Comparative analysis of brain and fat body gene splicing patterns in the honey bee, *Apis mellifera*

(Supplementary Materials)

Supplementary methods

Modifying Amel 4.5 Gene models

Transcribed Islands. To obtain reliable 'transcribed islands' for the honey bee genome, we filtered TrueSight alignments for 10 samples by only retaining the best alignment (i.e., smallest number of mismatches for full alignment and highest TrueSight score for gapped alignment) for each RNA-seq read. After this filtering, we obtained read counts for each nucleotide position in the honey bee genome and searched for *transcribed islands* with the following criteria: (i) at least 5x coverage for each base of the island; and (ii) a transcribed island should be longer than 50bp. Boundaries for 'transcribed islands' identified at this stage were determined independently from splicing signals or SJs inferred by TrueSight. The boundaries were further refined by the subsequent Gene models modification process. All transcribed islands were compared with exons of Amel 4.5 Gene models; only islands non-overlapping with existing exons were retained for identifying novel exons.

TrueSight splice junction. SJs from independent TrueSight runs on ten RNA-seq samples were clustered together, and the highest TrueSight score was assigned for each junction if it was detected in multiple samples. SJs with score higher than 0.5 were retained as TrueSight SJs and were utilized to improve the current Gene models.

Improving Gene models with iterative algorithm. We used the following procedure to update the existing Gene models.

- **Initiation:** By comparing TrueSight SJs with SJs from the amel 4.5 Gene models ($model^0$, we define a set of exons and SJs as a primary gene$model$), TrueSight SJs are categorized into four subsets: (i) known SJs, which are already in $model^0$; (ii) novel SJs with both splice sites known ($novel_0^0$), which reflect exon skipping; (iii) novel SJs with only one known splice site ($novel_1^0$); and (iv) novel SJs with two novel splice sites ($novel_2^0$).

- **Iteration:** In t^{th} (t ≥ 1) iteration

  Adding new link (SJ) to existing exons with $novel_0^{t-1}$

  SJs in $novel_0^{t-1}$ provide novel links to existing exons in $model^{t-1}$, and are strong supports for cassette exons. $novel_0^{t-1}$ is added into $model^{t-1}$ to construct a new version of Gene models $model^t$.

  Modifying exon coordinates with $novel_1^{t-1}$

  The original junction linking two exons (*a* ~ *b*; *c* ~ *d*) is *b* ~ *c*. If there is junction in $novel_1^{t-1}$: $b' \sim c$, such that $\|b' - b\| < 200, b' > a$, exon $a \sim b$ would have alternative boundary $a \sim b'$. If there is junction in $novel_1^{t-1}$: $b \sim c'$, such that $\|c' - c\| < 200, c' < d$, exon $c \sim d$ would have alternative boundary $c' \sim d$.

  SJs used in modifying exon coordinates should be in the same strand as the exon. Both the SJs and modified exon boundaries are added into $model^t$. All junctions in $novel_1^{t-1}$, which are not utilized in modifying exon boundaries and are not added into $model^t$, are treated as $novel^{t-1}$. Junctions in $novel_2^{t-1}$are also added into $novel^{t-1}$. Junctions in the new set $novel^{t-1}$ are compared with $model^t$and categorized into$novel_0^t$, $novel_1^t$ and $novel_2^t$, which are used in the $t + 1^{th}$ iteration.

- **Termination:** The modification process would terminate at T^{th}iteration when there is no junction in $novel_0^T$. SJs in $novel_1^{\mathrm{T}}$ and $novel_2^{\mathrm{T}}$ are utilized to add new exons and SJs to$model^T$. The newly added junctions are signals for two types of AS: cassette exons (CE) and alternative exon boundaries (AEB).

Adding new exons and splice junctions. To be conservative in adding novel exons and SJs into the modified amel 4.5 Gene models, we only use $novel_1^{\mathrm{T}}$ and $novel_2^{\mathrm{T}}$ with TrueSight score greater than 0.9.

- **Adding new exons and splice junctions with $\boldsymbol{novel_1^T}$.** For exon $a \sim b$, if there is a junction in $novel_1^{t-1}$: $b \sim p'$ such that we can find a transcribed island ($p \sim q$) satisfying $\|p' - p\| < 100, p' < q$, we can label the transcribed island ($p' \sim q$) as a new exon, with one boundary (p') fixed and the

other (q) undetermined. Symmetrically, for exon $a \sim b$, if there is a junction in $novel_1^{t-1}$: $q' \sim a$ such that we can find a transcribed island ($p \sim q$) satisfying $\|q' - q\| < 100, q' > p$, we can label the transcribed island ($p \sim q'$) as a new exon, with one boundary (q') fixed and the other (p) undetermined. The new exons identified in this process are in two subsets: (i) new exons with both boundaries fixed, since both ends of these exons are linked to known exons by junctions in $novel_1^T$. (ii) new exons with only one end defined (*Novel Terminal Exons*). These *Novel Terminal Exons* are either first/last exons of the whole transcripts, or linked to further novel exons by SJs in $novel_2^T$.

