Coding Exon-Structure Aware Realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation

Virag Sharma[1,2], Anas Elghafari[1,2,3] and Michael Hiller[1,2*]

[1]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

[2]Max Planck Institute for the Physics of Complex Systems, Dresden, Germany

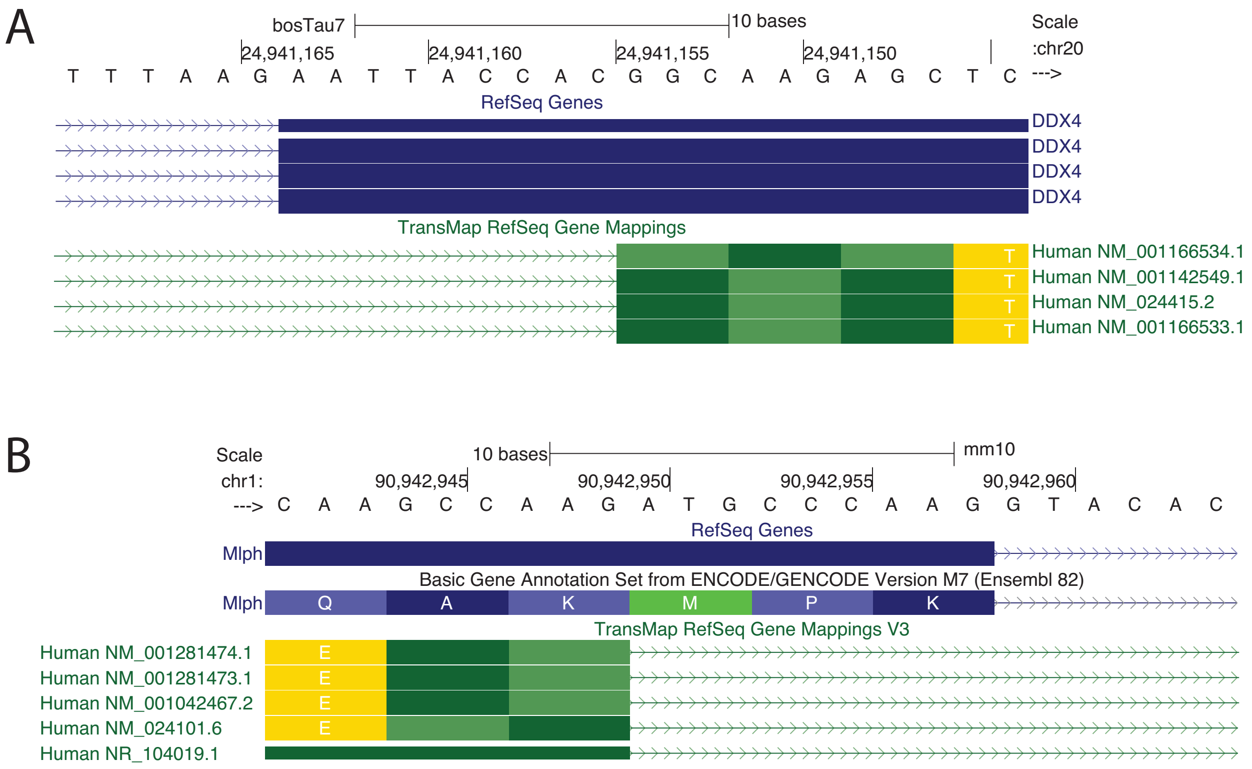
[3]Technical University, Dresden, Germany

*To whom correspondence should be addressed:
hiller@mpi-cbg.de

Supplementary Material

The Supplementary Material contains

- Figures S1 – S23
- Tables S1 – S5

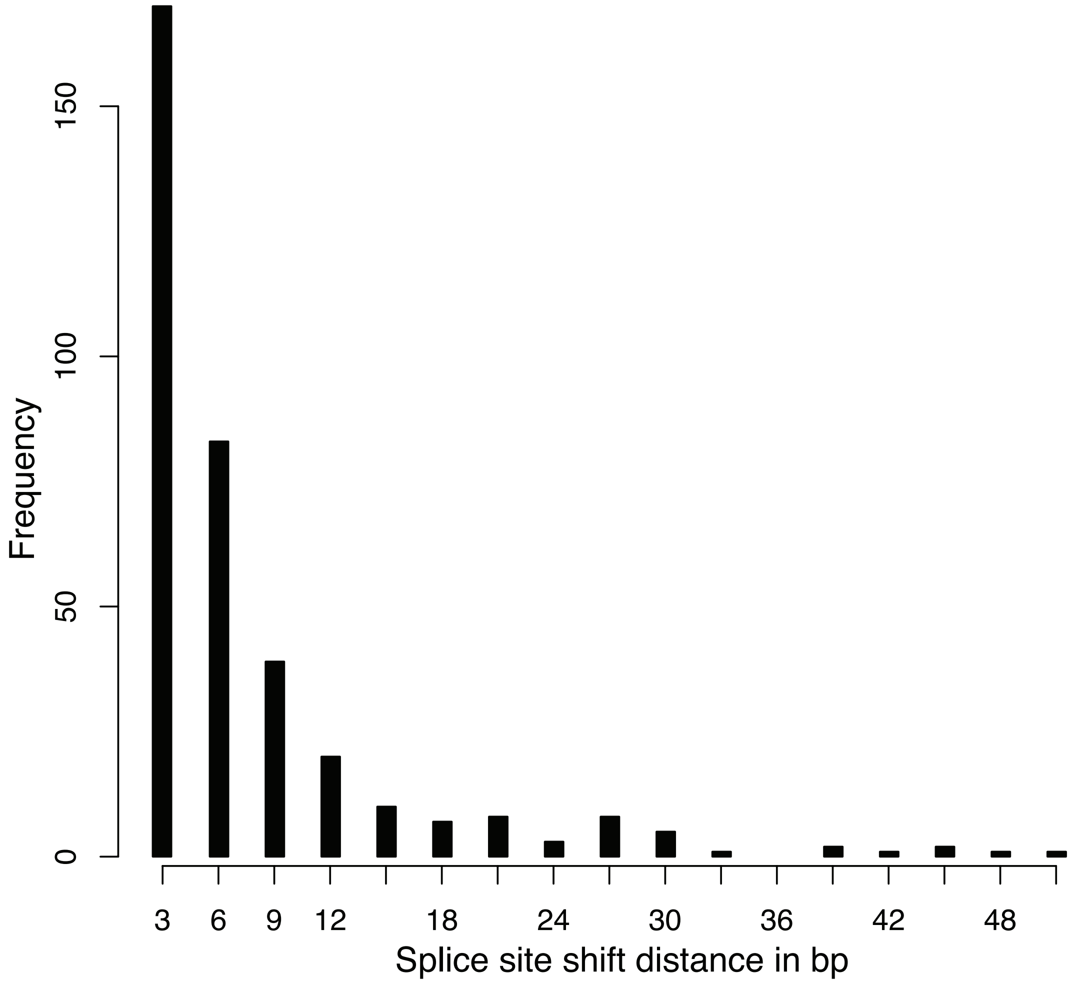


Supplementary Figure S1: Projecting the coordinates of aligned exons in a genome alignment does not identify the correct splice sites for both cases shown in Figure 1C main text.

(A) TransMap [1-3] annotation of the human *DDX4* gene in the cow genome. The acceptor splice site of cow *DDX4* is shifted by 9 bp upstream (cow RefSeq annotation). TransMap shows a non-consensus AC acceptor.

(B) TransMap annotation of the human *MLPH* gene in the mouse genome. The donor site is shifted by 9 bp downstream (mouse RefSeq and GENCODE annotation). TransMap shows a non-consensus AT donor.

1. Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH, Haussler D: **Comparative genomics search for losses of long-established genes on the human lineage.** *PLoS Comput Biol* 2007, **3:**e247.
2. Stanke M, Diekhans M, Baertsch R, Haussler D: **Using native and syntenically mapped cDNA alignments to improve de novo gene finding.** *Bioinformatics* 2008, **24:**637-644.
3. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, et al: **The UCSC Genome Browser Database: update 2009.** *Nucleic Acids Res* 2009, **37:**D755-761.



Supplementary Figure S2: Distances of splice site shifts.
Histogram of the distance between the human splice site and the shifted splice site is shown. The data are 360 exons where a splice site shift happened in the mouse, rat, dog or cow genome according to their RefSeq annotation.

A true alignment

```
GACGTTAGGAAGGCAGAGGAGGAGCTGGGTGAGCTGGAGGCTAAGCT
GACGTTAGGAAAGCAGAG---GAGCTGGGTGAGGTGGAGGCTAAGCT
```

reported alignment

```
GACGTTAGGAAGGCAGAGGAGGAGCTGGGTGAGCTGGAGGCTAAGCT
GACGTTAGGAAAGCA---GAGGAGCTGGGTGAGGTGGAGGCTAAGCT
```

B true alignment

```
ACCTGGAGAATGCACTTTTGATCAAGATGAATGTGCGTTTACACAGG
ACCTGGAGAATGCACTTTTGATCAAGAG---TGTGCATTTACACAGG
```

reported alignment

```
ACCTGGAGAATGCACTTTTGATCAAGATGAATGTGCGTTTACACAGG
ACCTGGAGAATGCACTTTTGATCAA---GAGTGTGCATTTACACAGG
```

C true alignment

```
GGAAGTAGGAGCTGAACAGACAGACTTTCTGCGAGGGCCATTAGAG
GGAAGTAGGAGCTGAA---------CTT---CGAGCGCCATTAGAG
```

reported alignment

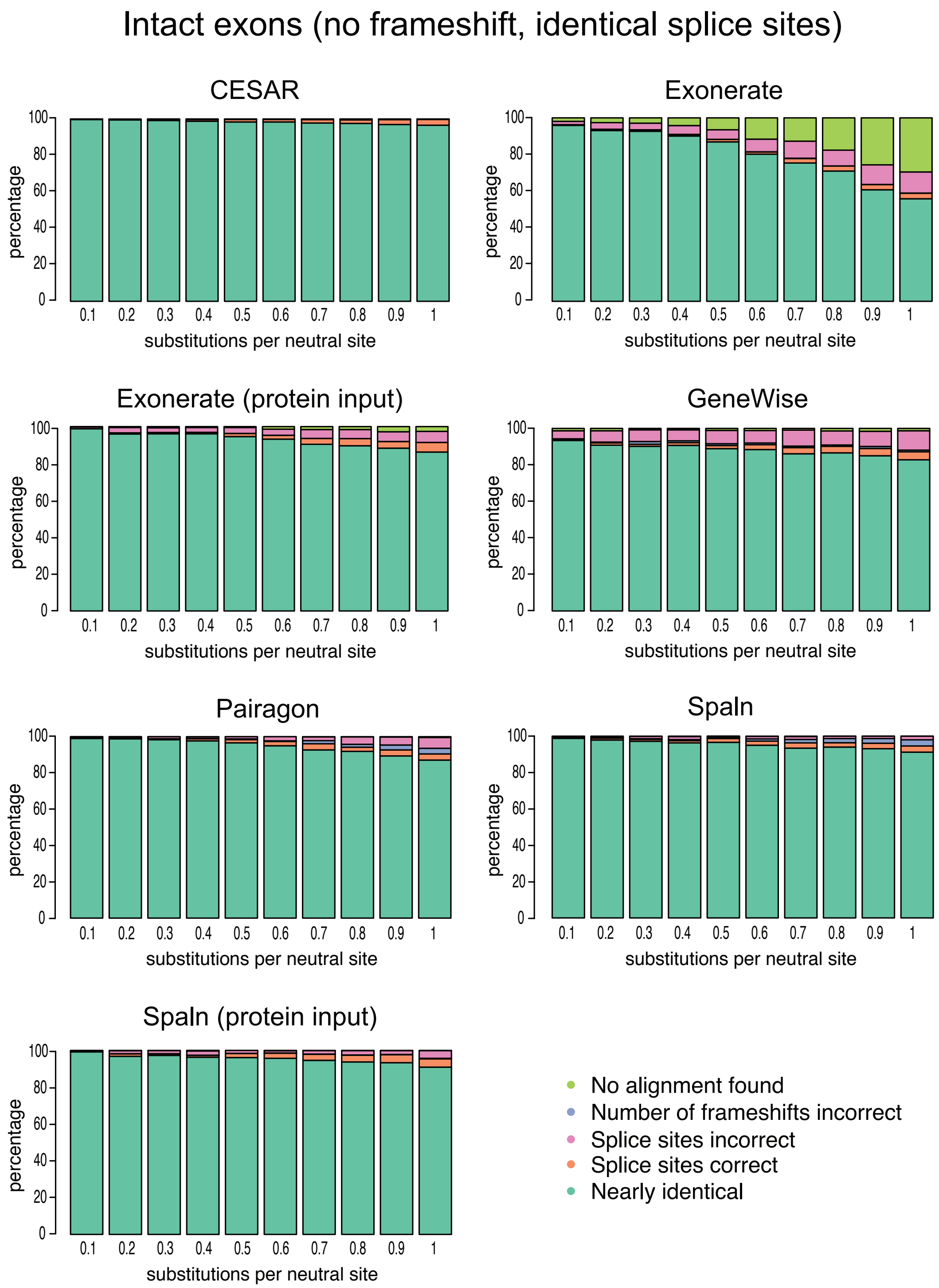
```
GGAAGTAGGAGCTGAACAGACAGACTTTCTGCGAGGGCCATTAGAG
GGAAGTAGGAGCTGAA------------CTTCGAGCGCCATTAGAG
```

Supplementary Figure S3: Alignment ambiguities and difficulties in locating the exact position of insertion/deletions.

