Genome-wide identification of zero nucleotide recursive splicing in Drosophila

Michael O. Duff1, Sara Olson1*, Xintao Wei1*, Sandra C. Garrett1, Ahmad Osman1, Mohan Bolisetty1, Alex Plocik1, Susan E. Celniker2 & Brenton R. Graveley1

Recursive splicing is a process in which large introns are removed in multiple steps by re-splicing at ratchet points—5’ splice sites recreated after splicing. Recursive splicing was first identified in the Drosophila Ultrabithorax (Ubx) gene and only three additional Drosophila genes have since been experimentally shown to undergo recursive splicing. Here we identify 197 zero nucleotide exon ratchet points in 130 introns of 115 Drosophila genes from total RNA sequencing data generated from developmental time points, dissected tissues and cultured cells. The sequential nature of recursive splicing was confirmed by identification of intron ratchet points generated by splicing to and from the ratchet points. We also show that recursive splicing is a constitutive process, that depletion of U2AF inhibits recursive splicing, and that the sequence and function of ratchet points are evolutionarily conserved in Drosophila. Finally, we identify four recursively spliced human genes, one of which is also recursively spliced in Drosophila. Together, these results indicate that recursive splicing is commonly used in Drosophila, occurs in humans, and provides insight into the mechanisms by which some large introns are removed.

Recursive splicing was first identified in the Drosophila melanogaster Ultrabithorax (Ubx) gene. The 73-kb intron within Ubx contains two alternative microexons (mI and mII) that both contain the consensus 5’ splice site sequence GTAAGA immediately downstream of the 3’ splice site. In addition, this intron contains a ratchet point, a zero nucleotide exon consisting of juxtaposed 3’ and 5’ splice sites. It has been shown that rather than being removed in a single step, the 73-kb Ubx intron is removed in four steps in which the upstream constitutive exon is spliced to exon mI, and subsequently re-spliced to exon mII, the ratchet point, and finally the downstream constitutive exon. A previous genome-wide computational search for potential ratchet points observed between D. melanogaster and D. pseudoobscura predicted 160 potential ratchet points in 124 introns of 106 genes. Of these, only seven ratchet points in three genes (kuzbanian (kuz), outspread (osp) and frizzled (fz)) have been reported to be experimentally validated.

We generated 10.9 billion uniquely mapped reads of rRNA-depleted, paired-end, strand-specific RNA sequence from 183 D. melanogaster individual RNA samples comprising 35 dissected tissue samples, 24 untreated and 11 ecdysone treated cell lines, 30 distinct developmental stages and males and females of four strains from the D. melanogaster Genetic Reference Panel (Supplementary Table 1). The majority of these RNA samples were previously used to generate poly(A)- RNA sequence data. As the current libraries were prepared without poly(A) selection, they contain a mixture of mRNA, pre-mRNA and nascent RNA. Co-transcriptional splicing can be observed in total, nuclear, or nascent RNA-seq data by the saw-tooth pattern of repeatedly decreasing read density across introns in the 5’ to 3’ direction of transcription (Fig. 1a). While visually inspecting these data on a genome browser, we noticed several large introns that lacked internal annotated exons yet possessed saw-tooth patterns of read density suggestive of co-transcriptional splicing, including the introns from Ubx (Fig. 1b), kuz, osp and fz that were previously shown to undergo recursive splicing. We hypothesized that such saw-tooth patterns could be indicative of recursive splicing, and therefore performed a genome-wide search for ratchet points supported by the RNA-seq data.

To identify potential zero nucleotide exon-type ratchet points, we parsed the RNA-seq alignments to identify novel splice junctions where the reads mapped to an annotated 5’ splice site and an unannotated 3’ splice site, and the genomic sequence at the 3’ splice site junction was AG/GT (Extended Data Fig. 1a). We also aligned the total RNA-seq data to a database of splice junctions between annotated exons and all potential ratchet points (AG/GT sequences) in the downstream intron that did not correspond to annotated 3’ splice sites. We then identified ratchet point junctions where reads mapped without any mismatches, with at least three distinct offsets, and with an overhang of at least eight nucleotides (Extended Data Fig. 1b). We then visually inspected each ratchet point independently identified by both methods on the genome browser, removing candidates that did not display an obvious saw-tooth pattern of read density or which clearly corresponded to an unannotated exon.

We identified a total of 197 ratchet points in 130 introns of 115 genes (Supplementary Table 2). Two of these ratchet points were missed by our computational approaches, but identified during the course of manual inspection on the browser, validated, and subsequently included in the remainder of these analyses. This provides the first experimental verification of 91 of the 160 ratchet points computationally predicted by ref. 2 based on comparative genomics (Supplementary Table 3). Of the 69 unverified ratchet points predicted by ref. 2, 34 correspond to previously unannotated exons, 23 lacked convincing saw-tooth patterns, 7 did not pass our recursive junction thresholds, and 5 could not be identified in the current assembly of the genome. Although it is difficult to conclude that these are not true ratchet points, we have not included them in our subsequent analysis as their supporting evidence is inconclusive. The other 106 (53.8%) of the ratchet points we identified are described here for the first time.

