGAGAAGGC Reverse: 5'-GCTAAGCAGTTGGTGGTGCA

Table S3. Other primers used in this study, Related to STAR Methods.

Purpose	Sequence
PCR and Sanger sequencing to validate IPA site KO	Forward: 5'-GGGTATAGGGAATTGGCCCC Reverse: 5'-ACTGTGTGGGGGCAAACTATT
sgRNA cloning, upstream of *Pcf11* IPA site	Forward: 5'-CACCGACCGTCTCTAAACAACATAT Reverse: 5'-AAACATATGTTGTTTAGAGACGGTC
sgRNA cloning, downstream of *Pcf11* IPA site	Forward: 5'-CACCGCTGCTTCACAGGCATTCGAC Reverse: 5'-AAACGTCGAATGCCTGTGAAGCAGC
Pcf11 IPA site strength analysis	Forward: 5'-ATATATCTCGAGGCAGAGTGAACCCCTT Reverse: 5'-CCGGAATTCCACACAACCACAAAAGT
Pcf11 proximal 3'UTR PAS strength analysis	Forward: 5'-ATATATCTCGAGCCTTCAGCTATCATTTGG Reverse: 5'-CCGGAATTCATCATGTTAAGATTGCTGTTC
Pcf11 3'UTR distal PAS strength analysis	Forward: 5'-ATATATCTCGAGGGTTTTGATTGAATTAGATGGG Reverse: 5'-CCGGAATTCCTGGGAGTATGGCTCATAACT
Control (*Pcf11* intron 1 sequence) for PAS strength analysis	Forward: 5'-ATATATCTCGAGAGTTTTTGAGCACATGTTTCCT Reverse: 5'-CCGGAATTCCAAGTGAAGTCCTTCCATGTAAT

