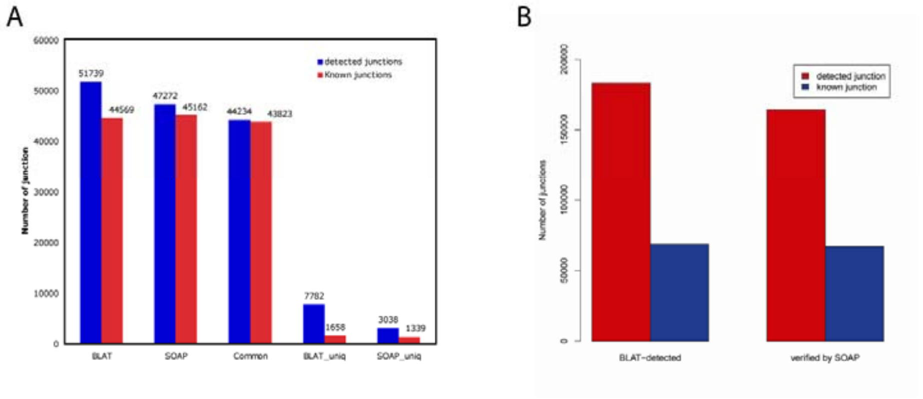
Supplemental Figures

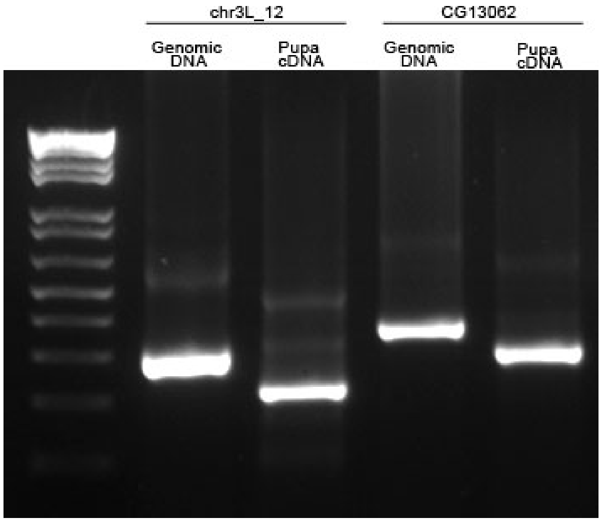
Supplemental Tables

SUPPLEMENTAL FIGURES

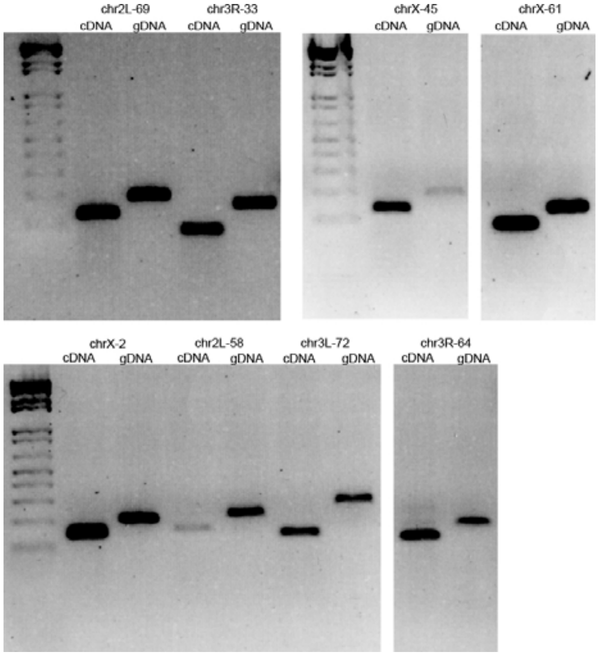


Supplemental Figure 1, related to Figure 1 - Junction reads identify annotated and novel introns
To maximize sensitivity for detecting junction reads two independent computational approaches were used: BLAT based alignment to the reference genome and SOAP alignment to an assembled ‘junctionome (see Experimental Procedures). (A) In total 46820 annotated junctions are observed with 93.6% consistency between the approaches. (B) BLAT identified 120,000 candidate novel junctions; these were added to the junctionome and support by the SOAP approach was indicated for more than 87%.