Most genes (100) contain only one recursively spliced intron, although 15 genes contain two. The number of ratchet points in an intron ranges from one to six (Extended Data Fig. 2a). The recursively spliced introns range in size from 11,341 bp to 132,736 bp, with an average size of 45,164 bp. Recursive-splice-site-containing introns are enriched in large introns (97% of all introns are smaller than the smallest recursive intron), although not all large introns contain recursive splice sites (Extended Data Fig. 2b). In fact, only 6% of introns larger than the smallest recursive intron are recursively spliced. The segments of the introns removed by recursive splicing range from 2,596 bp to 63,580 bp, with an average size of 17,953 bp (± 9,039 bp) and median size of 16,368 bp (Extended Data Fig. 2c). The luna gene contains a 108-kb intron with five ratchet points, such that the intron is...
removed in six stepwise recursive splicing events (Fig. 1c). The five ratchet points are supported by the saw-tooth pattern of read density across the intron, reads that map to the exon–ratchet point splice junctions (Fig. 1c), and have been validated by RT–PCR and Sanger sequencing (Fig. 1d). In total, RT–PCR and Sanger sequencing validated 24 ratchet points from 14 genes in Drosophila S2 cells (Extended Data Fig. 3).

Ratchet points are zero nucleotide exons, and therefore do not exist in the mRNA. However, direct evidence of recursive splicing can be obtained by identifying lariat introns—by-products of all splicing reactions that contain a 2’–5’ linkage between the first nucleotide of the intron and the branch point. Because reverse transcriptase can occasionally traverse the branch point, reads corresponding to the 5’ splice site–branch-point junction may be present in the total RNA-seq data (Fig. 2a). To identify putative recursive lariat introns, we generated a set of potential 5’ splice site–branch-point junctions for all recursively spliced introns, and all possible permutations, and aligned the total RNA-seq reads to them (Methods). Although rare, we identified 46 reads that mapped uniquely to 27 recursive lariat introns in 20 genes (Supplementary Table 4). Directed RT–PCR and sequencing experiments independently verified 14 recursive lariats in 9 genes (Extended Data Table 1) for a total of 41 recursive lariat introns in 26 genes. Ten of the lariat introns detected correspond to the first segment of the recursive introns and are also supported by standard splice junction reads. However, the remaining lariat introns detected correspond to internal segments, further supporting the sequential nature of recursive splicing. For example, couch potato (cpo) contains an intron that is removed in three recursive splicing events involving two ratchet points. We obtained evidence for all three lariats from both the total RNA-seq data and directed RT–PCR sequencing experiments (Fig. 2b). This analysis also identified the putative branch points used for these recursive splicing events. All but five of these branch points reside from −42 to −19 upstream of the 3’ splice site with a peak at −29 (Fig. 2c). Six of the 3’ splice sites appear to use two different branch points. We observed that 81% have an A at the branch point, while 12%, 5% and 2% have a T, C, or G, respectively (Fig. 2d).

The nucleotide sequences of ratchet points resemble juxtaposed 3’ and 5’ splice sites (Fig. 3a), and the regions immediately flanking the ratchet points are much more highly conserved than those flanking non-ratchet point AG/GT sequences in the same introns (Fig. 3b). However, the ratchet points have a more prominent pyrimidine tract, and a significantly (P ≤ 0.0001) higher frequency of a TT dinucleotide at positions −5 and −6 relative to the 3’ splice site when compared to introns genome-wide. Whereas only 43.76% (30,151 out of 68,898) of all introns have Ts at positions −5 and −6, 99.5% (196/197) of ratchet points do. The only ratchet point lacking a TT dinucleotide at positions −5 and −6 is in CG15360 which has a C at position −6 that is conserved in other Drosophila species. Notably, the majority of Caenorhabditis elegans 3’ splice sites have this sequence and it has been shown that the large U2AF subunit (encoded by U2af50 in Drosophila) interacts with these bases. Thus, the strong preference for the TT dinucleotide at positions −5 and −6 of Drosophila ratchet points could represent high affinity U2AF binding sites so that the ratchet points are efficiently recognized.

To test this hypothesis, we sequenced total RNA from untreated S2 cells as well as cells treated with double-stranded RNA (dsRNA) to knockdown expression of lariat debranching enzyme (ldbr) as a control, U2af38, U2af50, or both U2af38 and U2af50 (Extended Data Table 2). We observed ~20 recursive junction reads per million mapped reads that mapped to 119 and 100 distinct ratchet points in untreated controls and ldbr depleted cells, respectively (Fig. 3c). Depletion of U2af38 or U2af50 alone reduced the frequency of recursive junction reads 3–4-fold (corresponding to 81 and 64 distinct ratchet points) (Fig. 3c). Notably, depletion of both U2af38 and U2af50 resulted in a complete absence of detectable recursive junctions reads (Fig. 3c), although similar fractions of non-recursive junction reads were observed in all samples (Fig. 3d). Depletion of U2AF might have such a strong impact on recursive splicing—but not on non-recursive splicing—because recursive junction reads can only be generated from nascent RNA while non-recursive junction reads can be generated from stable mRNAs. Additionally, exon or intron definition may not be possible for zero nucleotide exons in Drosophila due to the combination of large introns and non-existent exons. This would eliminate many of the cooperative interactions normally involved in splice site recognition, making recursive splicing particularly sensitive to decreased U2AF levels. Although additional work will be required to fully elucidate the role for U2AF in recursive splicing, this result strongly suggests that U2AF is required for efficient recognition of ratchet points in S2 cells.