Figure S1

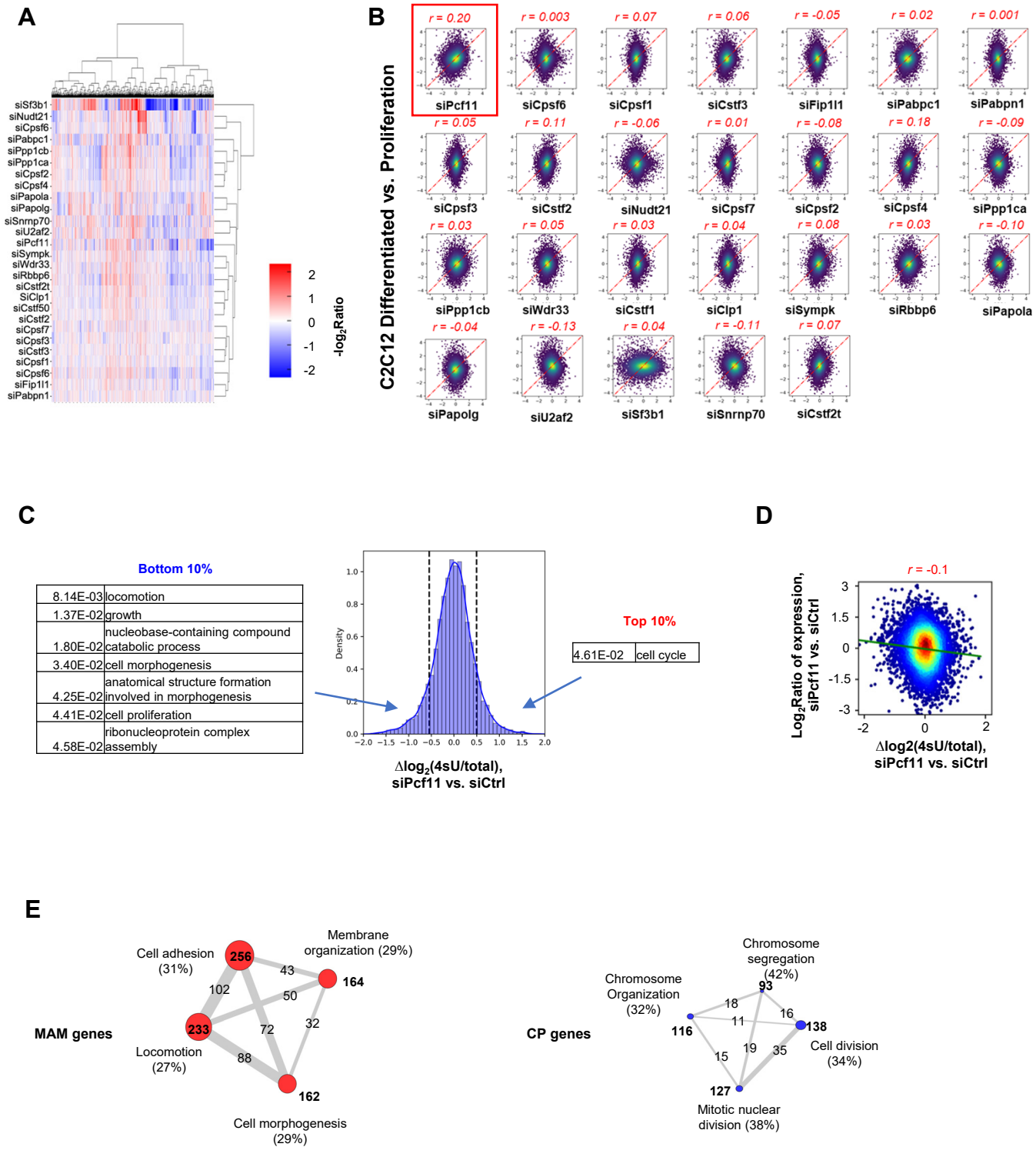

A
siSf3b1
siNudt21
siCpsf6
siPabpc1
siPpp1cb
siPpp1ca
siCpsf2
siCpsf4
siPapola
siPapolg
siSnrnp70
siU2af2
siPcf11
siSympk
siWdr33
siRbbp6
siCstf2t
siClp1
siCstf50
siCstf2
siCpsf7
siCpsf3
siCstf3
siCpsf1
siCpsf6
siFip1l1
siPabpn1
2
1
0
-1
-2
-log2Ratio
B
C2C12 Differentiated vs. Proliferation
r = 0.20
siPcf11
r = 0.003
siCpsf6
r = 0.07
siCpsf1
r = 0.06
siCstf3
r = -0.05
siFip1l1
r = 0.02
siPabpc1
r = 0.001
siPabpn1
r = 0.05
siCpsf3
r = 0.11
siCstf2
r = -0.06
siNudt21
r = 0.01
siCpsf7
r = -0.08
siCpsf2
r = 0.18
siCpsf4
r = -0.09
siPpp1ca
r = 0.03
siPpp1cb
r = 0.05
siWdr33
r = 0.03
siCstf1
r = 0.04
siClp1
r = 0.08
siSympk
r = 0.03
siRbbp6
r = -0.10
siPapola
r = -0.04
siPapolg
r = -0.13
siU2af2
r = 0.04
siSf3b1
r = -0.11
siSnrnp70
r = 0.07
siCstf2t
C
Bottom 10%
8.14E-03 locomotion
1.37E-02 growth
1.80E-02 nucleobase-containing compound catabolic process
3.40E-02 cell morphogenesis
4.25E-02 anatomical structure formation involved in morphogenesis
4.41E-02 cell proliferation
4.58E-02 ribonucleoprotein complex assembly
Density
Δlog2(4sU/total), siPcf11 vs. siCtrl
Top 10%
4.61E-02 cell cycle
D
r = -0.1
Log2Ratio of expression, siPcf11 vs. siCtrl
Δlog2(4sU/total), siPcf11 vs. siCtrl
E
MAM genes
Cell adhesion (31%)
Membrane organization (29%)
Locomotion (27%)
Cell morphogenesis (29%)
256
164
233
162
43
102
50
72
32
88
CP genes
Chromosome segregation (42%)
Chromosome Organization (32%)
Cell division (34%)
Mitotic nuclear division (38%)
93
116
138
127
18
16
11
15
19
35

Figure S1, Related to Figure 1.

(A) Heatmap of gene expression changes by various KDs. Only significantly regulated genes in at least one KD sample are shown. Genes and samples were clustered based on Pearson Correlation. Replicates were combined.

(B) Correlation of gene expression changes in KD cells and those in C2C12 differentiation. siPcf11 sample is highlighted.

(C) Distribution of $\log_2$(4sU/total) difference between siPcf11 and siCtrl samples. The data are based on two replicates. GO terms enriched for the top and bottom 10% of genes are indicated. A high ∆log2(4sU/total) value indicates mRNA destabilization, whereas a low value stabilization.

(D) Correlation between ∆log2(4sU/total) and expression change by *Pcf11* KD.

(E) Two GO term groups (MAM and CP) associated with regulated genes. For each group, the number of upregulated (MAM group) or downregulated genes (CP group) associated with each GO term is indicated, and number of overlap genes between GO terms is indicated on the edge connecting them. Percentage values indicate percent of genes shared with other GO terms in the group.