A



B

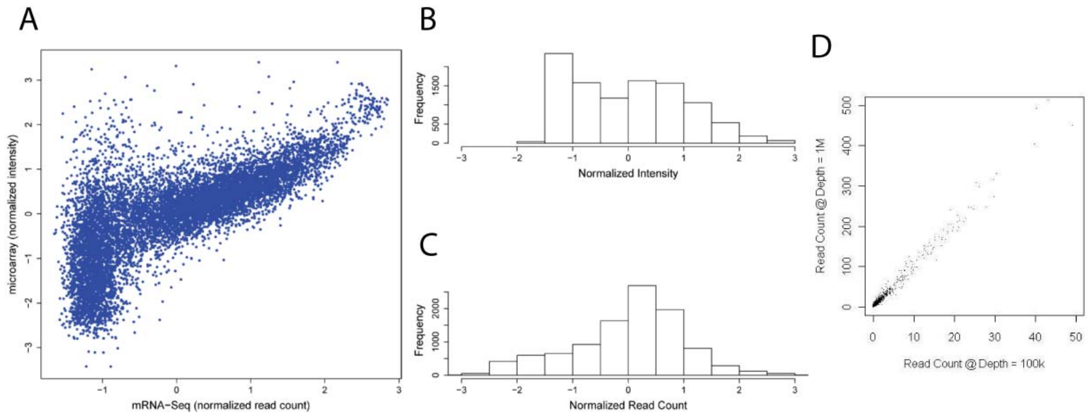


Supplemental Figure 2 - RT-PCR validation of novel junctions predicted by RNA-Sequencing

(A) chr3L_12 is a predicted novel transcript, PCR primers were designed to amplify a 250bp product from genomic DNA. The expected splice junction associated with this transcript, 3L:16099002-16099076, results in a 177bp RT-PCR product when amplified from Pupa cDNA. Above, CG13062 suggests a splicing junction which drastically alters the annotated gene model. PCR primers were designed to amplify 324 bp product from genomic DNA. The expected splice junction associated with this transcript, 3L:16307638-16307703, results in a 260bp RT-PCR product when amplified from Pupa cDNA. (B) RT-PCR was used to verify splicing in eight additional novel transcripts: chr3R_33, chr2L_69, chrX_61, chrX_45, chrX_2, chr2L_58, chr3L_72 and chr3R_64.

>chr2L_12 Pupa cDNA RT-PCR product
TGTACCAGGTGGACTTCGTGAAAAGGCAAGCACATGGCAGTGTAAAATTT
GCCGCCAAAAGCAAAGTTTGCTCAAGGAATTCTTTCGAGGATCGGCAGCG
GAATGCCGGGTTAAGGTGCAGCATCTCAACTTAGAGCGCGG

>CG13062 Pupa cDNA RT-PCR product
AACCACCCACAAGATGATGAAACTGGTAGTGTCGCTACTCTCAATTTGCG
CCTTGACGGCAGCTCGTCCTGGTTTCCTGCATGGCCATCACTATCCGGAG
ATCCCTTACTATCCACACCACCATCATGTGGAACCACTGCACTACCATCT
GCCCGCCGCTGTCTCCCACCAGAGCTCCACGGTGGTGCACAGTGTGCCGC
ACCACATAATCAAGCCGGTCCTGTCATAGCTGTTTCCTGAGCA



Supplemental Figure 3, related to Figure 3 - RNA-Seq accurately determines gene expression levels
(A) mRNA-Seq and microarray were performed on identical RNA libraries. Expression levels were calculated in platform specific ways and correlated. (B) The distribution of normalized intensity scores for the microarray platform for all genes. (C) The distribution of normalized expression levels for the mRNA-Seq platform for all genes. (D) Reads from E2-4hr library were randomly sampled at two depths, 1M and 100K reads. The correlation between these depths was calculated R=0.99.

SUPPLEMENTAL TABLES

Supplemental Table 1, related to Figure 1 – Observed junctions by stage

All junctions used in this study for gene model modification and alternate splicing detection are listed. This includes all observed annotated junctions and well-supported novel junctions. The amount of support (read counts) for each junction is indicated across all sequencing libraries.

Supplemental Table 2, related to Figure 2 – Gene Model Modifications

3' UTR	5' UTR	Novel Exon	Novel Coding	Genes
804	**775**	**3692**	**2028**	**4418**
5.5%	5.3%	25.3%	13.9%	30.3%

Analysis of coverage and junction data led to the modification of many genes by addition of exonic sequences, and extension of UTRs. The number of genes modified in each category is listed.

Supplemental Table 3, related to Figure 3 – Expression of Genes without a *Drosophila* Orthologue

This table has calculated expression levels for the stages in Figure 3G which shows genes without a Drosophila orthologue are highly expressed in males.

