**Methods**

**Simulated data**

A simulation using Polyester [1] was performed: 10,000 RefSeq transcripts were randomly selected. The only restriction that was placed on transcript selection was that it could not come from a gene that had an exonic sequence shared across more than one gene (10,006 genes in the mm10 genome have exons with genomic coordinates that overlap at least one other gene). These transcripts represented 7,678 genes. Paired end reads were simulated for 6 independent replicates at 100× coverage. Read size is 56bp (matching the mouse neural data used in the manuscript). A list of the 10,000 transcripts is in Supplementary Table S1.1 (see file: Supplementary\_table\_S1\_1\_simulation1\_transcript\_list.xlsx).

A second simulation using Polyester [1] was performed using all 476 annotated RefSeq transcripts from 59 genes, and represents the scenario where multiple transcripts of a gene are expressed. Genes were selected to represent a range of scenarios (1) some transcripts with no unique events; (2) at least one exon unique to at least one transcript; (3) at least one exon fragment unique to at least one transcript; (4) a transcript with only a unique junction; (5) only two transcripts, of which one has a unique exon and the other has a unique junction; (6) only two transcripts with at least one junction each with alternative donors or acceptors; (7) five transcripts each with a unique exon; (8) gene has five transcripts, each with a unique donor site. These criteria represent a diverse set of alternative splicing event types. We also created nine “hybrid” transcripts, where exons from two transcripts of the same gene are joined by an unannotated junction but where the donor and acceptor sites are present in the annotation. Reads from all of these transcripts for the 59 genes were simulated as per the 10,000-transcript simulation outlined above. A list of the 59 genes is in supplementary table S1.2 (see file: Supplementary\_table\_S1\_2\_simulation2\_gene\_list.xlsx).

Both simulated datasets used in this study are included as supplementary data. Supplementary files X,Y

**Alignments, detection of events and identification of transcripts using Event Analysis**

For both simulation datasets, simulated reads were first aligned as single-ended reads to the set of cataloged junction sequences using the Bowtie algorithm (version 0.12.9) [2], allowing for only a single alignment per read (parameter “-m 1”) and for up to three mismatched nucleotides (“-v 3”). The "--tryhard" parameter was used to force Bowtie to find as many valid alignments as possible, and the "best" alignment in terms of stratum were reported using the options "--best --strata --chunkmbs 1024". Reads unmapped to junctions are then mapped to the genome (GRCm38/mm10 version for the mouse data, GRCh37/hg19 version (release 73) for the T1D data) using BWA-MEM (version 0.7.12) [3]. Reads mapping to exon fragments were identified using the BED file for the genomic regions defining these events. Coverage is summarized as the average number of reads per nucleotide (APN = number of reads aligning to region / region length). Event detection was defined as an APN > 0. Genes with no evidence of expression (i.e. all associated exonic regions had no coverage in any sample) were excluded, and the proportion of events detected for each transcript of the remaining set of genes was calculated.

**Junction identification with STAR**

For both simulation datasets, simulated reads were aligned to the RefSeq mm10 genome using the STAR aligner (version 2.5.4b) [4]. STAR references were generated using RefSeq mm10 GFF3 annotations, and a maximum junction overhang size of 40 bp either side of the junction was set for the creation of STAR's splice junction database, analogous to that of the junction catalog used by Event Analysis. To ensure that comparisons between STAR and Event Analysis were appropriate, STAR alignment parameters were also set to mimic those used by Bowtie for Event Analysis: reads were aligned to the mm10 genome using the alignment mode “EndToEnd” to force STAR to not soft-clip reads, the minimum allowable read junction overhang was set to 16 bp, up to three mismatches were allowed per alignment, and multimapping alignments were disallowed.

**Reassembly and estimation of simulated transcripts using iReckon**

For the 10,000-transcript simulation only, transcripts were reassembled and estimated using iReckon, version 1.0.8 [5]. Genome alignments and junctions identified using STAR (see above) were used as input. Annotations consisting of a 12-column BED file of all RefSeq transcripts (including genomic start positions and lengths of each exon) was used to guide transcript assembly. Alignments to transcript sequences were carried out using BWA-MEM (version 0.7.12) [3] as per the iReckon user guide, and all parameters were left at their default settings.