Figure S2

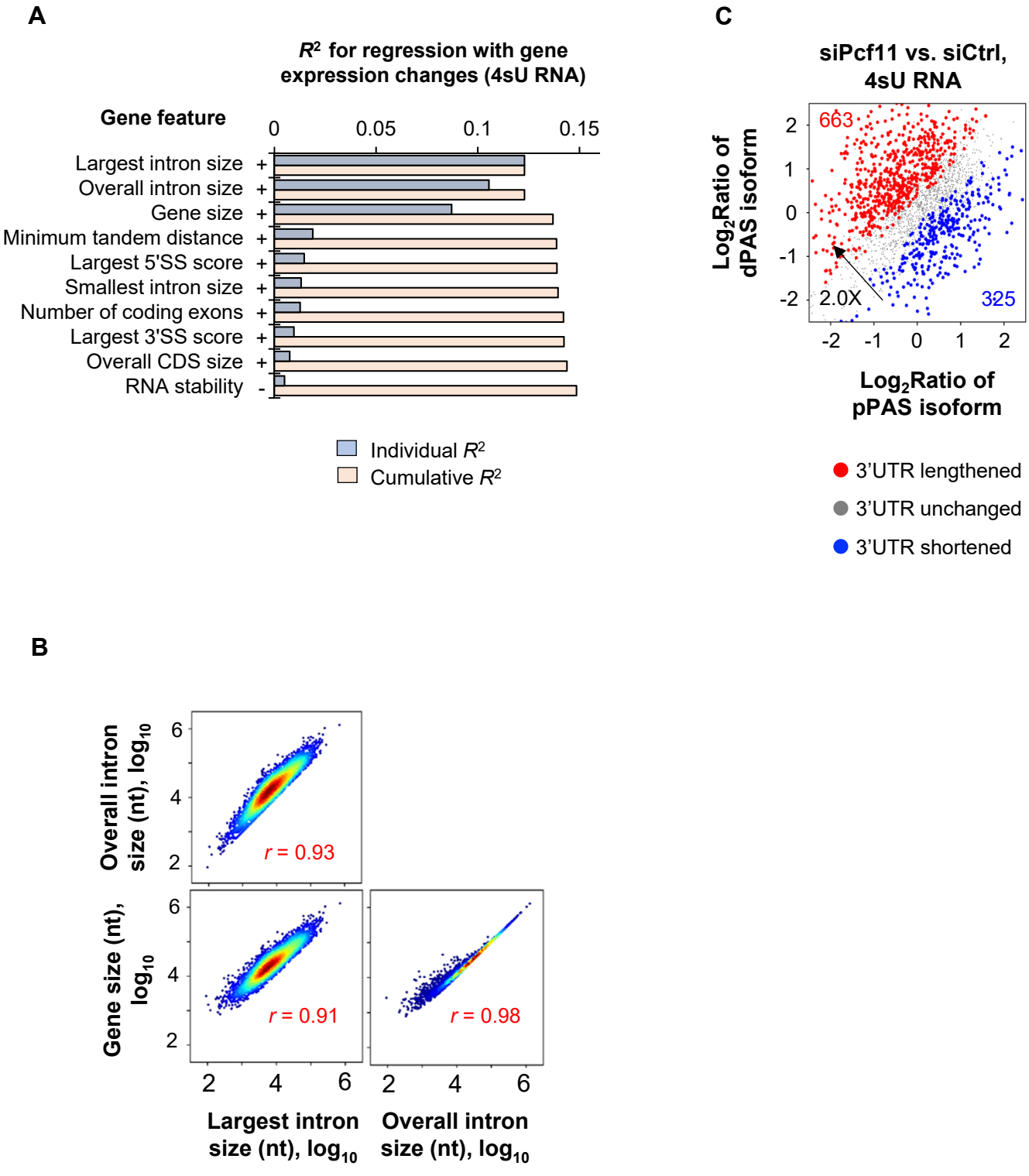

A
R^2 for regression with gene expression changes (4sU RNA)
Gene feature
0
0.05
0.1
0.15
Largest intron size +
Overall intron size +
Gene size +
Minimum tandem distance +
Largest 5'SS score +
Smallest intron size +
Number of coding exons +
Largest 3'SS score +
Overall CDS size +
RNA stability -
Individual R^2
Cumulative R^2
C
siPcf11 vs. siCtrl, 4sU RNA
Log_2Ratio of dPAS isoform
663
325
2.0X
Log_2Ratio of pPAS isoform
3'UTR lengthened
3'UTR unchanged
3'UTR shortened
B
Overall intron size (nt), log_{10}
Gene size (nt), log_{10}
r = 0.93
r = 0.91
r = 0.98
Largest intron size (nt), log_{10}
Overall intron size (nt), log_{10}

Figure S2, Related to Figure 2.

(A) Summary of regression analysis of different gene features vs. gene expression changes in *Pcf11* KD cells. Expression data are based on 4sU-labeled RNA. Top features are sorted according to individual R^2. Cumulative R^2 for a feature is based on the feature and all other features with a better individual R^2. '+', positive correlation; "-", negative correlation.

(B) Correlations among the size of largest intron, gene size and overall intron size. Pearson correlation coefficient (*r*) is indicated.

(C) 3'UTR APA changes in *Pcf11* KD cells (4sU-labeled RNA). The numbers of genes with significantly lengthened 3'UTRs (red) or shortened 3'UTRs (blue) are indicated.