To determine whether recursive splicing is evolutionarily conserved, we generated rRNA-depleted, stranded RNA-seq data from Drosophila simulans, Drosophila sechellia, Drosophila yakuba, Drosophila pseudoobscura and Drosophila virilis adults (Extended Data Table 3). We aligned these data to the corresponding reference genomes and searched for splice junction reads whose 3’ splice sites
mapped to positions orthologous to the identified *D. melanogaster* ratchet points. Despite having two orders of magnitude fewer reads from these species, 131 of the 197 *D. melanogaster* ratchet points (66.5%) were identified in at least one of the five other *Drosophila* species, 69 of which were identified in at least two species (Extended Data Table 4 and Supplementary Table 5). Together, these observations demonstrate that the nucleotide sequence and function of ratchet points are conserved among *Drosophila* species, indicating that recursive splicing is evolutionarily conserved.

We also searched for zero nucleotide exons in humans by generating and analysing >1.1 billion reads of total RNA data from 20 tissues (Extended Data Table 5). Our analysis pipeline identified 76 putative ratchet points, but upon further inspection all but five ratchet points in four genes were eliminated because they either lacked an obvious saw-tooth pattern of read density or corresponded to unannotated exons. The five ratchet points that we identified in *PDE4D, HS6ST3, CADM2* and *ROBO2* (Supplementary Fig. 1 and Supplementary Table 6) were independently identified by an accompanying study that also demonstrated that recursive splicing in humans involves recursive exons rather than true zero nucleotide exons as in *Drosophila*. This suggests that although recursive splicing occurs in both *Drosophila* and humans, the precise nature of recursive splicing differs between these organisms. Nonetheless, the presence of ratchet points in both *Drosophila* *Hs6st* and its human orthologue *HS6ST3* indicates that recursive splicing is either very ancient or evolved independently.

The host genes containing recursively spliced introns are expressed in a broad spectrum of developmental time points, tissues and cell types—the recursive host genes are expressed at fragments per kb per million mapped reads (FPKM) >1 in 72%, 93% and 83% of cell lines, developmental time points, and tissues, respectively. However, host gene expression levels are quite dynamic throughout development and 63% have their peak expression in nervous system tissues (Supplementary Table 7).

Several lines of evidence suggest that recursive splicing is constitutive—specifically, when the host gene is transcribed, it is recursively spliced. First, we have been unable to detect lariat introns that would be generated by ratchet point skipping or the direct splicing of the flanking constitutive exons without recursive splicing. In our directed

Figure 2 | Identification of recursive lariat introns in *Drosophila*. a, RNA-seq reads (red and orange) that traverse a 5′ splice site—branch-point junction would align to the linear intron as out-of-order split reads. b, Example of recursive lariat introns in *cpo*. Shown are the recursive junctions identified (blue), the lariat junction reads (orange) and the overall RNA-seq read density from all samples (red). A magnification of each branch-point region is also shown along with the conservation among 16 insects. The positions of the branch points are indicated by the vertical arrows. c, Distribution of the distance of the recursive lariat intron branch points from the 5′ splice sites. d, Sequence logo of the recursive lariat intron branch-point sequences.

Figure 3 | Characteristics of *Drosophila* ratchet points. a, Sequence logos of 5′ splice sites, 3′ splice sites, ratchet points and non-ratchet point AG/GT sequences located in the same introns as ratchet points (top to bottom). b, Sequence conservation of ratchet points. Average PhastCons scores of ratchet points (green) and non-ratchet points (blue). Solid line indicates the average PhastCons score; shaded regions indicate the 95% confidence interval. c, d, Normalized recursive junction (c) reads and per cent non-recursive junctions (d) observed in untreated S2 cells and cells treated with the indicated dsRNAs.
Sites are associated with chromatin marks that would be expected specifically associated with ratchet points, yet the recursive splice indexing intron–exon boundaries. Of particular note, H3K4me3 (ref. 10), chromatin marks and particular features of gene architecture, including constitutive, although it remains possible that regulated or alternative highest and lowest number of mapped reads, respectively (Extended tissue samples and weakest among the cell lines, which have the tissue samples and weakest among the cell lines, which have the highest and lowest number of mapped reads, respectively (Extended Data Fig. 4), indicating that low correlation is related to sequencing depth. Together, these results strongly suggest that recursive splicing is constitutive, although it remains possible that regulated or alternative ratchet points may be identified in the future. Recent studies have demonstrated strong associations between chromatin marks and particular features of gene architecture, including intron–exon boundaries. Of particular note, H3K4me3 (ref. 10), H3K79me2 (ref. 11) and H3K36me3 (ref. 12) have been shown to specifically transition near intron–exon boundaries in humans. We inspected ChIP-seq data obtained from whole larvae to determine whether any chromatin marks are associated with ratchet points (Extended Data Fig. 5). None of the chromatin marks we examined is specifically associated with ratchet points, yet the recursive splice sites are associated with chromatin marks that would be expected given their position relative to canonical exons.

Here we provide experimental evidence that 130 Drosophila introns, 26 times the number previously known, are removed in multiple, sequential steps by recursive splicing, rather than by a single splicing event. We also identified five ratchet points in four human genes, including one case of orthologous Drosophila and human genes, indicating that recursive splicing evolved long ago. We do note, however, that recursive splicing in Drosophila involves true zero nucleotide exons but that an accompanying paper9 has demonstrated that recursive splicing in humans involves recursive exons, pointing to mechanistic and perhaps functional differences in this process between flies and humans. The ratchet points involved in recursive splicing are highly conserved and share sequence similarity with one another. While recursive splicing clearly occurs in Drosophila, its function and mechanism remain elusive, although we provide evidence that U2AF is required for recursive splicing. It also remains unknown why some Drosophila introns are recursively spliced and others are not. Further investigation will be necessary to determine whether recursive splicing is required for the function of the host gene in Drosophila and how the upstream exons re-engage in subsequent splicing reactions.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Acknowledgements** This work was supported by NHGRI grant U54HG006994 to S.E.C. and B.R.G. (co-PI), and R01 GM095296 to B.R.G.