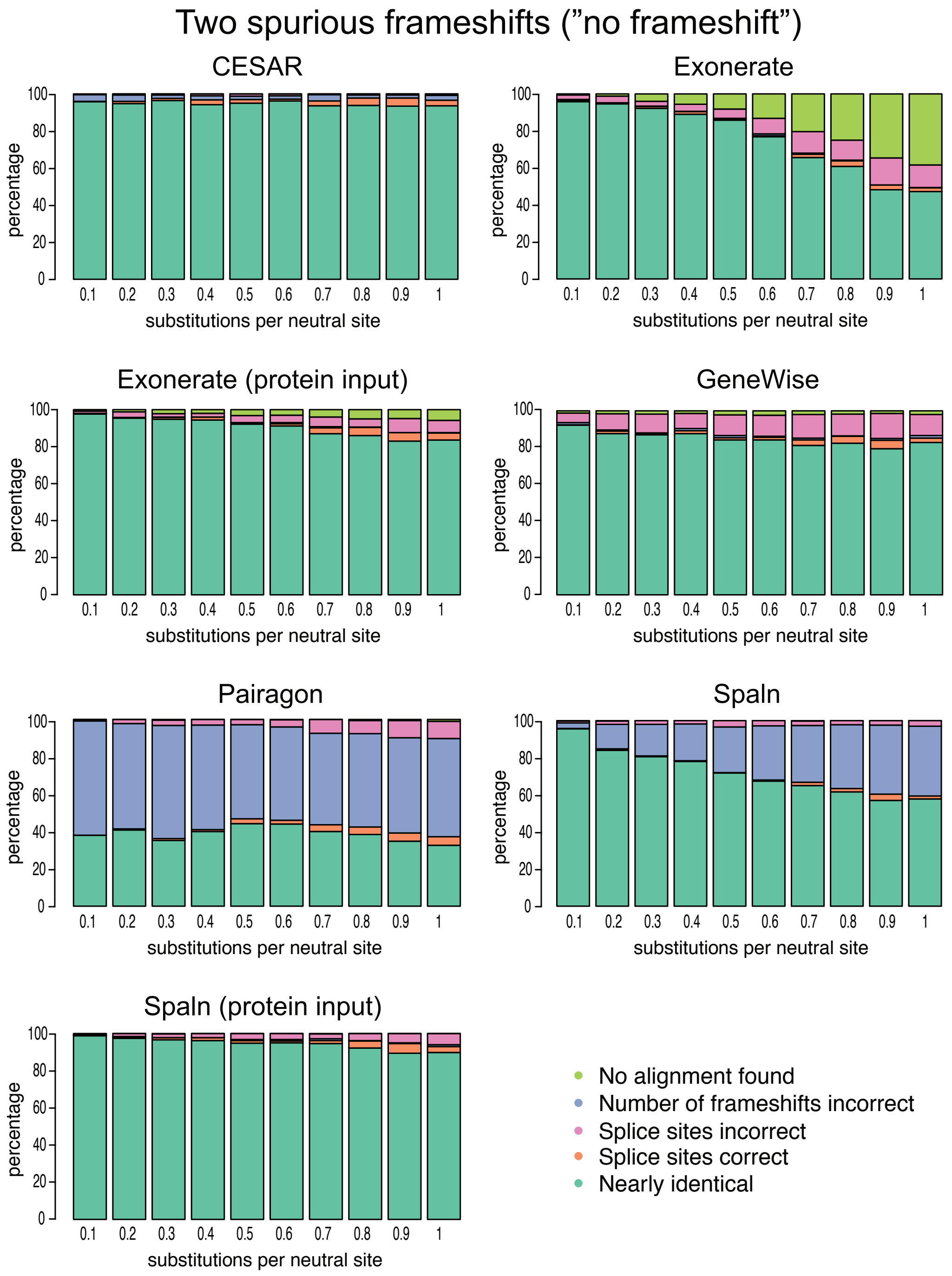
(A) There are 3 equivalent ways of placing the deleted GAG.

(B) Two equivalent ways of placing the 3 bp deletion. Both GAT or GAA in the reference can align to GAG in the query. (A) and (B) are regarded as "nearly identical" in Figure 4 main text and Supplementary Figures S4-S8.

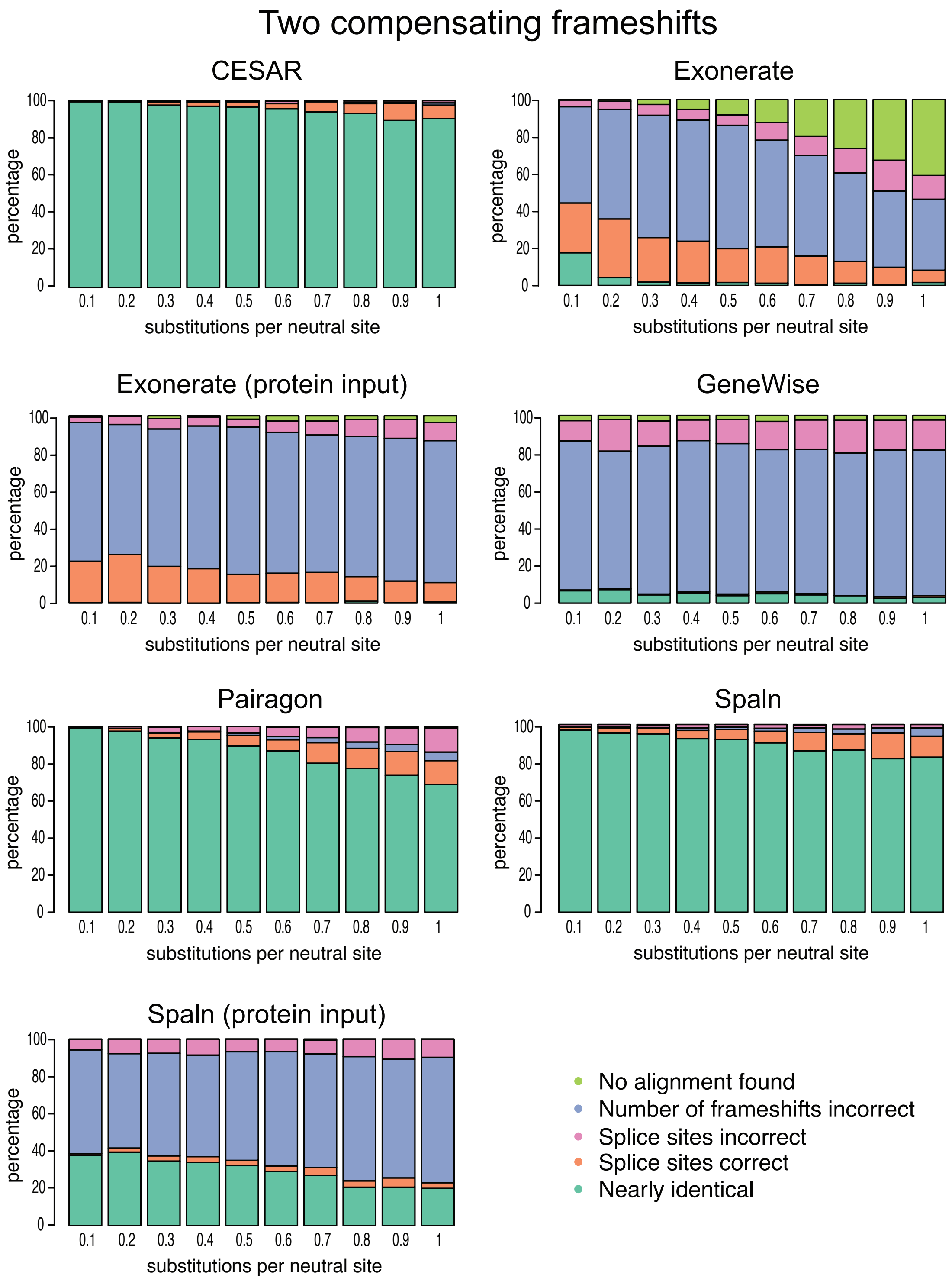
(C) A 9 bp and a nearby 3 bp deletion in the true alignment are reported as a single 12 bp deletion. This alignment is regarded as "splice sites correct" in Supplementary Figures S4-S8 if the splice sites are correctly identified.



Supplementary Figure S4: Detailed breakdown of differences between the reported and true alignment for intact exons. Intact exons are defined as exons with identical splice sites and without any frameshift. “Nearly identical” alignments are alignments that are either identical to the true alignment or differ from the true alignments only in the position of an indel that we allow to be shifted by at most 6 bases up- or downstream. “Splice sites correct” are alignments where both splice sites are correctly aligned and the correct number of frameshifts (0 here) is reported but indel positions are shifted by more than 6 bp or a different number of indels is reported.

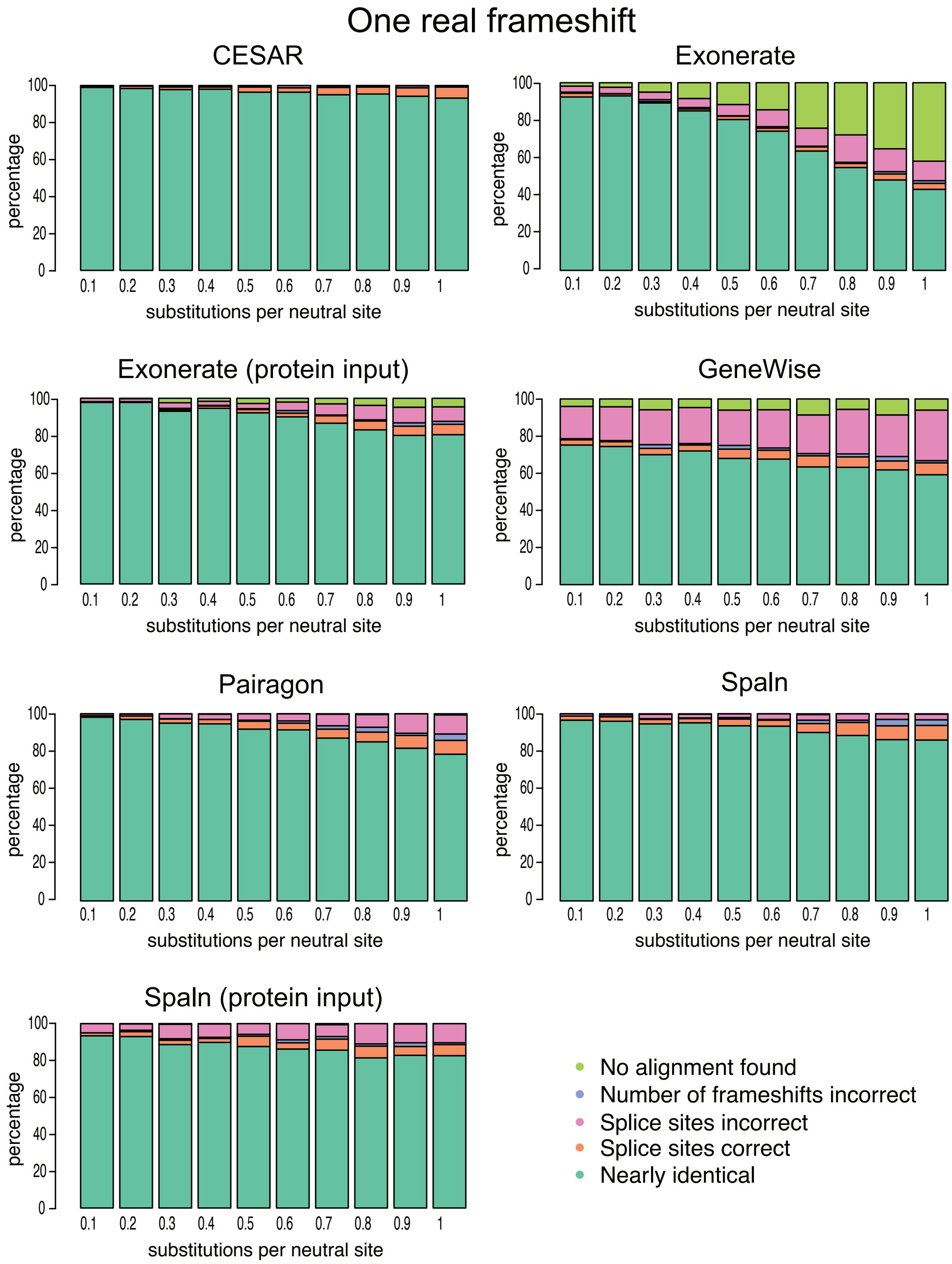


Supplementary Figure S5: Detailed breakdown of differences between the reported and true alignment for exons from the “no-frameshift” dataset. This dataset mimics the numerous cases of two nearby compensating frameshifts that we observed in genome alignments. Given that a single frameshift inactivates an exon, an alternative alignment with no frameshifts but a few codon substitutions is more plausible. In this dataset, we introduced two compensating frameshifting indels that are separated by 6 to 12 bp. Since these two close compensating frameshifts result affect only 2 to 4 codons in an otherwise intact exon, we regard them as spurious and an aligner that is aware of the reading frame should not report any frameshift. The true alignment is therefore the alignment that does not have any frameshift.