Figure S3

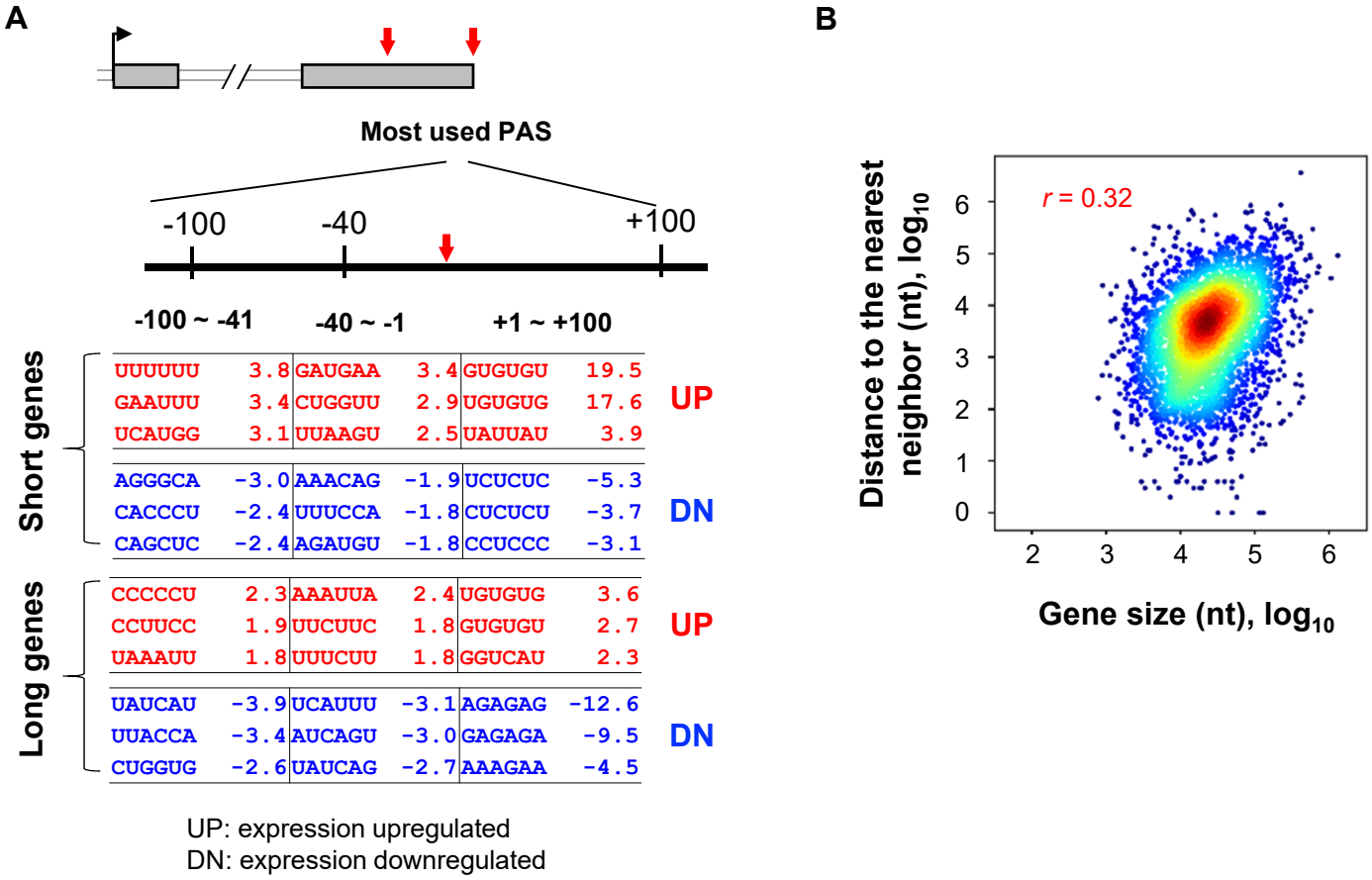

A
Most used PAS
-100
-40
+100
-100 ~ -41
-40 ~ -1
+1 ~ +100
Short genes
UUUUUU 3.8 GAUGAA 3.4 GUGUGU 19.5
GAAUUU 3.4 CUGGUU 2.9 UGUGUG 17.6
UCAUGG 3.1 UUAAGU 2.5 UAUUAU 3.9
UP
AGGGCA -3.0 AAACAG -1.9 UCUCUC -5.3
CACCCU -2.4 UUUCCA -1.8 CUCUCU -3.7
CAGCUC -2.4 AGAUGU -1.8 CCUCCC -3.1
DN
Long genes
CCCCCU 2.3 AAAUUA 2.4 UGUGUG 3.6
CCUUCC 1.9 UUCUUC 1.8 GUGUGU 2.7
UAAAUU 1.8 UUUCUU 1.8 GGUCAU 2.3
UP
UAUCAU -3.9 UCAUUU -3.1 AGAGAG -12.6
UUACCA -3.4 AUCAGU -3.0 GAGAGA -9.5
CUGGUG -2.6 UAUCAG -2.7 AAAGAA -4.5
DN
UP: expression upregulated
DN: expression downregulated
B
r = 0.32
Distance to the nearest neighbor (nt), log10
Gene size (nt), log10
0
1
2
3
4
5
6
C
Gene size
Distance to the nearest neighbor
Downregulated
Upregulated
-0.4 -0.2 0 0.2 0.4
Median log2Ratio,
siPcf11 vs. siCtrl

