Legends to Supplementary Tables and Figures

Supplementary Table S1. All Putative PTES structures identified by High Throughput Sequencing. Accession number, gene name, donor and acceptor exon (relative to longest RefSeq isoform), samples where structure was identified, read number, read names and read sequences are provided for all putative PTES structures identified. The sizes of donor and acceptor exons/introns are also shown.

Supplementary Table S2. PTES structures validated by RT-PCR. R= number of reads. F= Frame of novel exon-exon junction: In-in frame, Out-frameshift, UTR-non coding sequence at start of exon). M- multiple amplicons observed in RT-PCR. DNA-amplicon of expected size present in DNA. Expression levels in cDNA derived from Total/PolyA+ RNA are as follows: O-absent, X- present with no clear enrichment, XX-present and enriched. For all primers used see Supplementary Table S5.

Supplementary Table S3. *In silico* analysis of PTES read frequency. The number of sequence reads spanning PTES and canonical splice junctions in all 4 ALL samples is shown, together with the total number of reads mapping to each gene, average coverage per base, and total coverage of each gene at 1X depth. Totals for all 4 samples, and the average read depth/coverage for each gene are also shown.

Supplementary Table S4. Structures common to Dixon et al. PTES structures identified in this study and by Dixon et al (2005) are shown, together with genes common to both studies where distinct PTES structure have been identified.

Supplementary Table S5. All RT-PCR primers. Gene name, accession number, exons spliced, primer sequences and expected amplicon sizes are shown for all PTES structures analysed in the validation study.

Supplementary Table S6. Other Primers. Primers used for the UTR analysis (Figures 3 and S3), Real Time PCR (Figures 5, S4 and S5), Southern analysis (Figure 6), Comparative analysis (Figure 4), and the subcloning of templates used in the *in vitro* transcription (Figures 5 and S5) are all shown. Where appropriate, amplicon size, amplicon sequence, efficiency of reaction, and primer concentration used is also provided.

Supplementary Figure S1. Frequency of PTES donor and acceptor exons. Frequency plot of donor and acceptor PTES exons with respect to transcription start (exon 1). Exon numbers above 22, which occur in 5 PTES structures, are not shown (*BPTF* E28-E23, *KIAA1109* E55-E22, *UBR5* E37-E36, *UTRN* E44-E28 and *WDFY4* E39-E35, see Supplementary Table S1).

Supplementary Figure S2. Sequence and secondary structure of *CCDC66* PTES transcript.

A. Sequence relationships within *CCDC66* intron 10 and intron 4. Dot plot generated using Megalign (Lasergene software, DNAStar Inc. Madison, WI, USA) - window size: 5bp, match: 100%. Matches are coloured dynamically to highlight the longest matches (red) relative to medium (green) and short (blue). Alignment of these introns identifies a 50bp region of 98% identity between an *Alu*Y element present at the breakpoint in I4, and an *Alu*Y element lying in the opposite orientation

approximately 200bp downstream of the breakpoint in I10 (diagonal 1). The breakpoint in I10 is defined by a related, but degenerate sequence which creates an 18-21bp region of high sequence identity in the same orientation as the sequence in intron 4. The sequence identity at the I10-I4 breakpoint is also highlighted (diagonal 2). Approximate positions of exons and the inverted intronic *Alu*Y elements are also shown. Direction of transcription is indicated with arrows. Sequence alignment of the 3 regions highlighted by boxes 1 and 2 is also shown and includes the terminal ~30bp of the 50bp identity within the *Alu*Y elements [I10(2rev) and I3(rev)], the imperfect 18-21bp match to this sequence in I10 [I10(1)] and the *CCDC66* I4-I10 cDNA. The 18-21bp imperfect repeat is shown in bold and sequences flanking the breakpoint region in the cDNA are shown in red to highlight their origins in I10 and I4. Rev indicates reverse complement. Because of the overlapping sequence identity, the precise breakpoint within the I10-I4 cDNA cannot be defined, although no canonical splice donor or acceptor sites are present. In RNA, complementary base pairing between the two 50bp sequences could bring E10 and E5 together to promote the E10-E5 splice which was originally identified by high throughput sequencing and subsequently validated. However, displacement of this complementary base pairing by the degenerate sequence match in E10 could give the secondary structure shown in Figure 3B, which brings the I10 and I4 sequences fused in the *CCDC66* I10-I4 cDNA into alignment as single stranded RNAs.

B. Secondary structure aligns the I10 and I4 fusion points and brings E10 and E5 into close proximity. The I10-I4 cDNA sequence is created by joining the two single stranded sequences shown (sequence in red is to facilitate orientation with Fig. 3A). The I10 hairpin beginning AGCCACC can be displaced by the highly similar sequence from I4 to restore the 50bp match between the two *Alu*Y elements. While it is possible that this secondary structure could promote aberrant splicing to generate the I10-I4 cDNA, it has been shown that such structures which contain short regions of sequence identity can promote template switching during reverse transcription (Coquet et al. 2006; Houseley and Tollervey 2010).

Supplementary Figure S3. Amplification to terminal exons using PTES junction specific primers. Amplicons generated using primers specific for *LARP1B* E4-E2 and *UBAP2* E10-E5 junctions are shown. In both cases primer pairs 1 and 2 amplify from the terminal 5' exon to the junction specific primer. Primer pairs 3 and 4 amplify from the junction specific primer to the terminal 3' exon. Templates are NB3 - cDNA where the PTES structures were originally identified, +ive - unrearranged (canonical) cDNA (AK092764 and AK294952 respectively), -ive - no template. The exon organisations of the inferred PTES RNAs, of the full length canonical genes from RefSeq (canonical), and of the +ive control templates are shown, together with the expected amplicon sizes. Due to the incomplete nature of the control templates, UTR primers specific for the NB3 and control templates were used in some cases (indicated as **a** and **b**). For clarity of presentation, exons 14-27 of *UBAP2* are not shown. For all primers see Supplementary Table S6.

Supplementary Figure S4. Correlation of Canonical and PTES exon junction levels. Δ-Ct values obtained for canonical (x-axis) and PTES (y-axis) exon junction products from 5 human genes in a variety of fetal and adult human tissues are shown. Values shown are relative to the mean expression level of 3 housekeeping genes (Δ-ct = target Ct-control Ct, see Methods). Because all genes analysed are expressed at lower levels than the controls, higher Δ-Ct value indicate lower expression levels. The

diagonal corresponds to a PTES threshold 3 cycles higher than canonical. Tissues used were: Fetal and adult spine, thalamus, lung and heart, and adult cerebral cortex. The levels of PTES and canonical junctions within transcripts are correlated ($r=0.674$, $p= 6.7 \times 10^{-6}$ for all data combined), correlation coefficients ranging from 0.96 ($p=0.0001$) for *RTN4* to 0.37 ($p=0.48$) for *TLE4*. The most extreme outliers with high PTES expression relative to canonical are circled. These are fetal heart and fetal brain (*PHC3*) and fetal spine and thalamus (*CDK13*). For all primers see Supplementary Table S6.

Supplementary Figure S5. Real Time Standard Curves. Standard curves for 5 PTES transcripts found to be highly expressed by *in silico* or Δ-Ct analyses are shown, together with curves for the corresponding canonical transcripts. The slope is shown in each case. Templates were generated by *in vitro* transcription of clones PTES and canonical spice junctions (see methods). For all primers see Supplementary Table S6.