**Author Contributions** B.R.G. and S.E.C. supervised data production. S.O., A.O., M.B. and B.R.G. performed computational analysis. B.R.G. wrote the paper with input from all authors.

**Author Information** Sequences are available from http://www.encodeproject.org and have been deposited in the Short Read Archive under accession numbers SRP056962, SRP056966 and SRP056969 and those listed in Supplementary Table 1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.R.G. (graveley@uchc.edu).
METHODS

No statistical methods were used to predetermine sample size.

RNA collections. The D. melanogaster RNA samples used for this study have been described previously17. RNA was isolated from D. simulans, D. sechellia, D. yakuba, D. pseudoobscura, D. virilis, and D. willistoni, using TRIzol. Total RNA from 20 human tissues was obtained from Clontech (catalogue no. 636643). RNA from the RNAi experiment was extracted from Drosophila S2 cells (Life Technologies, catalogue no. R690-07) treated with 20 μg of dsRNA for 5 days. The S2 cells were not authenticated or tested for mycoplasma contamination. The following dsRNA sequences were used for the RNAi experiments: Idrb (AAGCTTAGACATGCTAACCTTCAGAAGCAGTGAGGAAGGACGAGGAAAAGAGGGAGATGATGTAAGATATGTCGATCTTAGCATCACAG). CTGGTTAAGGATGCCGCTACTGGGTTGAGTAAGGGTTATGCTTTCTGACGATTACCAGCCCATGCCGGGTATAACAGATACGCCGGCAATTAAGTTCCTCGAATTCCGATCGATTGATGAAACCACCCAGGCCATGGCATTGGCAGTCCCGTCTTGGCATGCCAAATTAACTTGGACAAAAACTTTGCTAGTTCTTCAACCAACAGATGCATTTAGTTGGGCTCGCCCAGGCGGCCCTCCGGTGCCGACGTCGTGCTTTCCTGCAAGTGGTAGAGGTACCCAGTGACGTGAAGACGAGGAAAGGGAGAAGGTAAAGAAAGCTGCTCCTGTACCTC.

We generated a set of potential ratchet points. We next compared the lists of potential ratchet point junctions identified by analysis of the TopHat alignments and by alignment to all potential ratchet point junctions. This resulted in a list of 356 potential ratchet point junctions that were then individually examined on the genome browser to verify their identity. To facilitate this analysis we merged the bedGraphs from all of the TopHat alignments into one positive- and one negative-strand-specific bedGraph file. For each intron containing potential ratchet points, we calculated a robust regression line of the read density of each segment of the intron from the merged bedGraph profile using the robustfit function of MATLAB. This required masking out repeatMasker regions and over-lapping annotated features, both of which confused the regression process, and calculating the robust regression lines based on the remaining unmasked portions of the segment. We used MATLAB to generate a ‘flip-book’ of browser-like images of each potential ratchet junction with bedGraph and local robust regression plots superimposed to aid in the manual inspection of each ratchet point for verification.

The merged bedGraphs were loaded into the genome browser along with tracks of all 356 potential ratchet point splice junctions. In addition, we loaded the FlyBase 5.45 annotation, which was the most recent version of FlyBase at the time of this analysis, as well as the most recent modEncode annotation (MDv3) to identify ratchet points that match to exons identified more recently than the modEncode annotation (MDv1) used to seed the alignments. Finally, we loaded in the modEncode CAGE data to assess whether any potential ratchet points corresponded to previously unannotated promoters, which could also give rise to a saw-tooth pattern of RNA-seq read density. Potential ratchet points were removed if the 3′ ratchet point junction corresponded to an annotated exon, or if the saw-tooth pattern of read density was not apparent on the browser or from the local robust regression plots. During the course of this manual inspection we identified two ratchet points in lunar and mbl that were not identified in this computational analysis. Both of these were present in introns that contained other computationally identified ratchet points. We identified these based on their strong pattern of saw-tooth read density in both the browser and the local robust regression plots and the fact that they had conserved AG/GT sequences at the ratchet point junctions. These were removed in the computational analysis because they did not pass the stringent filters used. Both of these were experimentally validated and included in the analysis of all ratchet points. In total, the final list of ratchet points consists of 197 ratchet points (Supplementary Table 2). The same approach was used to manually review the putative ratchet points identified in the human RNA-seq data.

Validation of recursive splice sites by RT–PCR. 500 ng of total RNA isolated from S2 cells with TRIzol was used to synthesize cDNA using SuperScript II Reverse Transcriptase (RT) kits according to the manufacturer’s protocol. PCR amplification was performed using Phusion High-Fidelity DNA Polymerase (NEB), using the manufacturer’s instructions, with the following amplification program: 98 °C for 3 min, followed by 40 cycles 98 °C for 10 s, 55 °C for 30 s, and 72 °C for 20 s. Ratchet points were first confirmed by gel electrophoresis followed by Sanger sequencing.