Figure S3, Related to Figure 3.

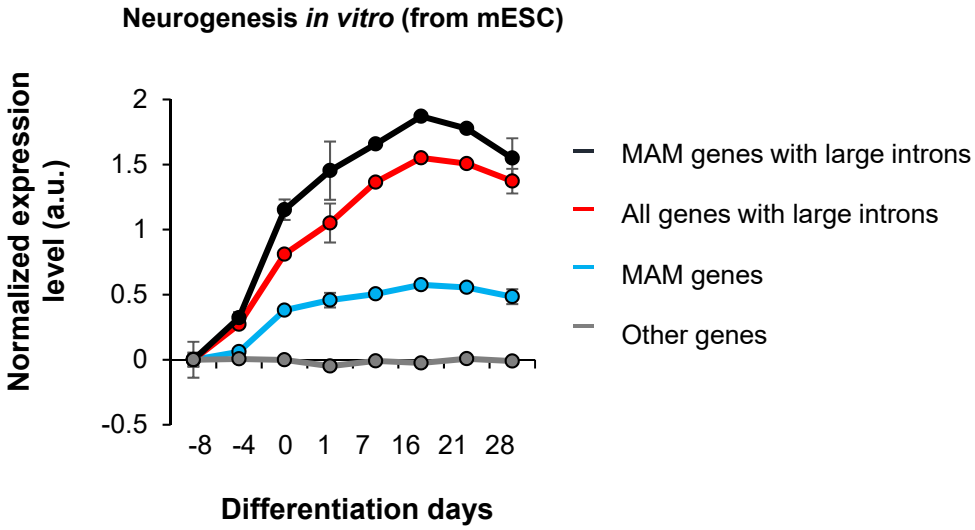
(A) Enriched motifs around the PAS for upregulated or downregulated genes in short and long gene groups. The most used PAS in the 3'-most exon of each gene based on 3'READS read count is used for analysis.

(B) Correlation between gene size and distance to the nearest neighbor. Pearson correlation coefficient (*r*) is indicated.

(C) Heatmap showing gene expression change (median $\log_2$Ratio, siPcf11 vs. siCtrl) in gene groups based on gene size and distance to the nearest neighbor.

Figure S4

A


Neurogenesis *in vitro* (from mESC)
Normalized expression level (a.u.)
2
1.5
1
0.5
0
-0.5
-8 -4 0 1 7 16 21 28
Differentiation days
MAM genes with large introns
All genes with large introns
MAM genes
Other genes

B

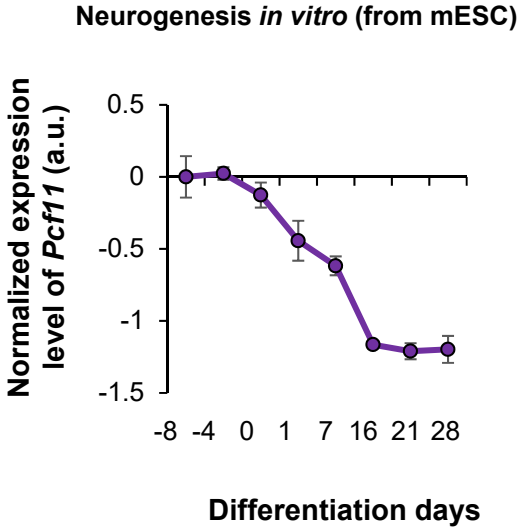

Neurogenesis *in vitro* (from mESC)
Normalized expression level of *Pcf11* (a.u.)
0.5
0
-0.5
-1
-1.5
-8 -4 0 1 7 16 21 28
Differentiation days

Figure S4, Related to Figure 4.

(A) Gene expression changes of different gene sets in neurogenesis (*in vitro* differentiation of mouse embryonic stem cells to mature neurons, based on SRP017778).

(B) *Pcf11* expression levels in neurogenesis as in (A).