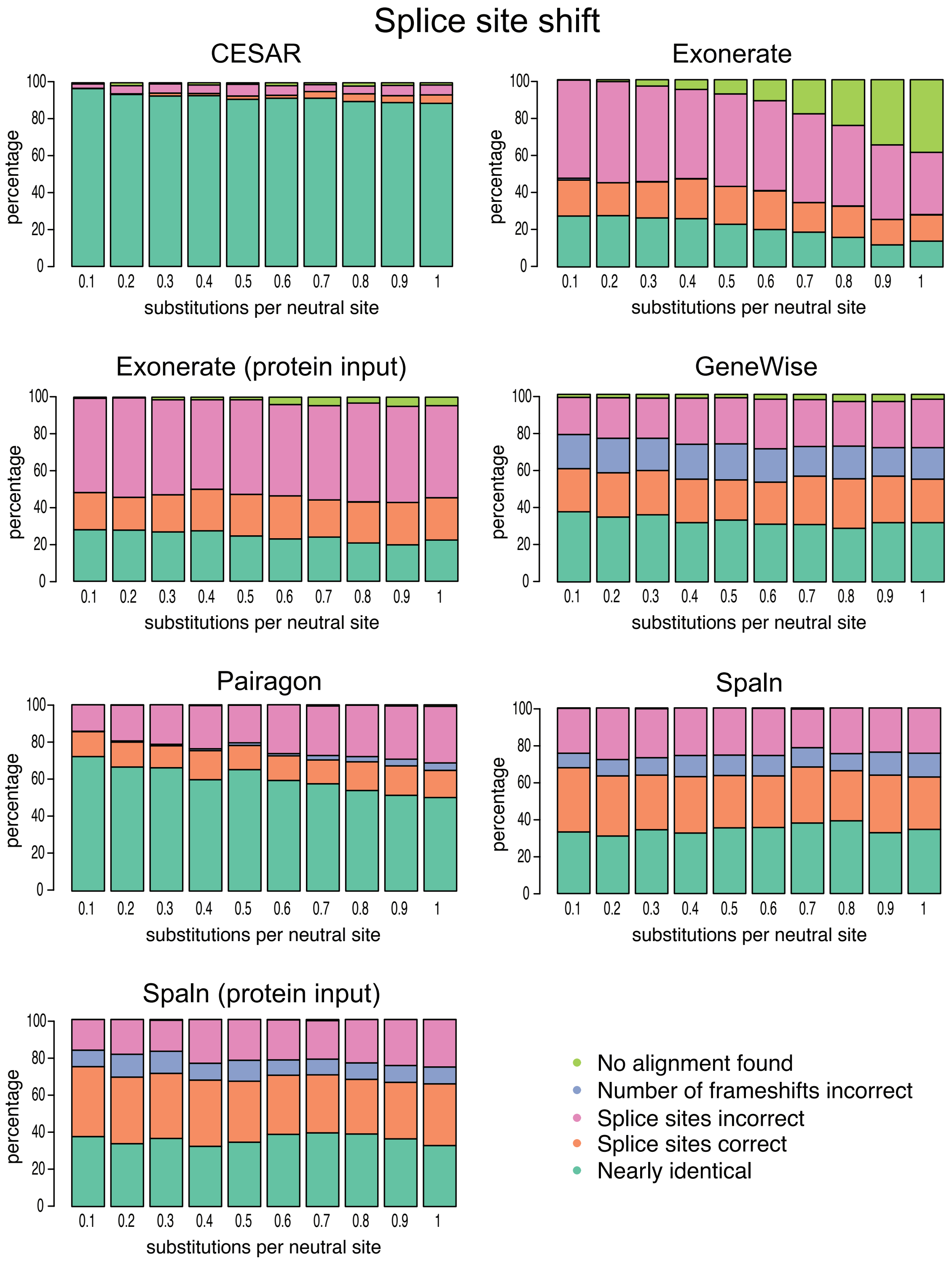
Supplementary Figure S6: Detailed breakdown of differences between the reported and true alignment for exons that have two real compensating frameshifts.

This dataset tests if methods report frameshifts that most likely did occur in evolution. In this dataset, we introduced two compensating frameshifting indels that are separated by a large distance (30 to 45 bp). The true alignment has exactly two frameshifts.



Supplementary Figure S7: Detailed breakdown of differences between the reported and true alignment for exons that have exactly one frameshift.

This dataset represents exons that are really inactivated by a frameshift and the true alignment has exactly one frameshift. This dataset also tests if methods avoid frameshifts by any means, which would result in incorrectly inferring exon conservation for exons that are not conserved.



Supplementary Figure S8: Detailed breakdown of differences between the reported and true alignment for exons where splice site shifts occurred.
In this dataset, one splice site was shifted by a distance obtained by sampling from the distribution of real splice site shifts. The true alignment here has no frameshift and the shifted splice site is aligned to the original splice site.

A hg19: chr17:72,527,497-72,527,586

human-mouse alignment with two compensating frameshifting indels

```
CAAGCCA---GAGCTCAGGCAGAACTTCCAGAGTGCATCTGGGATCTGCATTTGCCACTGGTTGCAGATC-AGGCGGACGAGGAGCCGGGAAGG
| || ||   ||| ||   |||  |||   |  | |  ||| ||| || ||||||  |||| ||   ||| ||| |  |    |   |||||||
CGAGTCAGCTGAGATC-TCCAGGGCTTTGGGTATCCCGCTGAGATTTGGATTTGCTGCTGGCTGTTCATCAAGGAGTGCAGCAAAGTGGGAAGG
```

alignment without frameshifting indels

```
CAAGCCA---GAGCTCAGGCAGAACTTCCAGAGTGCATCTGGGATCTGCATTTGCCACTGGTTGCAGATCAGGCGGACGAGGAGCCGGGAAGG
| || ||   ||| ||     |   ||    |   |     |  |  |  ||       | |        ||| |  |    |   |||||||
CGAGTCAGCTGAGATCTCCAGGGCTTTGGGTATCCCGCTGAGATTTGGATTTGCTGCTGGCTGTTCATCAAGGAGTGCAGCAAAGTGGGAAGG
```

B hg19: chr11:76,751,532-76,751,619

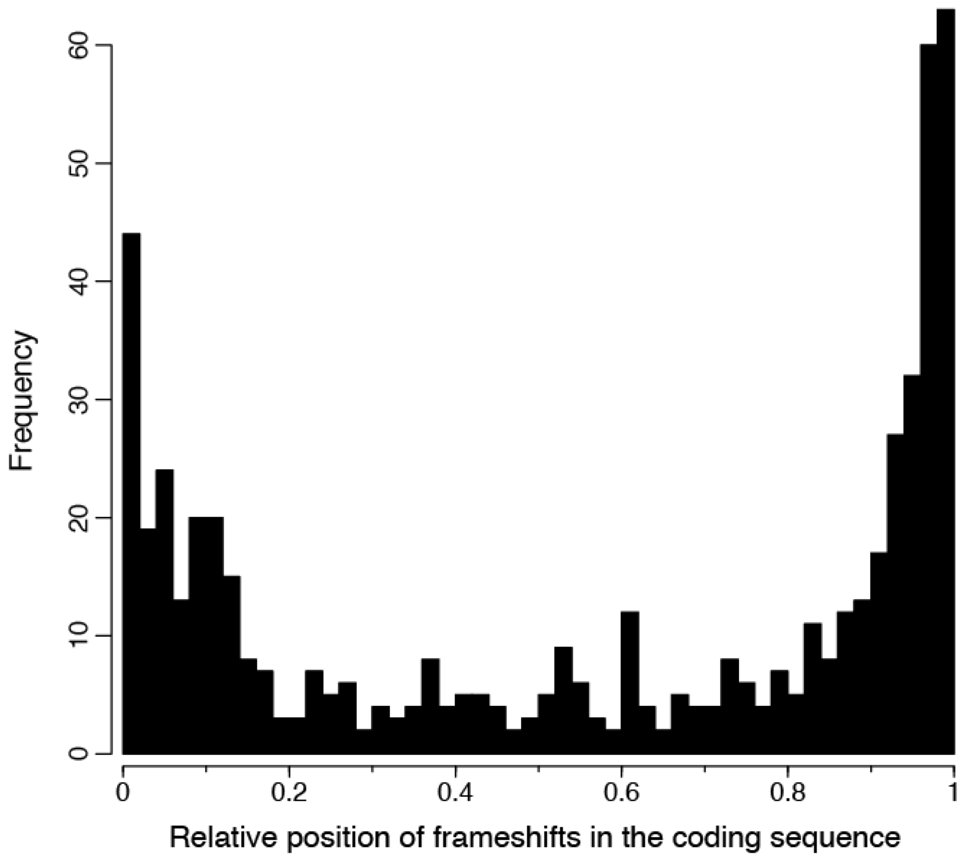
human-mouse alignment with three compensating frameshifting deletions

```
GCTGTGCGCCAGGCAAGCTGCACGCCGAAGGGCCAGGATGCCCTCGTGGCCGCTGGGCGCCAGGCCGGCGCGCTCCAAGACACATGCC
||||  |    ||||   ||||||||||||||||  |||||||| ||||| ||||||||||||||||||  |||   |||||||||||
GCTGCACATTGGGCAGC-TGCACGCCGAAGGGCCG-GATGCCCTGGTGGCTGCTGGGCGCCAGGCCGGCCTGCTG-CAGACACATGCC
```

alignment without frameshifting deletions but a 3 base pair deletion

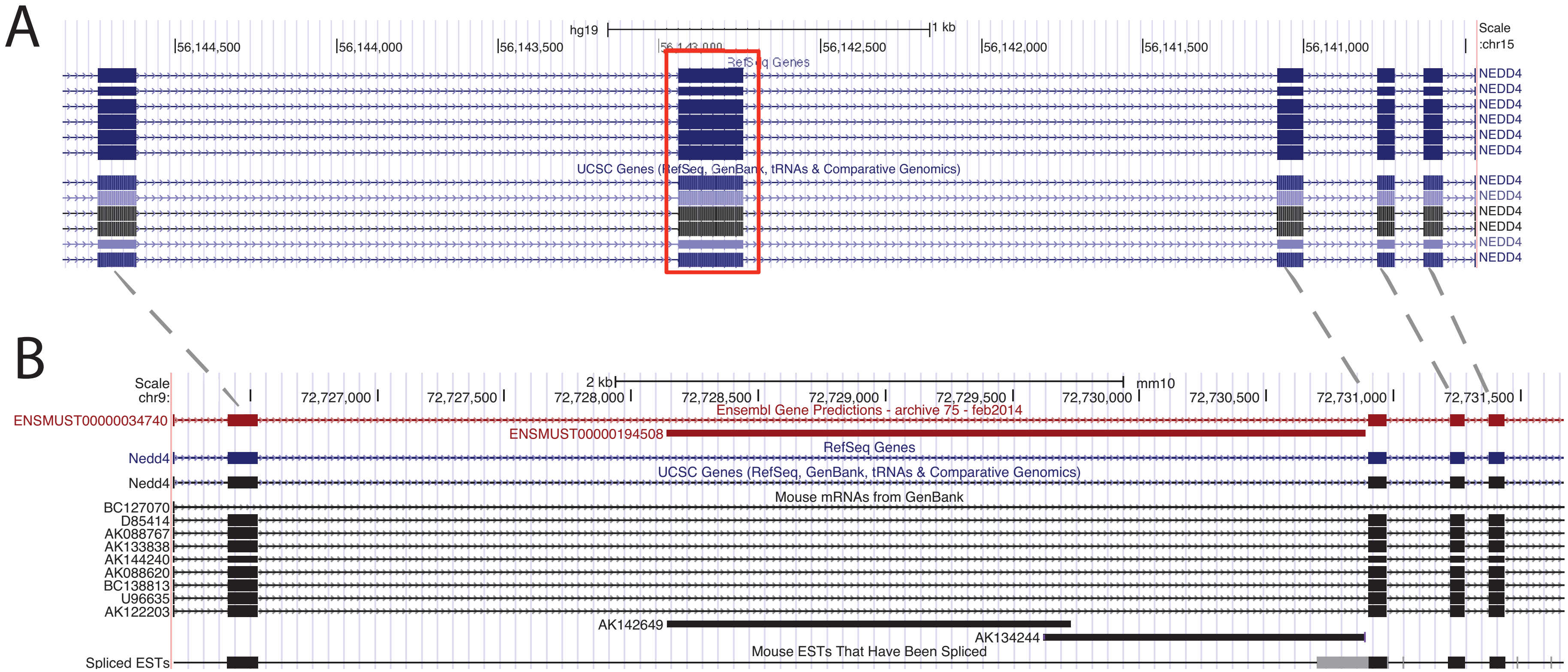
```
GCTGTGCGCCAGGCAAGCTGCACGCCGAAGGGCCAGGATGCCCTCGTGGCCGCTGGGCGCCAGGCCGGCGCGCTCCAAGACACATGCC
||||  |    ||||         |  | || |       |    |    |   | ||           ||      |||||||||||
GCTGCACATTGGGCAGCTGCACGCCGAAGGGCCGGATGCCCTGGTGGCTGCTGGGCGCCAGGCCGGCCTGCTG---CAGACACATGCC
```

Supplementary Figure S9: Examples of real compensating frameshifts.
Two (A) and three (B) frameshifts compensate each other and restore the original reading frame. Note that the sequence similarity is substantially lower without these frameshifts, strongly suggesting that these frameshifts did happen in evolution. The frameshifted part is shown in blue, the frameshifts are shown in red.