Validation of recursive splice sites by cross-species RNA sequencing. The coordinates of the D. melanogaster ratchet points, and 50 nt on either side, were lifted over to the D. simulans (droSim1), D. sechellia (droSecl1), D. yakuba (droYak2), D. pseudoobscura (dp4) and D. virilis (droVir3) genomes, respectively. We used the same method and parameters, but without a reference transcriptome annotation. The human RNA-seq data sets were aligned to hg19 using the same method and parameters and Gencode v19 as a reference annotation.

PARSING TOPHAT ALIGNMENTS TO IDENTIFY RATCHET POINTS. We identified sets of novel splice junctions from the TopHat alignments which share the same 5′ splice site. Ratchet point junctions were kept from a sample if the 3′ splice site of the junction is an AG/GT, the 3′ splice site was unannotated in the previous modENCODE annotation, and the distance to the previous splice junction and the next splice junction is longer than 2 kb.

De novo identification of ratchet points. We generated a set of potential ratchet point junctions by joining 95 nt of each exon to the 95 nt downstream of every unannotated 3′ splice site (AG/GT) in the downstream intron. We aligned reads 1 and 2 of the total RNA reads independently to the database of all possible ratchet point junctions using Bowtie version 0.12.7 with the following options: -v 2 -k 5 -M 5–best. As the paired-end reads were aligned separately, post-processing was used to enforce constraints on gene-strand and alignment-strand that are a consequence of the stranded protocol. For each potential ratchet-point junction, we tabulated the coordinates of the genomic regions that comprise the ratchet-point junction sequences, the intron(s) and gene(s) the ratchet site is derived from, the number of alignments to the ratchet-point junction, the average number of mismatches per alignment, detailed offset and mismatch information, the alignment offset entropy, and the number of distinct offsets for only perfect alignments with ≥8 nt overlap. This latter parameter is intended to be a robust and conservative measure of alignment diversity. Ratchet points contained in introns that overlap no other distinct introns or any exons were filtered to require ≥3 perfect alignments to three distinct offsets and overlap ≥8 (or two perfect alignments to two distinct offsets and overlap ≥8 and ≥10 general alignments with ≤2 mismatch and overlap ≥5). Ratchet points in introns that do overlap other introns or exons were filtered using slightly more restrictive criteria and required at least five zero-mismatch ratchet-junction alignments with distinct offsets.
the last 94 nt of an intron to the first 94 nt of the intron. We used 94 nt of each portion of the intron to enforce a minimum of a 6-nt overhang when aligning 100-nt reads. Because the precise location of the branch points are not known ahead of time, and because a previous study in human has shown that most known branch points occur between 18 and 35 nt upstream of the 3′ splice site,15, we generated 100 possible lariat junctions by sliding a window of the 3′ end of the intron in the 5′ direction at 1-nt intervals, and fusing each to the first 94 nt of the intron. We generated the potential lariat junction databases for each segment of all recursive introns, as well as all possible permutations of these segments. For example, for an intron with two ratchet points, we generated potential lariat junctions for the first, second and third recursive segments as well as the introns from the first 5′ splice site to the second ratchet point (segments 1 and 2), from the 5′ splice site of the first ratchet point to the last 3′ splice site (segments 2 and 3), and from the first 5′ splice site to the last 3′ splice site (segments 1, 2 and 3).

Bowtie version 0.12.7 was used to generate an index of the potential recursive lariat junctions and to align all of the total RNA-seq reads, where each mate pair was aligned separately, with the following parameters: -v 2 -p 8—all—quiet. In total, 72,712 alignments were reported. These were then filtered for reads that mapped uniquely to the lariat junction database, then sorted by the number of nucleotides overlapping the lariat junction. Randomly selected reads with various extents of overlap were aligned to the D. melanogaster genome using BLAT to determine whether or not they mapped elsewhere in the genome. From this we determined that all reads that overlapped the lariat junction with fewer than 14 nt also aligned elsewhere in the genome. We therefore used BLAT to align all reported reads from the Bowtie alignment and discarded those which mapped elsewhere resulting in a total of 46 reads. We identified the approximate locations of the branch points based on the coordinates of the 94-nt segment of the 3′ end of the intron that the reads aligned to.

Validation of recursive lariats. Thirty-eight recursively spliced genes were selected for lariat analysis, choosing genes that were highly expressed in the nervous system. These genes contain 95 distinct segments corresponding to 52 ratchet points. To detect splice lariats for these ratchet points, ‘outward-facing’ PCR primer pairs were designed to amplify through branch points. The PCR primers contained overhangs so that Illumina clustering, indexing and sequencing oligonucleotides could be added in a subsequent nested PCR. The same primers designed for ratchet point lariat amplification were also used in different combinations to attempt to amplify the branch-point lariats that would be generated by skipping one or more ratchet points.

Total RNA was extracted from whole D. melanogaster (Bloomington Drosophila Stock Center strain no. 2057) using TRizol reagent (Invitrogen, Grand Island, NY), followed by cDNA synthesis using Superscript II (Invitrogen) primed with random hexamers. The primer pairs described above were used for the first-round PCR, after which products were visualized on an agarose gel. For 13 genes, no product was detected for any of the sub-introns targeted and these genes were not analysed further. For the other 25 genes, which contained a total of 64 sub-introns, we obtained a PCR product for at least one sub-intron. For these 25 genes we also used the primers in combinations to amplify any potential lariats created if splicing skipped ratchet points in a total of 41 combinations. Regardless of whether or not a product was visualized we prepared sequencing libraries from all reactions in an effort to capture any low-level lariats. Nested PCR was performed to add the Illumina sequencing oligonucleotides to the first-round PCR products. The amplicon libraries were pooled, purified and size selected to select amplicons between 300 to 1,000 bases in length. The pooled amplicon library was then sequenced on an Illumina MiSeq using a V3, 600-cycle kit to produce 200 by 400 bp paired-end reads.