Supplemental Table 4, related to Figure 4 – Sex-biased splicing events

Each alternately splicing exon observed to be significantly differentially expressed between males and females is listed with the total gene counts of associated genes and the total exon counts for males and females.

Supplemental Table 5, related to Figure 4 – Observed NAGNAGs and GYNGYNs

Each observed candidate NAGNAG and GYNGYN splicing site is listed with its position and upstream/downstream splicing information.

SUPPLEMENTAL METHODS

Junction Detection

The first method for junction detection employed the SOAP algorithm to align reads to a constructed ‘junctionome’ database (Wang et al. 2009). We prepared two junction databases in this study: Exon Spliced Junction (ESJ) and Exon Random Junction (ERJ). ESJ is composed of all possible exon splicing junctions based on FlyBase gene model v5.23 with both intra-transcripts and inter-transcripts junctions considered, while ERJ is a negative control. ESJ database was constructed by pair-wise connection of exon sequences from every locus annotated by FlyBase. In principle, the last *n* bp of the upstream exon was connected to the first *n* bp of the corresponding downstream exon. We considered all possible combinations, e.g. exon *i* was connected to exon *i*+1, exon *i*+2, *etc*. ERJ database was built exactly the same way as we described for ESJ, except that the two exons joined together were selected randomly from separate loci. As a negative control, ERJ preserved inherent codon, dinucleotide and other compositional features. For our own convenience, the size of ERJ (i.e. total number of exon junctions) was set to be the same as ESJ. We generated 247,418 non-redundant exon junctions. All trimmed reads were mapped to ESJ and ERJ using SOAP (v1.11) with four mismatches allowed. We built a statistical model to assign each junction an empirical p-value, and then a critical p-value was selected based on a false discovery rate of 1%.

The second method for junction detection employed BLAT to map reads directly to the *Drosophila* reference genome. Paired reads were parsed to disambiguate multiple mapping locations if one of the two pairs mapped uniquely or agreed on a unique mapping location. Reads with unaligned mates or pairs which did not agree were treated as single-end reads. All reads which aligned with gaps in the reference sequence consistent with an annotated junction were considered ‘true’ junctions. Reads aligning with a gap suggesting novel introns were considered putative ‘novel’ junctions.

To identify a novel junction we required that supporting reads align in blocks of 10bp or greater on both sides of the alignment. Furthermore, after all potential junction reads were identified, alignments which occurred within 20bp of an annotated alignment were required to represent 5% or more of the local junction reads or the junction was discarded. Finally, the local coverage was considered and junctions without significant evidence in high or low coverage regions were discarded.

Approximating the contribution of sequencing error to false junction detection

Approximately 1/3 of candidate junctions suggest splicing at noncanonical sites, the majority of these cluster near annotated splicing sites. This observation suggests that it is possible a large portion of these splicing junctions are due to sequencing insertion and/or deletion error, because noncanonical splicing is extremely rare constituting less that 0.1% of annotated junctions in the fly genome. To approximate the contribution of sequencing error to false junction identification we considered all novel junctions with splice sites within 20 base pairs of an annotated site (Figure 4A). As a lower bound we calculate the frequency of noncanonical novel observed alignments near annotated junctions observed in the dataset: ~0.7% of reads. A more conservative estimate would be to assume that all junction reads 'near' an annotated alignment (noncanonical or otherwise) are false and thus the error rate is approximated as ~2.3%. From this we estimate the contribution of sequencing error to false junction detection as ~1-2% and suggest that for any annotated junction from 1-2% of its supporting reads might indicate a false 'novel' junction simply due to technical error. Our primary criteria for filtering out erroneous junctions, therefore, is to require a minimum threshold of 5% of reads associated by proximity with an annotated splice site when confirming a novel junction.

Alternate Splicing Detection

Annotated counts of alternate splicing were calculated according to our definitions for each event. Specifically, a 'skipped exon' is defined as the occurrence of any junction for a gene which completely contains an exon of the gene. A ‘retained intron’ is defined conversely as an exon which completely contains both a junctions donor and acceptor sites. Alternate donors and acceptors are identified by considering the orientation of each junction and the number of sites with which it shares a junction. Alternate first exons were defined as any exon without an upstream junction and alternate last exons conversely as any exon without a downstream junction. Mutually exclusive exons were defined as any consecutive internal exon pair (not first or last exons) which did not share a junction. Alternate splicing counts were calculated first from annotated junctions, second from observed annotated junctions, and finally incorporating novel junctions.