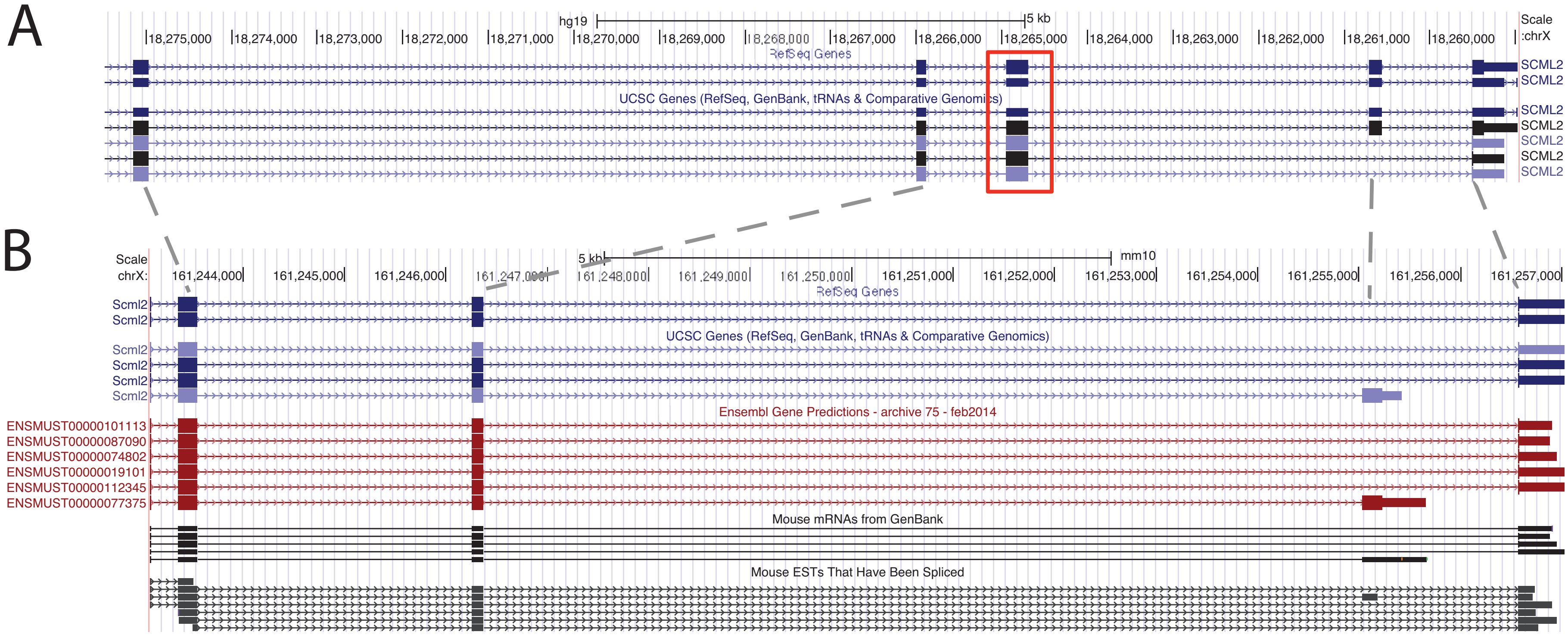


Supplementary Figure S10: Relative position of frameshifts in the mouse coding sequence after realigning with CESAR.

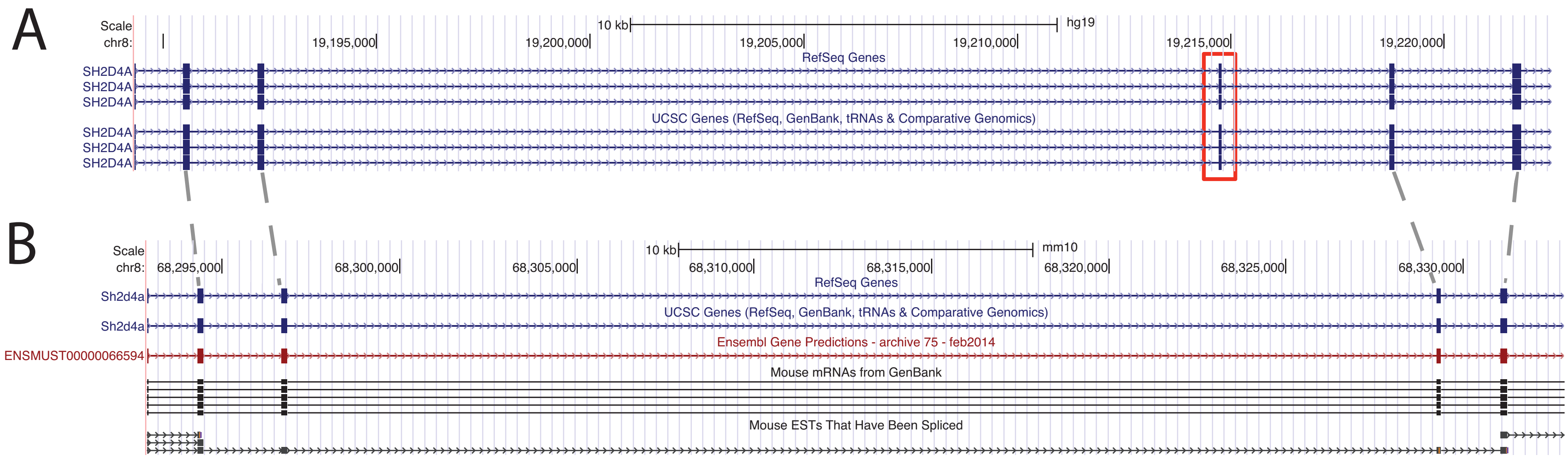
The histogram shows the relative position of 567 frameshifts that we detect in 149,331 realigned exons in mouse.



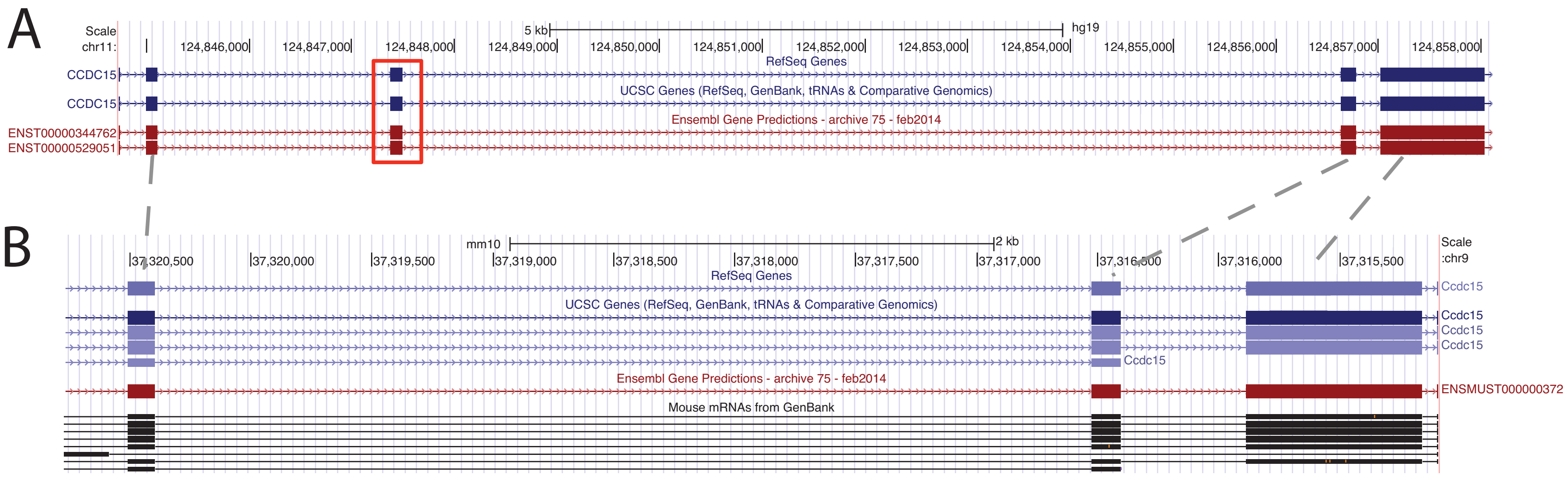
Supplementary Figure S11: Non-conserved exon in mouse *NEDD4*.
(A) Human genome browser: CESAR reports a frameshift in the highlighted exon in mouse.
(B) Mouse RefSeq, UCSC, Ensembl and MGC gene annotations and several mRNAs and ESTs show that this exon does not exist in mouse. Grey dashed lines indicate orthologous exons.



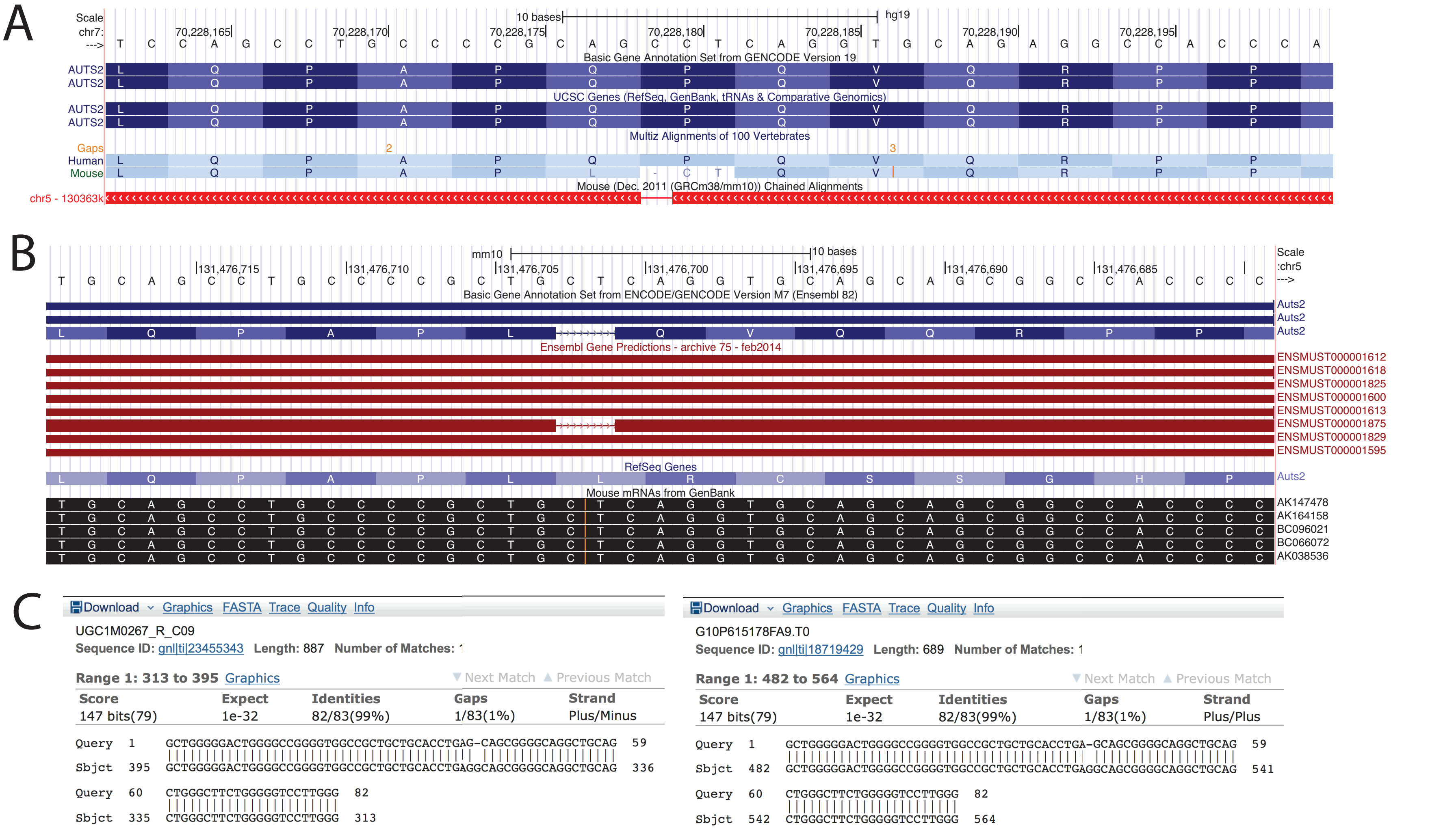
Supplementary Figure S12: Non-conserved exon in mouse *SCML2*.
(A) Human genome browser: CESAR reports a frameshift in the highlighted exon in mouse.
(B) Mouse RefSeq, UCSC and Ensembl gene annotations and mRNAs/ESTs show that this exon does not exist in mouse.
Grey dashed lines indicate orthologous exons.