Figure S5

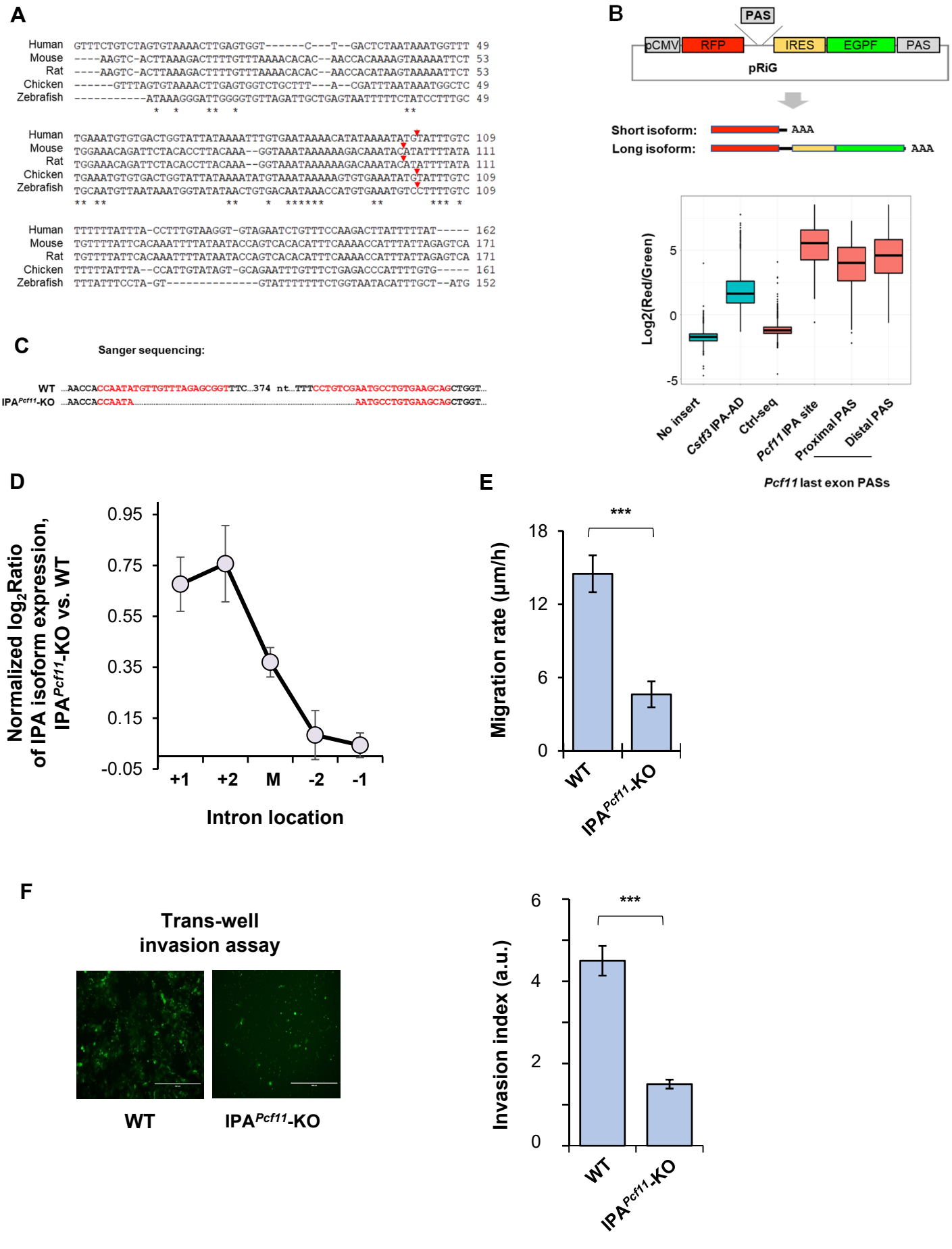

A
Human GTTTCTGTCTAGTGTAAAACTTGAGTGGT------C---T--GACTCTAATAAATGGTTT 49
Mouse ----AAGTC-ACTTAAAGACTTTTGTTTAAAACACAC--AACCACAAAAGTAAAAATTCT 53
Rat ----AAGTC-ACTTAAAGACTTTTGTTTAAAACACAC--AACCACATAAGTAAAAATTCT 53
Chicken ------GTTTAGTGTAAAACTTGAGTGGTCTGCTTT---A--CGATTTAATAAATGGCTC 49
Zebrafish ----------ATAAAGGGATTGGGGTGTTAGATTGCTGAGTAATTTTTCTATCCTTTGC 49
Human TGAAATGTGTGACTGGTATTATAAAATTTGTGAATAAAACATATAAAATATGTATTTGTC 109
Mouse TGGAAACAGATTCTACACCTTACAAA--GGTAAATAAAAAAGACAAATACATATTTTATA 111
Rat TGGAAACAGATTCTACACCTTACAAA--GGTAAATAAAAAAGACAAATACATATTTTATA 111
Chicken TGAAATGTGTGACTGGTATTATAAAATATGTAAATAAAAAGTGTGAAATATGTATTTGTC 109
Zebrafish TGCAATGTTAATAAATGGTATATAACTGTGACAATAAACCATGTGAAATGTCCTTTTGTC 109
Human TTTTTTATTTA-CCTTTGTAAGGT-GTAGAATCTGTTTCCAAGACTTATTTTTAT----- 162
Mouse TGTTTTATTCACAAATTTTATAATACCAGTCACACATTTCAAAACCATTTATTAGAGTCA 171
Rat TGTTTTATTCACAAATTTTATAATACCAGTCACACATTTCAAAACCATTTATTAGAGTCA 171
Chicken TTTTTATTTA--CCATTGTATAGT-GCAGAATTTGTTTCTGAGACCCATTTTGTG----- 161
Zebrafish TTTATTTCCTA-GT--------------GTATTTTTTTCTGGTAATACATTTGCT--ATG 152
B
PAS
pCMV
RFP
IRES
EGPF
PAS
pRiG
Short isoform:
AAA
Long isoform:
AAA
Log2(Red/Green)
5
0
-5
No insert
Cstf3 IPA-AD
Ctrl-seq
Pcf11 IPA site
Proximal PAS
Distal PAS
Pcf11 last exon PASs
C
Sanger sequencing:
WT ...AACCACCAATATGTTGTTTAGAGCGGTTTC...374 nt...TTTCCTGTCGAATGCCTGTGAAGCAGCTGGT...
IPAPcf11-KO ...AACCACCAATA.....AATGCCTGTGAAGCAGCTGGT...
D