- **Adding new exons and splice junctions with $\boldsymbol{novel_2^T}$.** For a SJ in $novel_2^T$: $q' \sim p'$, if there are two transcribed islands ($p_1 \sim q_1, p_2 \sim q_2$) such that: $q_1 < p_2, \|q' - q_1\| < 100, q' > p_1, \|p' - p_2\| < 100, p' < q_2$, we can link these two transcribed islands together by the junction $q' \sim p'$. There are two outcomes of this novel exon adding process: (i) novel multi-exon transcripts can be identified in inter-genic regions of the Gene models; (ii) novel exons might be connected to Novel Terminal Exons identified in the previous process, thus these novel exons serve as new *Novel Terminal Exons* for known genes.

Supplementary Tables and Figures

Method	All junctions	Both introns annotated		One splice site novel	Both Novel	SN (%)	SP (%)
		Known introns	Novel introns				
TrueSight Brain	157706	57210	57210	7933	47866	70	100
MapSplice Brain	129796	55436	55436	8957	32603	74.881	96.899
TopHat2 Brain	131353	55751	55751	7190	30489	76.789	97.45

Table S1. Performance of TrueSight, MapSplice and TopHat2 for Honey bee RNA-seq dataset from Brain.

**Other tables provided as excel spreadsheet

Supplementary figures

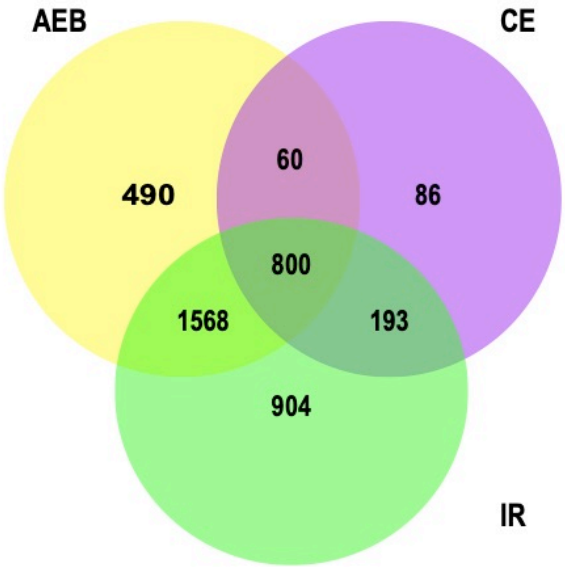

Fig. S1. Venn diagram showing overlap in genes having more than 1 AS events in fat body.

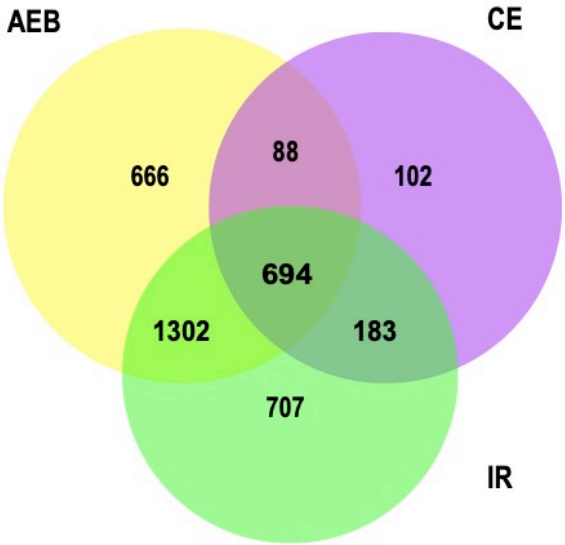

Fig. S2. Venn diagram showing overlap in genes having more than 1 AS events in Brain.