Supplementary Figure S13: Non-conserved exon in mouse *SH2D4A*.
(A) Human genome browser: CESAR reports an 11 bp frameshift in the highlighted exon in mouse.
(B) Mouse RefSeq, UCSC and Ensembl gene annotations and mRNAs/ESTs show that this exon does not exist in mouse.
Grey dashed lines indicate orthologous exons.



Supplementary Figure S14: Non-conserved exon in mouse *CCDC15*.
(A) Human genome browser: CESAR reports both frameshifts and a splice site mutation in the highlighted exon in mouse.
(B) Mouse RefSeq, UCSC and Ensembl gene annotations and mRNAs show that this exon does not exist in mouse.
Grey dashed lines indicate orthologous exons.



Download Graphics FASTA Trace Quality Info

UGC1M0267_R_C09

Sequence ID: gnl|ti|23455343 **Length:** 887 **Number of Matches:** 1

Range 1: 313 to 395 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
147 bits(79)	1e-32	82/83(99%)	1/83(1%)	Plus/Minus

```
Query  1    GCTGGGGACTGGGGCCGGGTGGCCGCTGCTGCACCTGAG-CAGCGGGGCAGGCTGCAG  59
            ||||||||||||||||||||||||||||||||||||||| ||||||||||||||||||
Sbjct  395  GCTGGGGACTGGGGCCGGGTGGCCGCTGCTGCACCTGAGGCAGCGGGGCAGGCTGCAG  336

Query  60   CTGGGCTTCTGGGGGTCCTTGGG  82
            |||||||||||||||||||||||
Sbjct  335  CTGGGCTTCTGGGGGTCCTTGGG  313
```

Download Graphics FASTA Trace Quality Info

G10P615178FA9.T0

Sequence ID: gnl|ti|18719429 **Length:** 689 **Number of Matches:** 1

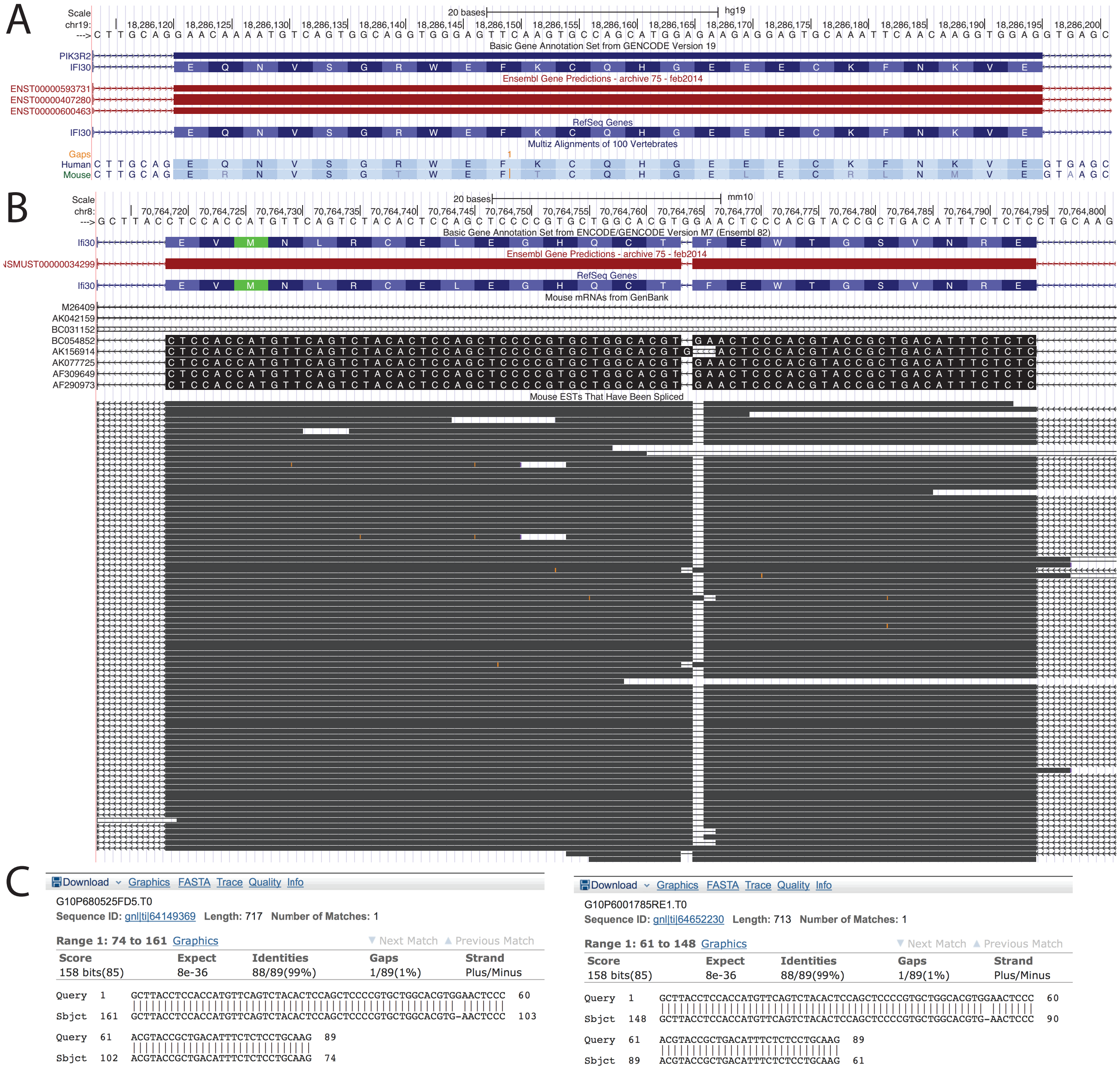
Range 1: 482 to 564 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
147 bits(79)	1e-32	82/83(99%)	1/83(1%)	Plus/Plus

```
Query  1    GCTGGGGACTGGGGCCGGGTGGCCGCTGCTGCACCTGA-GCAGCGGGGCAGGCTGCAG  59
            |||||||||||||||||||||||||||||||||||||| |||||||||||||||||||
Sbjct  482  GCTGGGGACTGGGGCCGGGTGGCCGCTGCTGCACCTGAGGCAGCGGGGCAGGCTGCAG  541

Query  60   CTGGGCTTCTGGGGGTCCTTGGG  82
            |||||||||||||||||||||||
Sbjct  542  CTGGGCTTCTGGGGGTCCTTGGG  564
```

Supplementary Figure S15: Assembly error in *AUTS2*.

(A) The human genome browser shows a 1 bp frameshifting deletion in *AUTS2* in mouse, visible in the multiple genome alignment and the pairwise alignment chain. (B) This 1 bp deletion is an assembly error in mouse. In the mouse genome browser, the GENCODE and Ensembl gene annotation show a 2 bp codon, which misses the single base. All four mRNAs that align to this locus have the base that is missing in the reference genome (orange tick mark). The RefSeq gene annotation is not aware of this assembly error and translates *AUTS2* in a different reading frame that leads to a premature stop codon at the end of this exon. (C) All seven aligning Sanger sequencing reads from the NCBI trace archive have the missing base. The screenshot shows two aligning reads.


A
Scale
chr19:
20 bases
hg19
Basic Gene Annotation Set from GENCODE Version 19
PIK3R2
IFI30
Ensembl Gene Predictions - archive 75 - feb2014
ENST00000593731
ENST00000407280
ENST00000600463
RefSeq Genes
IFI30
Multiz Alignments of 100 Vertebrates
Gaps
Human
Mouse
B
Scale
chr8:
20 bases
mm10
Basic Gene Annotation Set from ENCODE/GENCODE Version M7 (Ensembl 82)
Ifi30
Ensembl Gene Predictions - archive 75 - feb2014
NSMUST00000034299
RefSeq Genes
Ifi30
Mouse mRNAs from GenBank
M26409
AK042159
BC031152
BC054852
AK156914
AK077725
AF309649
AF290973
Mouse ESTs That Have Been Spliced
C

Download Graphics FASTA Trace Quality Info

G10P680525FD5.T0

Sequence ID: gnl|ti|64149369 **Length:** 717 **Number of Matches:** 1

Range 1: 74 to 161 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
158 bits(85)	8e-36	88/89(99%)	1/89(1%)	Plus/Minus

```
Query  1    GCTTACCTCCACCATGTTCAGTCTACACTCCAGCTCCCCGTGCTGGCACGTGGAACTCCC  60
            |||||||||||||||||||||||||||||||||||||||||||||||||||| |||||||
Sbjct  161  GCTTACCTCCACCATGTTCAGTCTACACTCCAGCTCCCCGTGCTGGCACGTG-AACTCCC  103

Query  61   ACGTACCGCTGACATTTCTCTCCTGCAAG  89
            |||||||||||||||||||||||||||||
Sbjct  102  ACGTACCGCTGACATTTCTCTCCTGCAAG  74
```

Download Graphics FASTA Trace Quality Info

G10P6001785RE1.T0

Sequence ID: gnl|ti|64652230 **Length:** 713 **Number of Matches:** 1

Range 1: 61 to 148 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
158 bits(85)	8e-36	88/89(99%)	1/89(1%)	Plus/Minus

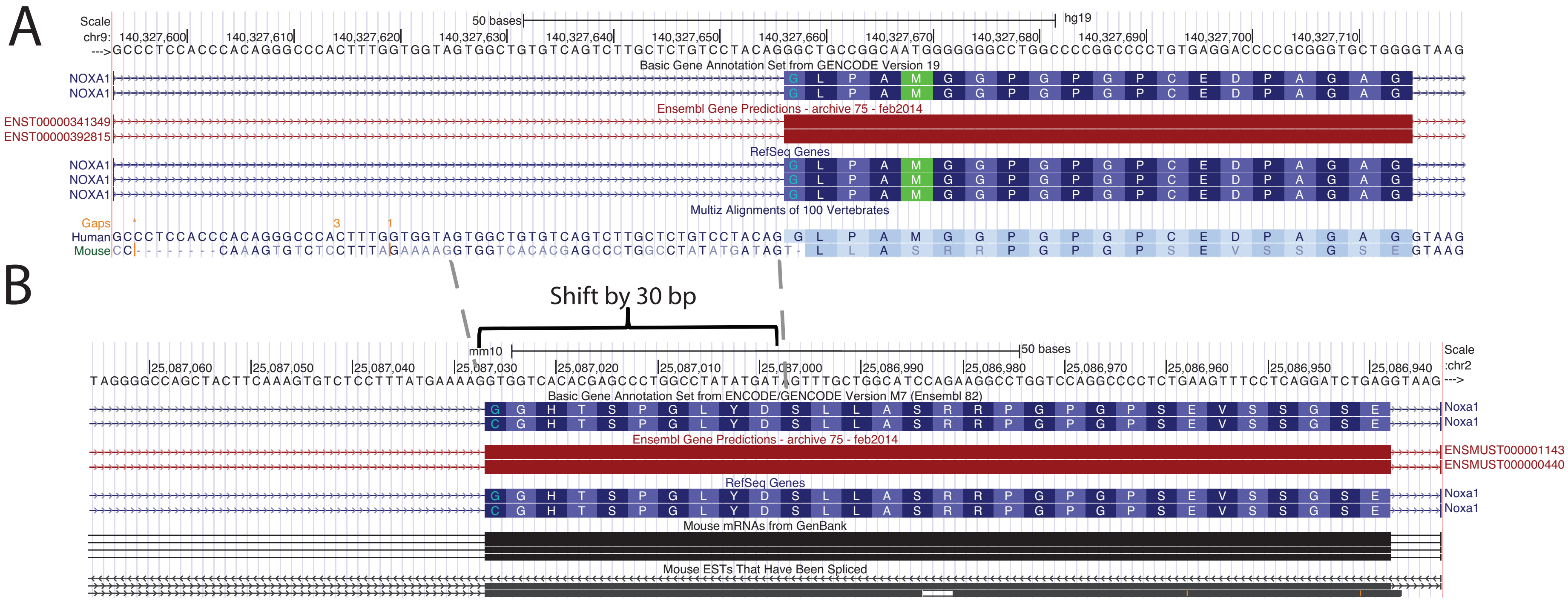
```
Query  1    GCTTACCTCCACCATGTTCAGTCTACACTCCAGCTCCCCGTGCTGGCACGTGGAACTCCC  60
            |||||||||||||||||||||||||||||||||||||||||||||||||||| |||||||
Sbjct  148  GCTTACCTCCACCATGTTCAGTCTACACTCCAGCTCCCCGTGCTGGCACGTG-AACTCCC  90

Query  61   ACGTACCGCTGACATTTCTCTCCTGCAAG  89
            |||||||||||||||||||||||||||||
Sbjct  89   ACGTACCGCTGACATTTCTCTCCTGCAAG  61
```

Supplementary Figure S16: Assembly error in *IFI30*.

(A) The human genome browser shows a 1 bp insertion (orange tick mark) in an exon of *IFI30* in mouse.

(B) Mouse genome browser: This insertion is an assembly error as the annotated reading frame simply ignores this insertion and all aligning mRNAs and ESTs do not have this extra base.