Reads were filtered to remove mis-primed sequences and then aligned to the target genes using BLAT. Mapped reads were manually reviewed on the genome browser and a lariat considered to be ‘confirmed’ if a portion of a read aligned precisely to the 5′ splice site (ratchet point or exon) and to a second region within 100 bp upstream of the 3′ splice site (ratchet point or exon) in a continuous manner. This analysis confirmed 14 of the 64 recursive segments examined. Importantly, none of the primer combinations used to detect ratchet point skipping yielded any reads corresponding to a lariat. The few sequence reads that were obtained for these controls were PCR artefacts that primarily contained Illumina sequencing oligonucleotides and had no homology to the target genes. Sequence logos were generated with WebLogo16.

Comparison of gene expression and recursive splicing. The total number of reads mapping to each ratchet point junction was tabulated for each library and then summed across biological and/or technical replicates for each biological sample. The recursive index was calculated as the number of ratchet point junction reads per mapped reads per billion reads ((ratchet point junction reads/mapped reads) × 1,000,000,000). The RNA-seq data generated for this study is from total RNA, a mixture of mRNA, nascent RNA and pre-mRNA. Owing to the complexity of RNA types and intronic reads it is difficult to accurately quantify gene expression levels when using most existing software. We therefore used the expression values calculated from the corresponding poly(A)+ RNA-seq data that was previously generated14 from the same RNA samples. Comparisons of the recursive index and gene expression levels were performed using R (http://www.r-project.org). GO analysis was performed using Funcasssorte 2.0 (ref. 17).

Analysis of chromatin marks. Visualizations of chromatin marks at recursively spliced genes were generated using custom R scripts (http://www.r-project.org). We obtained ChIP-seq scores (http://encode-x.med.harvard.edu/data_sets/chromatin/) and Affymetrix tiling array gene expression scores (http://intermine.modencode.org/) generated from 1.3 larvae via the modENCODE projects. For each feature of gene architecture illustrated, mean ChIP-seq scores were calculated for non-overlapping bins of 200 bp in length.

As expected, transcription-associated marks were specific to genes actively transcribed in larvae (Supplementary Fig. 5b). At these active genes, we observed low levels of H3K4me3 near recursive splice sites compared to first exons (Supplementary Fig. 5b), which suggests that the saw-tooth patterns observed by total RNA-seq were not due to cryptic transcription initiation or unannotated promoters, but rather, co-transcriptional splicing. We also observed lower levels of H3K36me3 near recursive splice sites compared to downstream exons (Supplementary Fig. 5b). Since the degree of H3K36me3 has been shown to increase with each internal exon in humans11, the low levels of H3K36me3 seen at recursive splice sites may reflect the fact that recursive splices are typically located in 5′ introns, and thus, preceded by few internal exons (Supplementary Fig. 5c). Indeed, recursive splice sites were associated with high levels of H3K79me2 exons, which is typical of long 5′ introns in humans11.

Code availability. Custom code used in this paper is available without restrictions at https://github.com/graveley/Recursive.

13. Trappnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111 (2009).
14. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).
15. Taggart, A. J., DeSimone, A. M., Shih, J. S., Filloux, M. E. & Fairbrother, W. G. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. Nature Struct. Mol. Biol. 19, 719–721 (2012).
16. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190 (2004).
17. Berriz, G. F., Beaver, J. E., Cenik, C., Tasan, M. & Roth, F. P. Next generation software for functional trend analysis. Bioinformatics 25, 3043–3044 (2009).
18. Tilgner, H. et al. Nucleosome positioning as a determinant of exon recognition. Nature Struct. Mol. Biol. 16, 996–1001 (2009).
Extended Data Figure 1 | Two approaches for identifying recursive splice sites. 

**a**, Identification of recursive splice sites by parsing alignments. RNA-seq reads were mapped to the genome using TopHat in a manner that allowed for novel splice junctions to be predicted. The alignments were then parsed for splice junction reads where the 5‘ splice site mapped to an annotated 5‘ splice site, but the 3‘ splice site was unannotated. 

**b**, De novo identification of recursive splice sites. A database was generated in which each annotated 5‘ splice site was spliced to all AG/GT sequences in an intron that did not correspond to an annotated 3‘ splice site. All RNA-seq reads were aligned to this database and the alignments parsed to find cases where reads mapped perfectly with at least three distinct offsets and at least an 8 nt overhang.
Extended Data Figure 2 | Characteristics of Drosophila ratchet points.

a. Distribution of the number of ratchet points per recursive intron. b. Size distribution (log10(bp)) of all (red) and recursive (blue) introns. c. Size distribution (in kb) of the individual intron segments removed by recursive splicing binned by the number of segments per intron.
Extended Data Figure 3 | RT–PCR validations of Drosophila recursive splicing events. a–m, RT–PCR validation of ratchet points (red dots) from the indicated genes using primers in the upstream constitutive exon and flanking the putative ratchet points. The RP primers are expected to yield RT–PCR products if the constitutive exon is spliced to the ratchet point. The URP primers, which are upstream of each ratchet point, serve as negative controls. The identity of all RT–PCR products was verified by Sanger sequencing. Although the URP control RT–PCR reactions yielded a product for hppy RP1 and pum RP2, we were not able to generate sequence from them and therefore consider them to be amplification artefacts.
Extended Data Figure 4 | Number of mapped reads per sample used for gene expression analysis.
Extended Data Figure 5 | Chromatin marks associated with recursive splice sites.  