Normalized log2Ratio of IPA isoform expression, IPAPcf11-KO vs. WT
0.95
0.75
0.55
0.35
0.15
-0.05
+1
+2
M
-2
-1
Intron location
E
***
Migration rate (μm/h)
18
12
6
0
WT
IPAPcf11-KO
F
Trans-well invasion assay
WT
IPAPcf11-KO
***
Invasion index (a.u.)
6
4
2
0
WT
IPAPcf11-KO

Figure S5, Related to Figure 6.

(A) Sequence alignment of the region around the IPA of *Pcf11* in five vertebrate species. Arrow indicates PAS annotated in PolyA_DB.

(B) Top, schematic of pRiG and two APA isoforms expressed (short and long isoforms). Bottom, PAS strength analyzed by log2(red fluorescence signal/green fluorescence signal) or log2(Red/Green). A high log2(Red/Green) value indicates a strong PAS. No insert is pRiG vector only. *Cstf3* IPA-AD is a mutated *Cstf3* IPA site, which is weak. Ctrl-seq is a random sequence from intron 1 of *Pcf11* inserted into pRiG. *Pcf11* proximal and distal PASs are two 3'UTR APA sites in the last exon of *Pcf11*. This result shows that the strength of IPA site is higher than those of PASs in the last exon.

(C) Sanger sequencing validation of the amplified PCR products from IPA^{Pcf11}-KO cells. sgRNA target sequences are highlighted in red.

(D) Intron location vs. IPA regulation. +1, +2, M, -2, -1 are first, second, middle, last but one, and last introns, respectively. Error bar is standard error of mean.

(E) Scratch assay analysis of cell migration.

(F) Trans-well invasion assay. Error bars in (E) & (F) are standard error of mean of ten randomly selected areas. ***, $P < 0.01$ (Wilcoxon test).