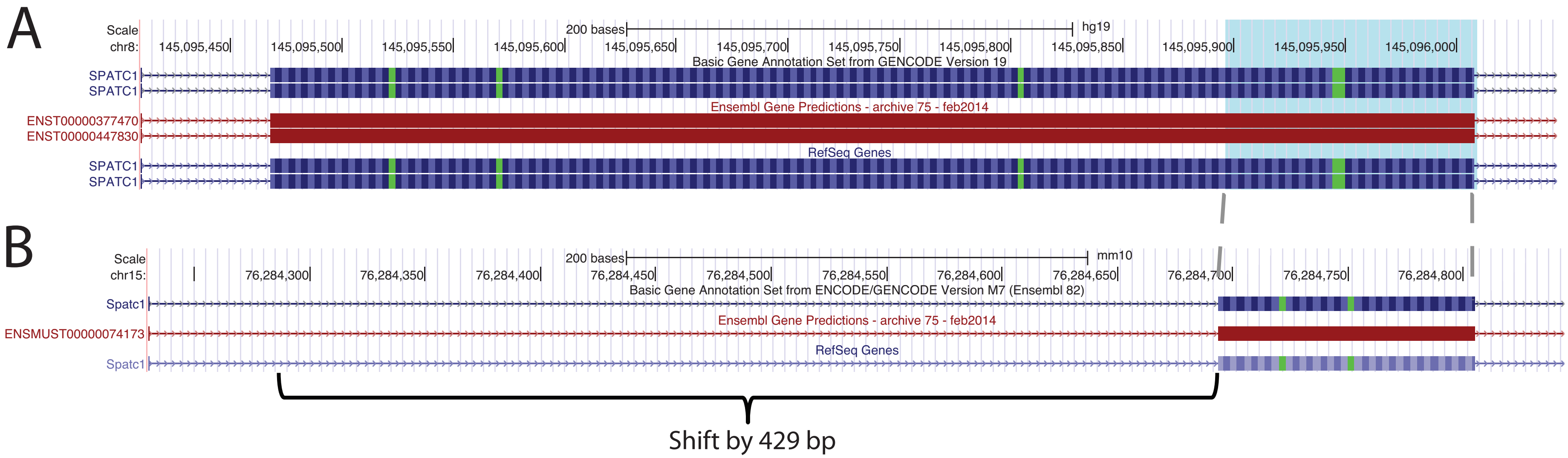
(C) All 11 Sanger reads from the NCBI trace archive that align to this region do not have the extra base. The screenshot shows two aligning reads.


A
Scale
50 bases
hg19
chr9: 140,327,600 140,327,610 140,327,620 140,327,630 140,327,640 140,327,650 140,327,660 140,327,670 140,327,680 140,327,690 140,327,700 140,327,710
---> GCCCTCCACCCACAGGGCCCACTTTGGTGGTAGTGGCTGTGTCAGTCTTGCTCTGTCCTACAGGGCTGCCGGCAATGGGGGGGCCTGGCCCCGGCCCCTGTGAGGACCCCGCGGGTGCTGGGGTAAG
Basic Gene Annotation Set from GENCODE Version 19
NOXA1
NOXA1
G L P A M G G P G P G P C E D P A G A G
Ensembl Gene Predictions - archive 75 - feb2014
ENST00000341349
ENST00000392815
RefSeq Genes
NOXA1
NOXA1
NOXA1
Multiz Alignments of 100 Vertebrates
Gaps * 3 1
Human GCCCTCCACCCACAGGGCCCACTTTGGTGGTAGTGGCTGTGTCAGTCTTGCTCTGTCCTACAG G L P A M G G P G P G P C E D P A G A G GTAAG
Mouse CC- - - - - - - -CAAAGTGTCTCCTTTAGAAAAGGTGGTCACACGAGCCCTGGCCTATATGATAGT- L L A S R R P G P G P S E V S S G S E GTAAG
B
Shift by 30 bp
mm10
50 bases
25,087,060 25,087,050 25,087,040 25,087,030 25,087,020 25,087,010 25,087,000 25,086,990 25,086,980 25,086,970 25,086,960 25,086,950 25,086,940
TAGGGGCCAGCTACTTCAAAGTGTCTCCTTTATGAAAAGGTGGTCACACGAGCCCTGGCCTATATGATAGTTTGCTGGCATCCAGAAGGCCTGGTCCAGGCCCCTCTGAAGTTTCCTCAGGATCTGAGGTAAG --->
Scale
:chr2
Basic Gene Annotation Set from ENCODE/GENCODE Version M7 (Ensembl 82)
G G H T S P G L Y D S L L A S R R P G P G P S E V S S G S E
C G H T S P G L Y D S L L A S R R P G P G P S E V S S G S E
Noxa1
Noxa1
Ensembl Gene Predictions - archive 75 - feb2014
ENSMUST000001143
ENSMUST000000440
RefSeq Genes
Noxa1
Noxa1
Mouse mRNAs from GenBank
Mouse ESTs That Have Been Spliced

Supplementary Figure S17: Splice site shift in mouse *NOXA1*.

(A) The genome alignment of this orthologous mouse exon has a frameshifting 1 bp deletion. Our re-alignment reports a slightly different alignment without the frameshift but with a TA acceptor site.

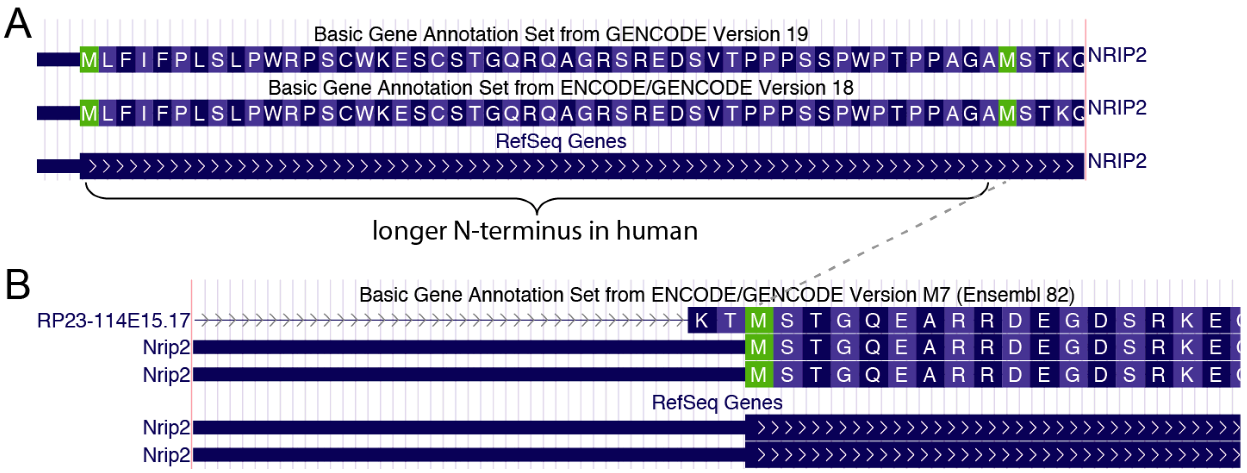
(B) The mouse genome browser show that the splice site has shifted by 30 bp upstream and is relatively weak (AAG) with a short polypyrimidine tract. The long distance and weak splice site explain why CESAR was not able to align this splice site.



Supplementary Figure S18: Splice site shift in mouse *SPATC1*.

(A) The human genome browser shows that this 540 bp exon of *SPATC1* has a splice site mutation in mouse. The mouse exon corresponds to the downstream part highlighted in light blue.

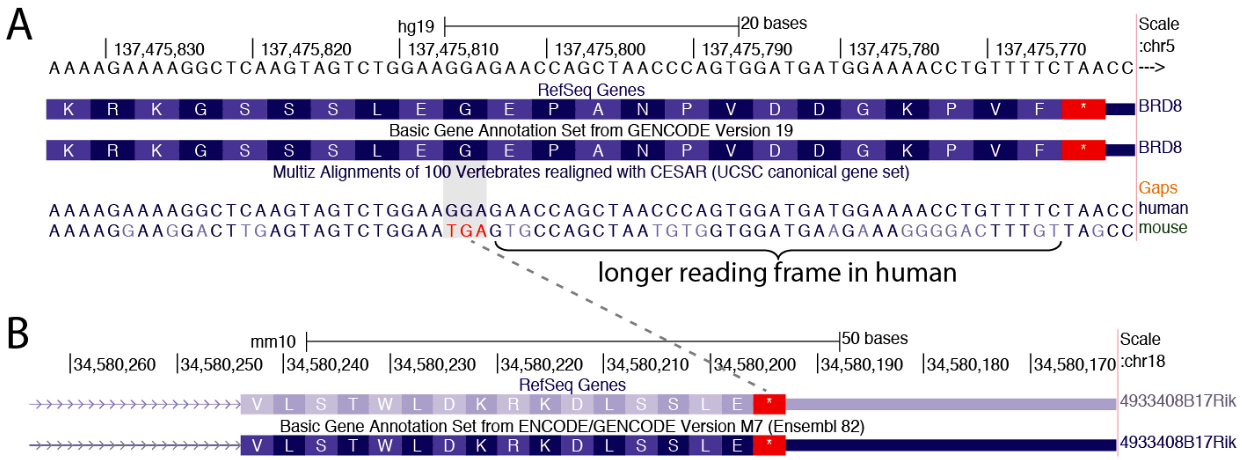
(B) Mouse genome browser shows that mouse has a considerably shorter 111 bp exon. That means the acceptor site that has shifted by 429 bp. The long distance explains why CESAR was not able to align this splice site.


A
Basic Gene Annotation Set from GENCODE Version 19
MLFIFPLSLPWRPSCWKESCSTGQRQAGRSREDSVTPPPSSPWPTPPAGAMSTKQ
NRIP2
Basic Gene Annotation Set from ENCODE/GENCODE Version 18
MLFIFPLSLPWRPSCWKESCSTGQRQAGRSREDSVTPPPSSPWPTPPAGAMSTKQ
NRIP2
RefSeq Genes
NRIP2
longer N-terminus in human
B
Basic Gene Annotation Set from ENCODE/GENCODE Version M7 (Ensembl 82)
RP23-114E15.17
KTMSTGQEARRDEGDSRKE
Nrip2
MSTGQEARRDEGDSRKE
Nrip2
MSTGQEARRDEGDSRKE
RefSeq Genes
Nrip2
Nrip2

Supplementary Figure S19: Start codon shift in mouse *NRIP2*.

(A) The human genome browser (hg19) shows that GENCODE/RefSeq annotation of the first coding exon of *NRIP2*.

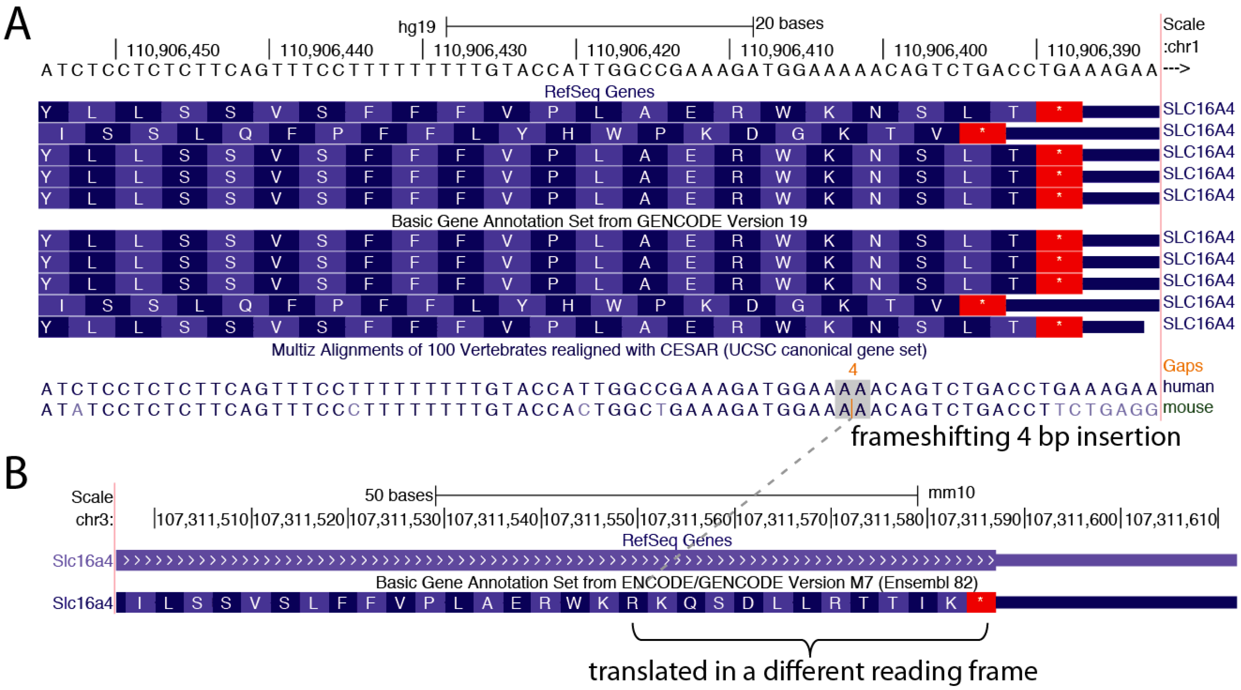
(B) The GENCODE/RefSeq annotations in the mouse show that a downstream ATG codon is used as the start codon in mouse (mm10). Compared to human, the N-terminus of Nrip2 in mouse is 51 amino acids shorter.