**a**, Examples of chromatin marks at the *luna* gene locus, which contains 5 recursive splice sites (red triangles) within a single long intron. **b**, Heat maps show relative ChIP-seq enrichment for H3K4me3 (top, red), H3K79me2 (middle, green) and H3K36me3 (bottom, blue), within 2 kb of the indicated gene features from 171 genes containing at least one ratchet point. Heat maps are centred around gene features, which include the transcription start site of the first exon (first exon, arrow), the 5’ splice site of the exon upstream of the recursive splice site (upstream exon, black rectangle), the ratchet point (red triangle), the 3’ splice site of the exon downstream of the recursive splice site (downstream exon, black rectangle), and the poly(A) site of the last exon (last exon, red octagon); the average exon of each gene feature is drawn to scale. Genes are sorted from top to bottom by decreasing expression level. For genes containing more than one ratchet point, the first, upstream, downstream and last exons are represented multiple times. **c**, Histogram illustrating the intron positions the ratchet points reside in based on RefSeq annotations.
Extended Data Table 1 | Summary of recursive intron lariats identified by directed RT–PCR and sequencing

| Gene | Segment | Coordinate of putative branchpoint | Distance upstream of 3' splice site | Sequence surrounding bp | Identified in Total RNA-Seq Data |
|------|---------|------------------------------------|-----------------------------------|-------------------------|---------------------------------|
| Sdc  | segment3 | 2R:17299944                        | 21                                | CATCTCACCTCATAATGTTTT    | No                              |
| apo  | segment1 | 3R:13803044                        | 34                                | AAGGTAACTAATATGTTTTTTTT  | Yes                             |
| apo  | segment2 | 3R:13815266                        | 31                                | CCAAATGCTAATTTATACCTT    | Yes                             |
| apo  | segment3 | 3R:13832619                        | 37                                | AGCAATCATCTAAGGTTCTTC    | Yes                             |
| bun  | segment2 | 2L:12498256                        | 32                                | CAACATACTTACAGAACCCTTT   | No                              |
| CG7029| segment2 | 3R:18592273                        | 55                                | GTTTGTCACACAGAGTTGCTGC   | No                              |
| nuf  | segment2 | 3L:14223246                        | 29                                | ATATAGACTTTACAGGTTCTCT   | No                              |
| CG31637| segment2| 2L:6525343                         | 23                                | GAGATTCTAAACAGGTTTCTC    | Yes                             |
| daily| segment1 | 3L:8843654                         | 26                                | CTAAATCTGTGCTTTATCTCT    | No                              |
| daily| segment2 | 3L:8855584                         | 45                                | AATTTGACACATCGTACATGCT    | Yes                             |
| daily| segment3 | 3L:8870223                         | 29                                | ATGCAAGCTGATCTGTTTCT     | Yes                             |
| Mmp2 | segment2 | 2R:5503040                         | 33                                | TAGCATGCTGATATCATGTGTT    | No                              |
| osp  | segment2 | 2L:14656399                        | 30                                | AAACCAAACTAATTTTCTACC    | No                              |
| osp  | segment1 | 2L:14677529                        | 41                                | ACATCTTCTTACTAATATT      | No                              |

For each recursive lariat confirmed by RT–PCR and sequenced, the following information is provided: gene, segment, coordinate of the putative branchpoint, distance upstream of the 3’ splice site the branch point is located, the sequence surrounding the branch point, whether the lariat intron is newly validated with respect to the total RNA analysis described in Supplementary Table 4a.
### Extended Data Table 2 | Summary of total RNA-seq data from *ldbr* and *U2AF* RNAi experiments

| Sample                          | S2 Untreated | ldbr dsRNA | U2af38 dsRNA | U2af50 dsRNA | U2af38 & U2af50 dsRNA |
|---------------------------------|--------------|------------|--------------|--------------|-----------------------|
| Total reads                     | 183,451,384  | 146,237,470| 174,626,770  | 162,012,182  | 170,417,376           |
| Mapped reads                    | 66,597,777   | 54,762,604 | 56,120,630   | 43,620,982   | 61,437,507            |
| % mapped                        | 36.30%       | 37.45%     | 32.14%       | 26.92%       | 34.24%                |
| # junctions                     | 44,860       | 42,736     | 52,228       | 46,184       | 50,883                |
| junctions/1M reads              | 673.60       | 780.39     | 930.64       | 1058.76      | 828.21                |
| # total junc reads              | 7,402,459    | 4,925,448  | 5,528,482    | 3,935,048    | 5,999,468             |
| % junction reads                | 11.12%       | 8.99%      | 9.65%        | 9.02%        | 9.77%                 |
| # novel junction reads          | 2,369,949    | 1,575,057  | 1,307,830    | 919,616      | 1,273,867             |
| % novel junction reads          | 3.56%        | 2.88%      | 2.38%        | 2.11%        | 2.07%                 |
| # annotated junc reads          | 5,032,510    | 3,350,391  | 4,190,632    | 3,015,432    | 4,725,601             |
| % annotated junction reads      | 7.56%        | 6.12%      | 7.47%        | 6.91%        | 7.69%                 |
| # novel junction cases          | 22,688       | 20,977     | 29,211       | 24,279       | 28,193                |
| novel/total junctions           | 0.51         | 0.49       | 0.56         | 0.53         | 0.55                  |
| # annotated junc cases          | 22,172       | 21,759     | 23,017       | 21,905       | 22,690                |
| annotated/total junctions       | 0.49         | 0.51       | 0.44         | 0.47         | 0.45                  |
| #RPs on both strands            | 119          | 100        | 81           | 64           | 0                     |
| RP/million reads                | 1.79         | 1.83       | 1.44         | 1.47         | 0.00                  |
| Total RP junc reads             | 1,449        | 985        | 377          | 259          | 0                     |
| RP junc reads/million           | 21.76        | 17.99      | 6.72         | 5.94         | 0.00                  |