Figure S6

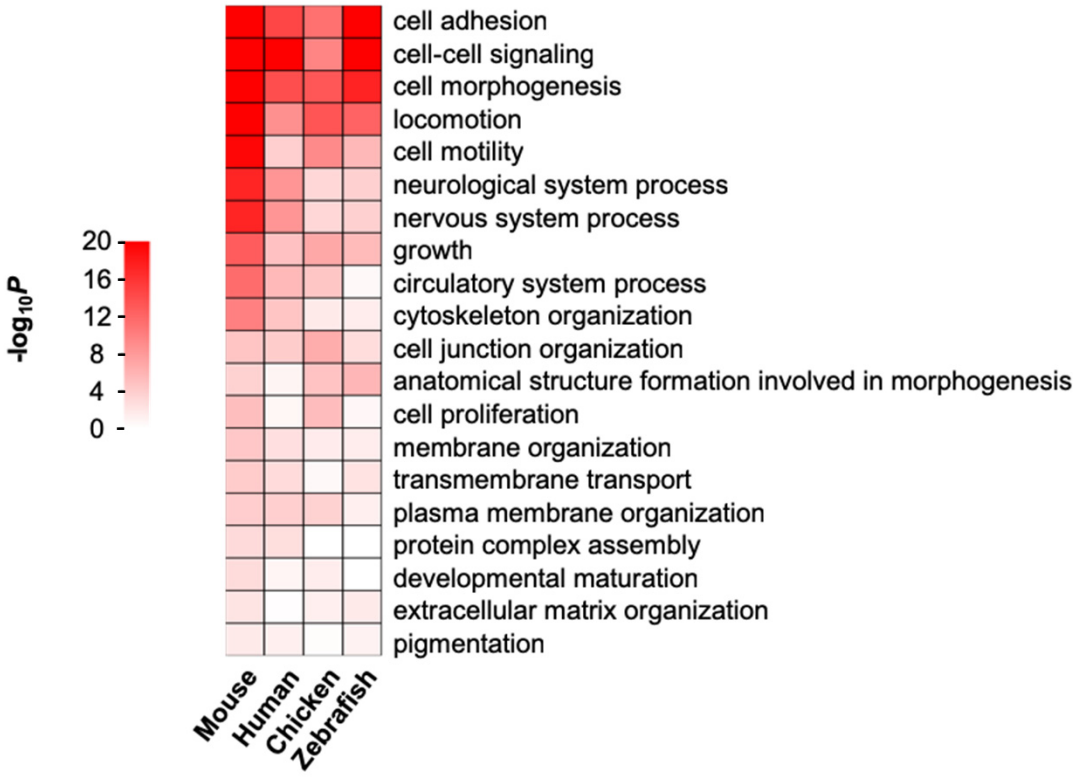
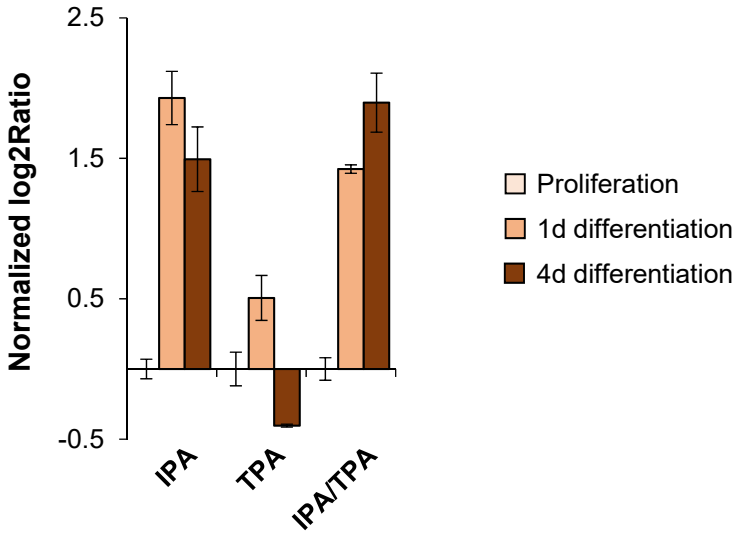

-log10P
20
16
12
8
4
0
cell adhesion
cell-cell signaling
cell morphogenesis
locomotion
cell motility
neurological system process
nervous system process
growth
circulatory system process
cytoskeleton organization
cell junction organization
anatomical structure formation involved in morphogenesis
cell proliferation
membrane organization
transmembrane transport
plasma membrane organization
protein complex assembly
developmental maturation
extracellular matrix organization
pigmentation
Mouse
Human
Chicken
Zebrafish

Figure S6, Related to Figure 7.

Top GO terms enriched for genes with large introns (top 10%) in several vertebrates. P-value (Fisher's exact test) is based on comparison with other genes in the genome.

Figure S7

A


2.5
1.5
0.5
-0.5
Normalized log2Ratio
IPA
TPA
IPA/TPA
Proliferation
1d differentiation
4d differentiation

B

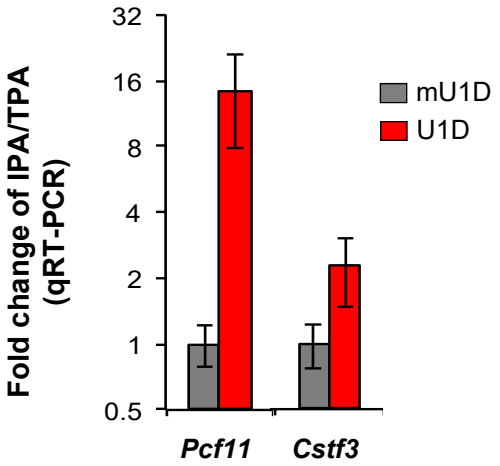

32
16
8
4
2
1
0.5
Fold change of IPA/TPA (qRT-PCR)
Pcf11
Cstf3
mU1D
U1D

Figure S7, Related to Figure 7.

(A) RT-qPCR analysis of IPA and TPA isoform expression during C2C12 differentiation.

(B) RT-qPCR analysis of IPA and TPA expression in C2C12 cells treated with U1D oligo to functionally inhibit U1. Both *Pcf11* and *Cstf3* IPA events were analyzed. mU1D, mutant U1D oligo (used as a control).