A
hg19
20 bases
Scale
:chr5
137,475,830
137,475,820
137,475,810
137,475,800
137,475,790
137,475,780
137,475,770
AAAAGAAAAGGCTCAAGTAGTCTGGAAGGAGAACCAGCTAACCCAGTGGATGATGGAAAACCTGTTTTCTAACC --->
RefSeq Genes
K R K G S S S L E G E P A N P V D D G K P V F *
BRD8
Basic Gene Annotation Set from GENCODE Version 19
K R K G S S S L E G E P A N P V D D G K P V F *
BRD8
Multiz Alignments of 100 Vertebrates realigned with CESAR (UCSC canonical gene set)
Gaps
AAAAGAAAAGGCTCAAGTAGTCTGGAAGGAGAACCAGCTAACCCAGTGGATGATGGAAAACCTGTTTTCTAACC human
AAAAGGAAGGACTTGAGTAGTCTGGAATGAGTGCCAGCTAATGTGGTGGATGAAGAAAGGGGACTTTGTTAGCC mouse
longer reading frame in human
B
mm10
50 bases
Scale
:chr18
34,580,260
34,580,250
34,580,240
34,580,230
34,580,220
34,580,210
34,580,200
34,580,190
34,580,180
34,580,170
RefSeq Genes
V L S T W L D K R K D L S S L E *
4933408B17Rik
Basic Gene Annotation Set from ENCODE/GENCODE Version M7 (Ensembl 82)
V L S T W L D K R K D L S S L E *
4933408B17Rik

Supplementary Figure S20: Stop codon shift in mouse *BRD8*.

(A) The human genome browser shows the last coding exon of *BRD8*. The alignment shows a stop codon mutation in mouse (GGA → TGA).

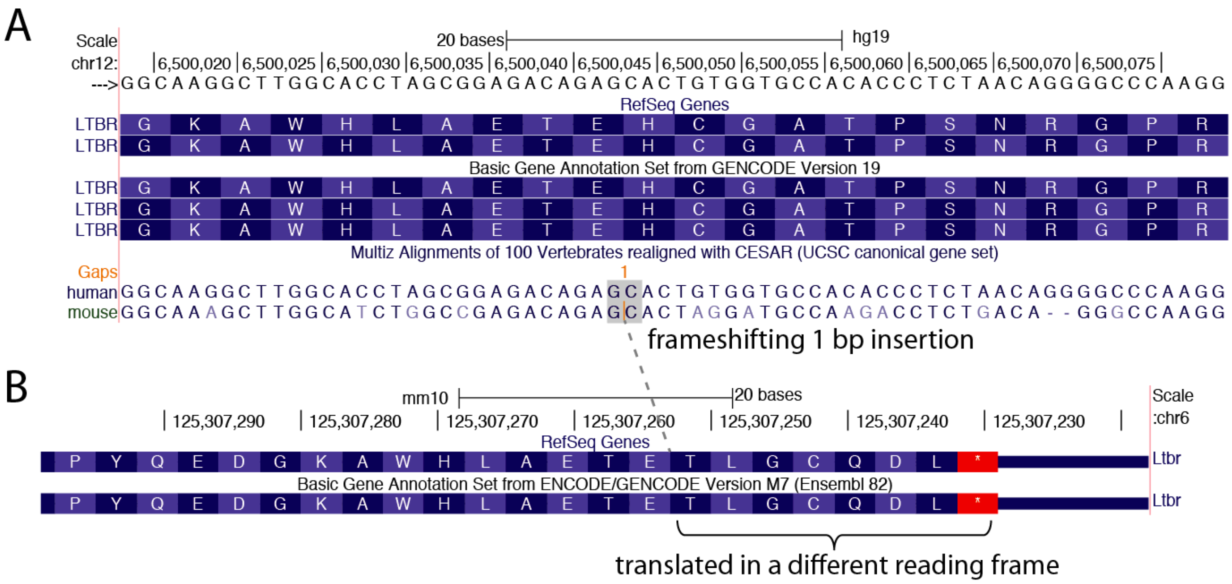
(B) The mouse genome browser shows the position of the annotated stop codon. Compared to human, the mouse Brd8 C-terminus is 14 amino acids shorter.


A
hg19
20 bases
Scale
chr1
110,906,450
110,906,440
110,906,430
110,906,420
110,906,410
110,906,400
110,906,390
ATCTCCTCTCTTCAGTTTCCTTTTTTTTTTGTACCATTGGCCGAAAGATGGAAAAACAGTCTGACCTGAAAGAA --->
RefSeq Genes
Y L L S S V S F F F V P L A E R W K N S L T *
I S S L Q F P F F L Y H W P K D G K T V *
SLC16A4
Basic Gene Annotation Set from GENCODE Version 19
Multiz Alignments of 100 Vertebrates realigned with CESAR (UCSC canonical gene set)
Gaps
4
human
mouse
ATCTCCTCTCTTCAGTTTCCTTTTTTTTTTGTACCATTGGCCGAAAGATGGAAAAACAGTCTGACCTGAAAGAA
ATATCCTCTCTTCAGTTTCCCTTTTTTTTTGTACCACTGGCTGAAAGATGGAAAAACAGTCTGACCTTCTGAGG
frameshifting 4 bp insertion
B
Scale
chr3:
50 bases
mm10
107,311,510
107,311,520
107,311,530
107,311,540
107,311,550
107,311,560
107,311,570
107,311,580
107,311,590
107,311,600
107,311,610
RefSeq Genes
Slc16a4
Basic Gene Annotation Set from ENCODE/GENCODE Version M7 (Ensembl 82)
I L S S V S L F F V P L A E R W K R K Q S D L L R T T I K *
translated in a different reading frame

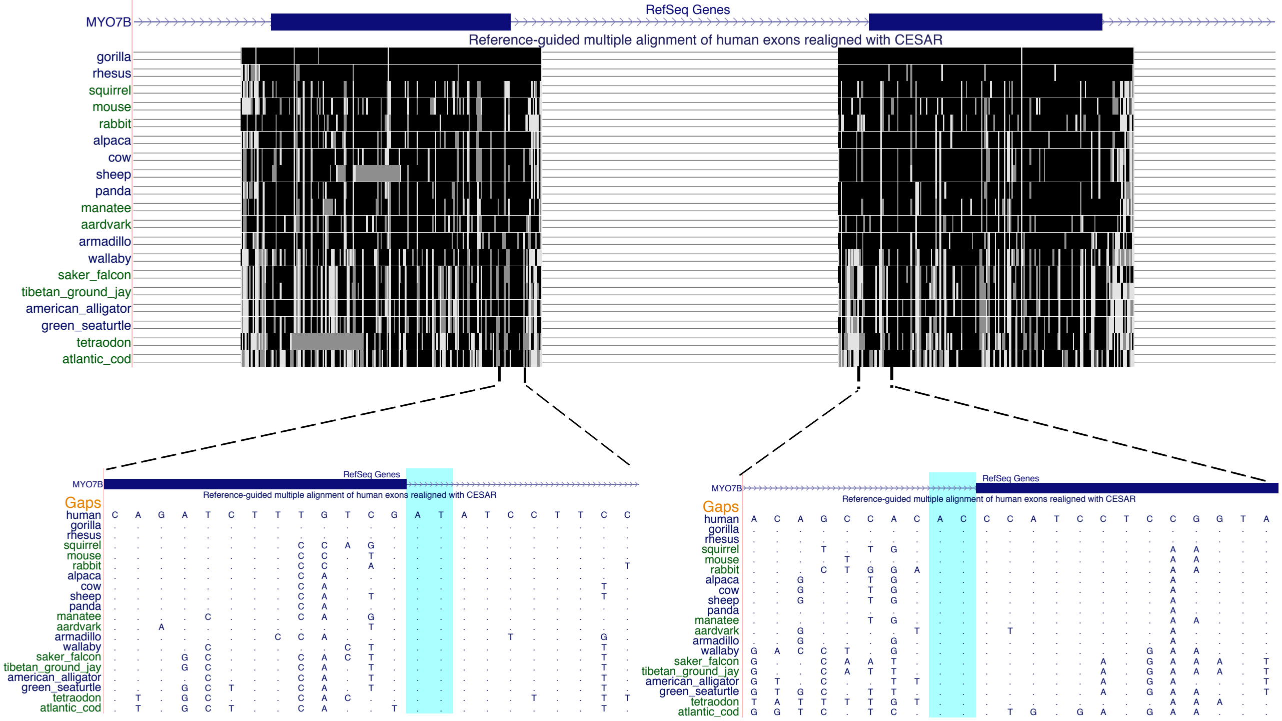
Supplementary Figure S21: Real frameshift in the last coding exon of *SLC16A4*.

(A) The human genome browser shows the last coding exon of *SLC16A4*. The alignment shows a frameshift in mouse (a 4 bp insertion – highlighted in gray).

(B) The mouse genome browser shows that the C-terminus of Slc16a4 is translated in a different frame. The dashed line shows the position of the frameshift.


A
Scale
20 bases
hg19
chr12:
6,500,020
6,500,025
6,500,030
6,500,035
6,500,040
6,500,045
6,500,050
6,500,055
6,500,060
6,500,065
6,500,070
6,500,075
---> GGCAAGGCTTGGCACCTAGCGGAGACAGAGCACTGTGGTGCCACACCCTCTAACAGGGGCCCAAGG
RefSeq Genes
LTBR G K A W H L A E T E H C G A T P S N R G P R
LTBR G K A W H L A E T E H C G A T P S N R G P R
Basic Gene Annotation Set from GENCODE Version 19
LTBR G K A W H L A E T E H C G A T P S N R G P R
LTBR G K A W H L A E T E H C G A T P S N R G P R
LTBR G K A W H L A E T E H C G A T P S N R G P R
Multiz Alignments of 100 Vertebrates realigned with CESAR (UCSC canonical gene set)
Gaps
1
human GGCAAGGCTTGGCACCTAGCGGAGACAGAGCACTGTGGTGCCACACCCTCTAACAGGGGCCCAAGG
mouse GGCAAAGCTTGGCATCTGGCCGAGACAGAGCACTAGGATGCCAAGACCTCTGACA--GGGCCAAGG
frameshifting 1 bp insertion
B
mm10
20 bases
125,307,290
125,307,280
125,307,270
125,307,260
125,307,250
125,307,240
125,307,230
Scale
:chr6
RefSeq Genes
P Y Q E D G K A W H L A E T E T L G C Q D L *
Ltbr
Basic Gene Annotation Set from ENCODE/GENCODE Version M7 (Ensembl 82)
P Y Q E D G K A W H L A E T E T L G C Q D L *
Ltbr
translated in a different reading frame

Supplementary Figure S22: Real frameshift in the last coding exon of *LTBR*.
(A) The human genome browser shows the last coding exon of *LTBR*. The alignment shows a frameshift in mouse (1 bp insertion – highlighted in gray).
(B) The mouse genome browser shows that the C-terminus of Ltbr is translated in a different frame. The dashed line shows the position of the frameshift.



Supplementary Figure S23: CESAR alignments of exons flanked by a U12 intron. The *MYO7B* gene (hg19: chr2:128,388,703-128,389,413) contains a U12 intron with an AT donor and an AC acceptor splice site. Using U12 specific splice site profiles, CESAR correctly identifies the conserved AT-AC splice sites.

Supplementary Table S1: The probability of deleting codon(s). These probabilities were derived from a human-rhesus alignment.

Number of codons deleted	**Probability**
1	0.00916
2	0.00396
3	0.00253
4	0.00191
5	0.00150
6	0.00135
7	0.00122
8	0.00116
9	0.00113
10	0.00108
Sum	**0.025**

Supplementary Table S2: The insertion probabilities associated with different sense codons.

Codon	**Codon insertion probability**
ACC	0.017981249
ATG	0.018984288
AAG	0.022613024
AAA	0.020731263
ATC	0.019797907
AAC	0.018084301
ATA	0.010602242
AGG	0.015901891
CCT	0.019572174
CTC	0.01682396
AGC	0.019045792
ACA	0.019485644
AGA	0.018170669
CAT	0.013882891
AAT	0.016456899
ATT	0.017789867
CTG	0.034270648
CTA	0.0084321
ACT	0.016427456
CAC	0.016768181
ACG	0.007151639
CAA	0.012221466
AGT	0.014247335
CCA	0.020065188
CCG	0.007220994
CCC	0.017115286
TAT	0.014817392
GGT	0.012572496
TGT	0.015166296
CGA	0.010368331
CAG	0.026072787
CGC	0.014941054
GAT	0.017237967
CGG	0.013522864
CTT	0.0149404
TGC	0.016183075
GGG	0.014124981
GGA	0.019038267
GGC	0.018230046
TAC	0.017342818
GAG	0.02179875
TCG	0.005771725
TTA	0.009876462
TTT	0.017288348
GAC	0.016802695
CGT	0.009603293
GAA	0.018091498
TCA	0.016484707
GCA	0.018616245
GTA	0.009152154
GCC	0.021175368
GTC	0.014428412
GCG	0.006742048
GTG	0.026216405
TTC	0.017482347
GTT	0.014378031
GCT	0.020157935
TTG	0.014331903
TCC	0.018710301
TGG	0.016215136
TCT	0.02007582
Sum	**1.000000**

Supplementary Table S3: The different parameters that were used to run the spliced aligners that were examined in this study.