For each data set the following information is included: total reads, mapped reads (from TopHat alignments), % mapped, number of total splice junctions, normalized junctions, total number of splice junction reads, per cent splice junction reads, number of novel junction reads, per cent novel junction reads, number of annotated junction reads, per cent annotated junction reads, number of novel junction cases, ratio of novel to total junctions, ratio of annotated to total junctions, number of ratchet points, normalized ratchet points (ratchet points per million reads), total ratchet point junction reads, ratchet point junction reads per million reads.
### Extended Data Table 3 | Summary of total RNA-seq data from related *Drosophila* species

| Species            | Total reads | Mapped reads | % mapped |
|--------------------|-------------|--------------|----------|
| *D. simulans*      | 92,949,544  | 45,797,569   | 49.27%   |
| *D. sechellia*     | 38,674,020  | 19,974,783   | 51.65%   |
| *D. yakuba*        | 46,002,798  | 18,300,604   | 39.78%   |
| *D. pseudoobscura* | 45,274,524  | 14,457,616   | 31.93%   |
| *D. virilis*       | 46,675,210  | 18,544,567   | 39.73%   |
| **Total**          | **269,576,096** | **117,075,139** |          |

For each data set the following information is included: species, total reads, mapped reads (from TopHat alignments), % mapped.
## Extended Data Table 4 | Summary of ratchet points experimentally validated in other *Drosophila* species

| Species       | Total liftovers | number of ratchet junction reads | distinct ratchets | % validated |
|---------------|----------------|----------------------------------|-------------------|-------------|
| *D. sechellia*| 163            | 93                               | 74                | 45.40%      |
| *D. simulans* | 166            | 106                              | 87                | 52.41%      |
| *D. yakuba*   | 150            | 52                               | 43                | 28.67%      |
| *D. pseudoobscura* | 78          | 15                               | 15                | 19.23%      |
| *D. virilis*  | 40             | 18                               | 18                | 45.00%      |

For each species analysed, the following information is provided: species, total leftovers (the number of *D. melanogaster* ratchet point coordinates that were successfully lifted over to the other species genome coordinates), number of ratchet junction reads (the number of reads that mapped to the AG/GT sequence of the lifted over ratchet points), distinct ratchets (the number of distinct ratchet points identified by the mapped reads), % validated (the per cent of lifted over ratchet points with at least one mapped read).

©2015 Macmillan Publishers Limited. All rights reserved
## Extended Data Table 5 | Summary of total RNA-seq data from human tissues

| Tissue               | Total Reads | Total Aligned Reads | Unique Aligned Reads | % Uniquely Aligned |
|----------------------|-------------|---------------------|----------------------|-------------------|
| adrenal-gland        | 75,843,326  | 66,482,852          | 59,958,580           | 79.06%            |
| brain-cerebellum     | 70,868,098  | 64,382,088          | 60,067,066           | 84.79%            |
| brain-whole         | 75,684,718  | 63,391,490          | 59,015,902           | 77.98%            |
| fetal-brain         | 77,532,892  | 65,028,630          | 58,748,974           | 75.77%            |
| fetal-liver         | 69,002,806  | 47,100,928          | 38,162,898           | 55.23%            |
| heart                | 79,200,464  | 70,490,860          | 65,388,604           | 82.56%            |
| kidney               | 65,109,248  | 57,411,382          | 52,917,602           | 81.28%            |
| liver                | 67,066,550  | 56,934,448          | 50,433,480           | 75.18%            |
| lung                 | 66,115,012  | 56,503,662          | 51,846,946           | 78.42%            |
| placenta             | 70,985,882  | 62,429,622          | 56,803,382           | 80.02%            |
| prostate             | 66,767,818  | 62,197,480          | 56,967,700           | 82.84%            |
| saliva-gland         | 77,811,154  | 73,634,924          | 55,954,594           | 71.91%            |
| skeletal-muscle      | 75,606,186  | 63,492,280          | 59,413,946           | 78.58%            |
| small-intestine      | 72,492,290  | 62,291,298          | 57,178,572           | 78.88%            |
| spleen               | 82,675,794  | 62,553,250          | 55,318,774           | 66.91%            |
| stomach              | 73,018,660  | 62,932,016          | 56,174,974           | 76.93%            |
| thymus               | 72,695,752  | 67,467,012          | 60,657,380           | 83.44%            |
| thyroid              | 75,794,702  | 68,200,264          | 63,259,560           | 83.46%            |
| trachea              | 70,890,274  | 58,922,766          | 53,022,564           | 74.80%            |
| uterus               | 72,876,872  | 61,974,024          | 56,835,240           | 77.99%            |
| total                | 1,460,148,498 | 1,257,831,176     | 1,128,146,738        |                   |

For each data set the following information is included: sample, total reads, total aligned reads, uniquely aligned reads, % uniquely aligned.