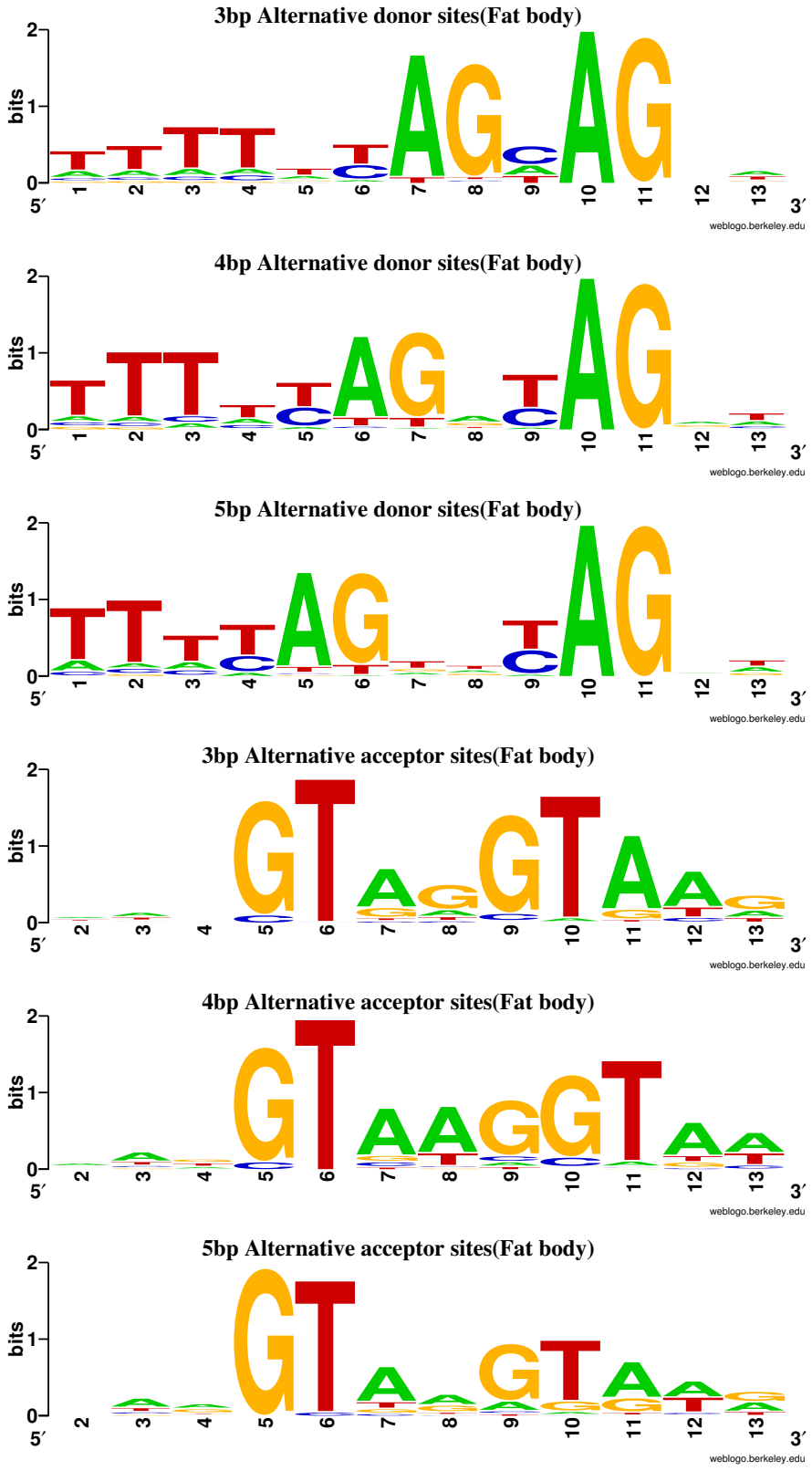

Fig. S3. Motif logos of splice sites in 3/4/5 bp displaced AEBs in both tissues in Fat body.