Program	**Parameters**
Exonerate Exonerate (protein sequence as input)	–model coding2genome –n 1 –model protein2genome –n 1
Spaln Spaln (protein sequence as input)	-01 -pw -yX -S1 -01 -pw -yX
Pairagon	-vulgar -cross -a GMap
GeneWise	-genes -pretty -quiet

Mutations

	Genome Alignment			realigned with CESAR		
Species	Exon-inactivating mutations	Frameshifts	Splice site mutations	Exon-inactivating mutations	Frameshifts	Splice site mutations
Mouse	11937	9046	2891	794	614	180
Rat	12248	9197	3051	1084	793	291
Cow	14377	11859	2518	1418	1185	233
Dog	14574	11834	2740	1307	1042	265

Mutated Exons

	Genome alignment								realigned with CESAR							
Species	Exons with inactivating mutations		Exons with only frameshifts		Exons with only splice site mutations		Exons with both frameshifting and splice site mutations		Exons with inactivating mutations		Exons with only frameshifts		Exons with only splice site mutations		Exons with both frameshifting and splice site mutations	
Mouse	5772	3,87%	2980	2,00%	1796	1,20%	996	0,67%	580	0,39%	408	0,27%	135	0,09%	37	0,02%
Rat	6099	4,15%	3154	2,15%	1870	1,27%	1075	0,73%	807	0,55%	557	0,38%	207	0,14%	43	0,03%
Cow	6226	4,24%	3763	2,56%	1602	1,09%	861	0,59%	1002	0,68%	816	0,56%	160	0,11%	26	0,02%
Dog	6296	4,27%	3634	2,46%	1732	1,17%	930	0,63%	959	0,65%	738	0,50%	183	0,12%	38	0,03%

Mutated Genes

	Genome alignment								realigned with CESAR							
Species	Genes with inactivating mutations		Genes with only frameshifts		Genes with only splice site mutations		Genes with both frameshifting and splice site mutations		Genes with inactivating mutations		Genes with only frameshifts		Genes with only splice site mutations		Genes with both frameshifting and splice site mutations	
Mouse	4425	32,8%	2149	15,9%	1084	8,0%	1192	8,8%	470	3,5%	348	2,6%	82	0,6%	40	0,3%
Rat	4597	34,1%	2216	16,4%	1074	8,0%	1307	9,7%	665	4,9%	485	3,6%	120	0,9%	60	0,4%
Cow	4666	34,6%	2603	19,3%	913	6,8%	1150	8,5%	858	6,4%	716	5,3%	97	0,7%	45	0,3%
Dog	4766	35,3%	2548	18,9%	1010	7,5%	1208	8,9%	842	6,2%	660	4,9%	122	0,9%	60	0,4%

Supplementary Table S4: A break down of the number of mutations observed in genome alignments and after realigning with CESAR.

				intact exons		genes with at least one intact exon		genes where all exons are intact	
	Assembly	Species	distance to human (substitutions per neutral site)	number	percent of 188788 exons	number	percent of 19865 genes	number	percent of 19865 genes
MAMMALS	panTro4	Chimp	0,013	179898	95,29	19102	96,16	15158	76,31
	gorGor3	Gorilla	0,017	174526	92,45	18581	93,54	13748	69,21
	ponAbe2	Orangutan	0,036	171362	90,77	18247	91,86	11848	59,64
	nomLeu3	Gibbon	0,042	174304	92,33	18440	92,83	12019	60,50
	rheMac3	Rhesus	0,070	177902	94,23	18506	93,16	13136	66,13
	macFas5	Crab-eating macaque	0,069	182427	96,63	18767	94,47	16231	81,71
	papHam1	Baboon	0,070	174640	92,51	17913	90,17	13145	66,17
	chlSab1	Green monkey	0,070	183027	96,95	18733	94,30	16471	82,92
	calJac3	Marmoset	0,122	173871	92,10	18058	90,90	12119	61,01
	saiBol1	Squirrel monkey	0,121	176982	93,75	18203	91,63	13169	66,29
	otoGar3	Bushbaby	0,272	177235	93,88	18041	90,82	13665	68,79
	tupChi1	Chinese tree shrew	0,324	172467	91,36	17517	88,18	10888	54,81
	speTri2	Squirrel	0,331	174631	92,50	17757	89,39	13000	65,44
	jacJac1	Lesser Egyptian jerboa	0,432	170146	90,13	17089	86,03	11720	59,00
	micOch1	Prairie vole	0,501	172989	91,63	17151	86,34	13454	67,73
	criGri1	Chinese hamster	0,482	168945	89,49	16897	85,06	9858	49,63
	mesAur1	Golden hamster	0,495	163008	86,34	16860	84,87	10861	54,67
	mm10	Mouse	0,514	176103	93,28	17499	88,09	14428	72,63
	rn5	Rat	0,520	172744	91,50	17187	86,52	12986	65,37
	hetGla2	Naked mole-rat	0,377	174191	92,27	17475	87,97	13694	68,94
	cavPor3	Guinea pig	0,433	171040	90,60	17214	86,66	11352	57,15
	chiLan1	Chinchilla	0,412	174014	92,17	17507	88,13	12628	63,57
	octDeg1	Brush-tailed rat	0,451	172527	91,39	17289	87,03	12604	63,45
	oryCun2	Rabbit	0,369	160895	85,23	16502	83,07	10749	54,11
	ochPri3	Pika	0,454	169085	89,56	17104	86,10	12167	61,25
	susScr3	Pig	0,367	159686	84,59	16818	84,66	10772	54,23
	vicPac2	Alpaca	0,362	173513	91,91	17680	89,00	11466	57,72
	camFer1	Bactrian camel	0,361	170853	90,50	17605	88,62	10245	51,57
	turTru2	Dolphin	0,333	162710	86,19	16582	83,47	8805	44,32
	orcOrc1	Killer whale	0,332	175913	93,18	17572	88,46	14041	70,68
	panHod1	Tibetan antelope	0,399	171872	91,04	17535	88,27	10765	54,19
	bosTau7	Cow	0,434	175414	92,92	17806	89,64	13121	66,05
	oviAri3	Sheep	0,403	168934	89,48	17472	87,95	9996	50,32
	capHir1	Domestic goat	0,402	167720	88,84	17482	88,00	9468	47,66
	equCab2	Horse	0,322	171452	90,82	17646	88,83	10322	51,96
	cerSim1	White rhinoceros	0,303	178141	94,36	17943	90,33	14231	71,64
	felCat5	Cat	0,348	173555	91,93	17793	89,57	11159	56,17
	canFam3	Dog	0,369	176094	93,28	17910	90,16	12866	64,77
	musFur1	Ferret	0,392	175775	93,11	17814	89,68	13046	65,67
	ailMel1	Panda	0,365	172388	91,31	17308	87,13	11723	59,01
	odoRosDiv1	Pacific walrus	0,358	178257	94,42	17807	89,64	14682	73,91
	lepWed1	Weddell seal	0,355	170627	90,38	17133	86,25	12659	63,73
	pteAle1	Black flying-fox	0,350	174642	92,51	17581	88,50	12431	62,58
	pteVam1	Megabat	0,351	153684	81,41	16789	84,52	6234	31,38
	myoDav1	David's myotis (bat)	0,392	164806	87,30	16872	84,93	8790	44,25
	myoLuc2	Microbat	0,381	158460	83,94	16358	82,35	9641	48,53
	eptFus1	Big brown bat	0,381	172205	91,22	17308	87,13	12523	63,04
	eriEur2	Hedgehog	0,467	161747	85,68	16015	80,62	11810	59,45
	sorAra2	Shrew	0,528	157817	83,60	15966	80,37	10026	50,47
	conCri1	Star-nosed mole	0,402	161113	85,34	16428	82,70	11876	59,78
	loxAfr3	Elephant	0,362	171382	90,78	17562	88,41	11064	55,70
	eleEdw1	Cape elephant shrew	0,497	167221	88,58	16453	82,82	12564	63,25
	triMan1	Manatee	0,345	175256	92,83	17689	89,05	13277	66,84
	chrAsi1	Cape golden mole	0,421	173975	92,15	17327	87,22	13672	68,83
	echTel2	Tenrec	0,507	165703	87,77	16912	85,14	10228	51,49
	oryAfe1	Aardvark	0,366	173459	91,88	17453	87,86	12946	65,17
	dasNov3	Armadillo	0,353	166767	88,34	17187	86,52	10644	53,58
	monDom5	Opossum	0,778	167167	88,55	17571	88,45	10347	52,09
	sarHar1	Tasmanian devil	0,797	160701	85,12	17227	86,72	7751	39,02
	macEug2	Wallaby	0,778	120532	63,85	15526	78,16	2669	13,44
	ornAna1	Platypus	0,968	129515	68,60	15033	75,68	4346	21,88

Supplementary Table S5 (part 1): Percent of human exons and genes that we annotate in 99 non-human vertebrates. Continued on the next page.

				intact exons		genes with at least one intact exon		genes where all exons are intact	
	Assembly	Species	distance to human (substitutions per neutral site)	number	percent of 188788 exons	number	percent of 19865 genes	number	percent of 19865 genes
	falChe1	Saker falcon	1,197	118446	62,74	13695	68,94	2971	14,96
S	falPer1	Peregrine falcon	1,197	119653	63,38	13862	69,78	3128	15,75
A	ficAlb2	Collared flycatcher	1,336	116148	61,52	13438	67,65	3310	16,66
U	zonAlb1	White-throated sparrow	1,375	111363	58,99	12924	65,06	3286	16,54
R	geoFor1	Medium ground finch	1,350	113874	60,32	13291	66,91	2650	13,34
O	taeGut2	Zebra finch	1,349	138125	73,16	15205	76,54	4644	23,38
P	pseHum1	Tibetan ground jay	1,313	124539	65,97	14241	71,69	4791	24,12
S	melUnd1	Budgerigar	1,235	115388	61,12	13443	67,67	3482	17,53
I	amaVit1	Parrot	1,260	109715	58,12	13319	67,05	2786	14,03
D	araMac1	Scarlet macaw	1,276	100007	52,97	13047	65,68	2016	10,15
A	colLiv1	Rock pigeon	1,227	116580	61,75	13687	68,90	2855	14,37
	anaPla1	Mallard duck	1,207	111240	58,92	13165	66,27	2446	12,31
	galGal4	Chicken	1,242	120518	63,84	13841	69,68	4135	20,82
	melGal1	Turkey	1,264	135159	71,59	15035	75,69	3514	17,69
	allMis1	American alligator	1,057	156967	83,15	16777	84,46	7405	37,28
	cheMyd1	Green seaturtle	0,998	149938	79,42	16415	82,63	5199	26,17
	chrPic1	Painted turtle	1,004	153944	81,54	16810	84,62	6897	34,72
	pelSin1	Chinese softshell turtle	1,053	142844	75,66	16358	82,35	4798	24,15
	apaSpi1	Spiny softshell turtle	1,104	128666	68,15	15497	78,01	3071	15,46
	anoCar2	Lizard	1,272	139306	73,79	15571	78,38	5056	25,45
	xenTro7	X. tropicalis	1,723	131987	69,91	15107	76,05	4413	22,22
	latCha1	Coelacanth	1,417	131129	69,46	15202	76,53	4150	20,89
	tetNig2	Tetraodon	2,376	101546	53,79	13225	66,57	1919	9,66
T	fr3	Fugu	2,355	108105	57,26	13526	68,09	2574	12,96
E	takFla1	Yellowbelly pufferfish	2,403	97014	51,39	12780	64,33	1849	9,31
L	oreNil2	Nile tilapia	2,200	113542	60,14	13669	68,81	2755	13,87
E	neoBri1	Princess of Burundi	2,237	109593	58,05	13743	69,18	2458	12,37
O	hapBur1	Burton's mouthbreeder	2,223	111824	59,23	13824	69,59	2601	13,09
S	mayZeb1	Zebra mbuna	2,224	111995	59,32	13808	69,51	2600	13,09
T	punNye1	Pundamilia nyererei	2,228	111582	59,10	13792	69,43	2550	12,84
	oryLat2	Medaka	2,349	108653	57,55	13531	68,12	2452	12,34
F	xipMac1	Southern platyfish	2,316	110660	58,62	13724	69,09	2535	12,76
I	gasAcu1	Stickleback	2,104	112467	59,57	13684	68,89	2662	13,40
S	gadMor1	Atlantic cod	2,133	97172	51,47	13354	67,22	2083	10,49
H	danRer7	Zebrafish	2,211	114799	60,81	14032	70,64	3042	15,31
	astMex1	Mexican tetra (cavefish)	2,105	111496	59,06	14147	71,22	2822	14,21
	lepOcu1	Spotted gar	1,793	125258	66,35	14834	74,67	3622	18,23
	petMar2	Lamprey	2,192	58439	30,96	9540	48,02	1069	5,38

Supplementary Table S5 (part 2): Percent of human exons and genes that we annotate in 99 non-human vertebrates. Non-mammalian species.