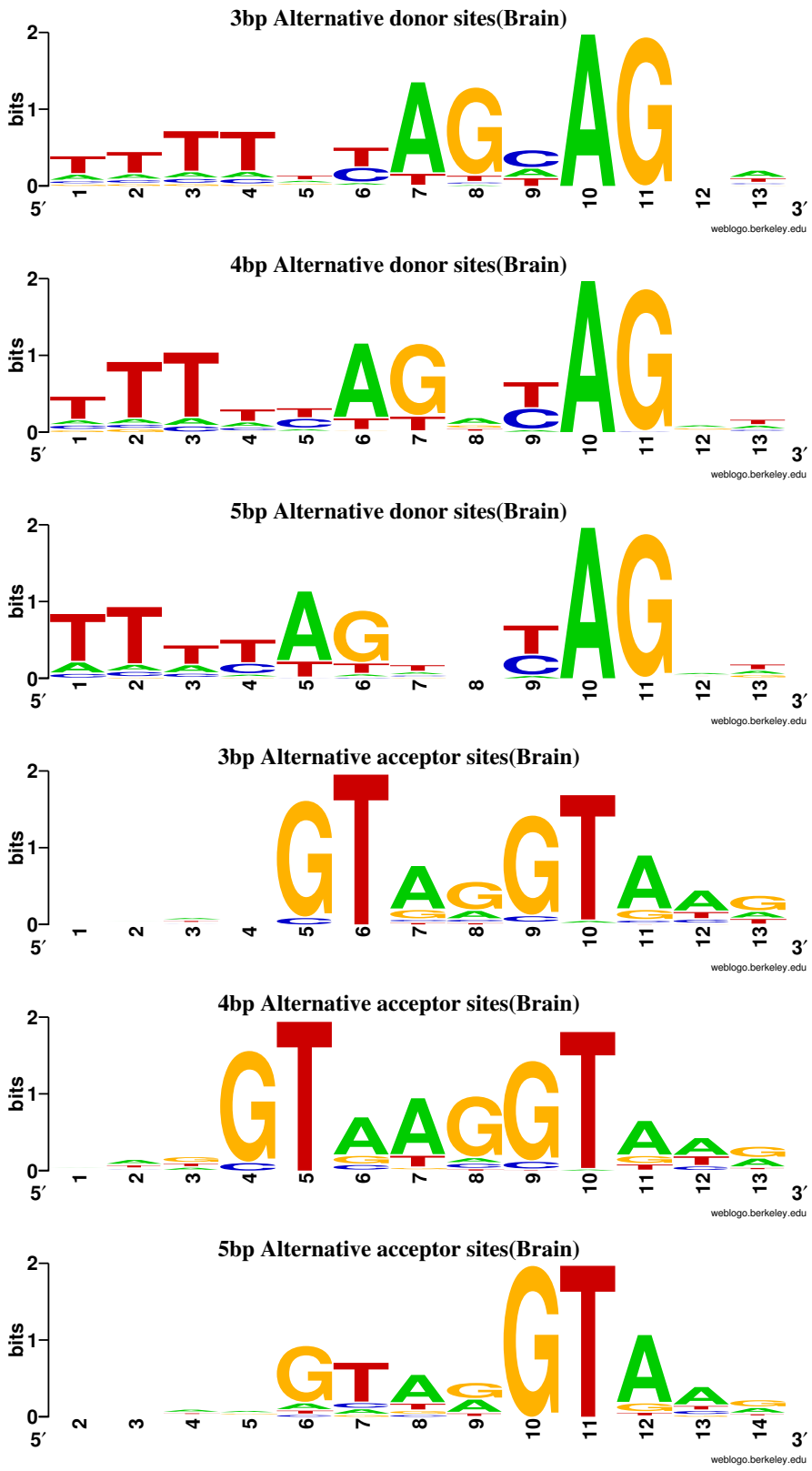

Fig. S4. Motif logos of splice sites in 3/4/5 bp displaced AEBs in both tissues in Brain.

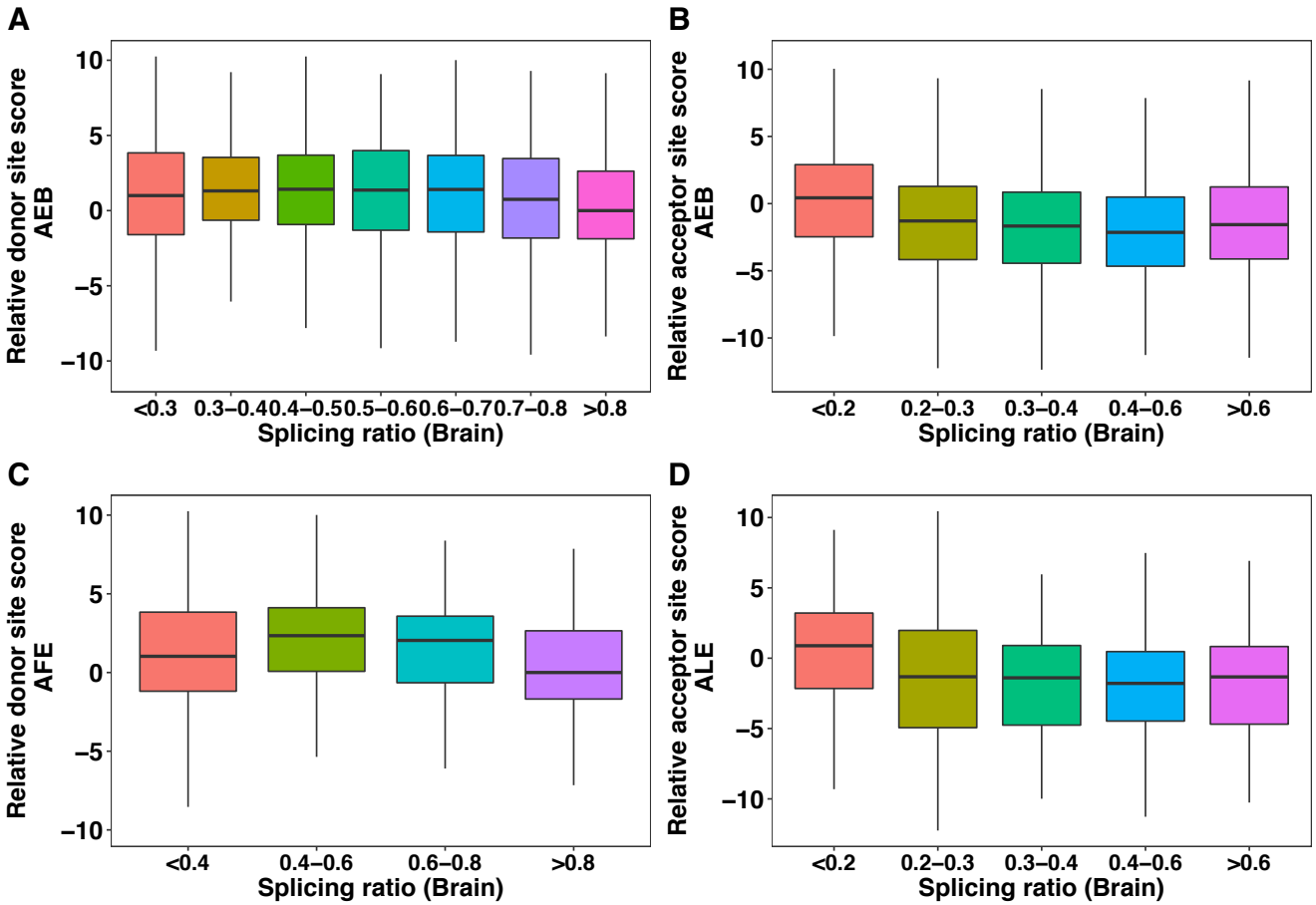

Fig. S5. (A-B) Impact of relative splice site score (major - minor) on AEB splicing ratio (minor/major) in brain. (C-D) Impact of relative splice site score (major-minor) on ATE splicing ratio (minor/major) in brain. P- values for Mann-Whitney-Wilcoxon tests for 1st and last splicing ratio categories: A = 0.0004402, B = 1.342e-15, C = 0.04852, D = 9.985e-05.

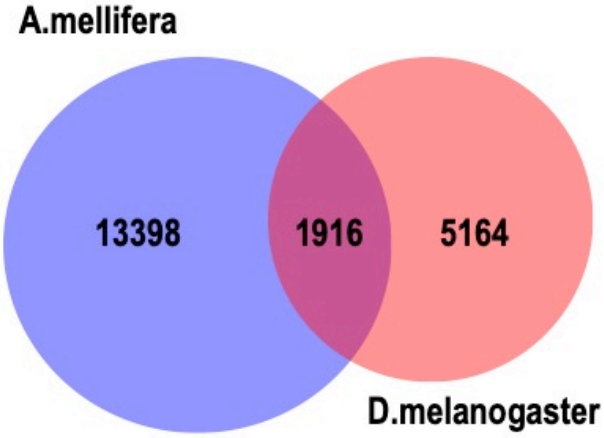

Fig. S6. Venn diagram showing cross-species conservation of AS genes between honeybee and Drosophila.

Additional Supplementary Files

Table S2.gff3. Improved honey bee gene model after prediction of alternative splicing patterns using TrueSight.

Table S3-S8.xlsx, which includes the following tables.

Table S3: All alternative splicing events in Fat body and Brain from TrueSight.

Table S4: *Drosophila* orthologs of honey bee alternatively spliced genes.

Table S5: *Drosophila* orthologs of honey bee alternatively spliced genes that undergo alternative splicing in *Drosophila* as well.

Table S6: Methylation and un-methylated fat body specific alternative splicing events along with drosophila orthologs.

Table S7: Functional categories of common genes undergoing AS in fat body and brain.

Table S8: Functional categories of genes involved in tissue specific alternative splicing events and AS events corresponding to brain regulatory network genes.