**Results Summary**

**Simulation 1: 10,000-transcript simulation**

In terms of the detection of junctions, we found that Event Analysis and STAR identify almost the same set of junctions, with Event Analysis detecting slightly more true junctions that STAR (Table S1.3). STAR detected an additional ~100 junctions that Event Analysis did not, and these are almost all complex multi-junction alignments involving microexons (Table S1.3).

We then compared the transcripts retained after eliminating unlikely transcripts with the reassembled transcriptome provided by iReckon. We found that Event Analysis correctly retains 93-99% of the simulated transcripts, depending on the level of detection (APN > 0, APN ≥ 5) and proportion of events detected (100% of events, ≥75% of events) (Table S1.4). By comparison, iReckon reassembled 80% of the 10,000 transcripts in at least one sample and only 66% in all samples, and instead assembled as many as ~9500 additional transcripts, including novel transcripts, intron-retaining transcripts and unspliced transcripts (Table S1.4).

**Simulation 2: 59-gene simulation**

Similar to the 10,000-transcript simulation, we found that Event Analysis and STAR identify almost all of the same junctions from the 476 RefSeq transcripts of the 59 genes selected for this simulation (Table S1.5), with Event Analysis detecting 17 annotated junctions that STAR did not. Of the nine unannotated junctions included in this simulation, Event Analysis identified all of them, while STAR missed three (Table S1.6).

For the 59 gene simulation, Event Analysis correctly retained more transcripts than iReckon correctly reassembles (Table S1.7). Using an event detection criteria of APN > 0 and requiring 100% of events detected, Event Analysis identified 90% of the 476 simulated RefSeq transcripts, as well one additional, unrelated transcript. When the event detection criteria was set to APN ≥ 5 and transcripts with at least 75% of events detected were retained, Event Analysis identified all but one of the 476 RefSeq transcripts and no unrelated transcripts (Table S1.7). iReckon correctly assembled 391 of the 476 simulated transcripts, and assembled an additional 565 transcripts of which most were novel (Table S1.7). When only transcripts in common to all samples were included, iReckon only correctly assembled 180 of the 476 transcripts (Table S1.7). However, in this case only 12 incorrect transcripts are assembled.

**Table S1.1. List of transcripts selected for Simulation 1**

See file: Supplementary\_table\_S1\_1\_simulation1\_transcript\_list.xlsx

**Table S1.2. List of genes and transcripts selected for Simulation 2**

See file: Supplementary\_table\_S1\_2\_simulation2\_gene\_list.xlsx

**Table S1.3. Summary of annotated junctions detected by Event Analysis and STAR for the 10,000-transcript simulation The total number of simulated junctions is 58,607. The number detected by Events Analysis, Star, the union and intersection of these approaches are given in the columns.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sample** | **Events total** | **STAR total** | **Events ∪ STAR** | **Events ∩ Star** | **Sensitivity (Events)** | **Sensitivity (STAR)** |
| Sample 1 | 58,190 | 57,833 | 58,281 | 57,742 | 99.3% | 98.7% |
| Sample 2 | 58,191 | 57,826 | 58,279 | 57,738 | 99.3% | 98.7% |
| Sample 3 | 58,196 | 57,833 | 58,284 | 57,745 | 99.3% | 98.7% |
| Sample 4 | 58,191 | 57,833 | 58,282 | 57,742 | 99.3% | 98.7% |
| Sample 5 | 58,195 | 57,834 | 58,284 | 57,745 | 99.3% | 98.7% |
| Sample 6 | 58,192 | 57,838 | 58,282 | 57,748 | 99.3% | 98.7% |
| All samples (intersection) | 58,167 | 57,739 | 58,249 | 57,657 | 99.2% | 98.5% |
| Any sample (union) | 58,216 | 57,928 | 58,320 | 57,824 | 99.3% | 98.8% |

**Table S1.4. Summary of Event Analysis and iReckon results for the 10,000-transcript simulation**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Method** | **Transcripts correctly identified** | **Related Refseq transcripts** | **Related non-Refseq transcripts** | **Unrelated transcripts** | **Missing transcripts** | **Precision** | **Sensitivity** |
| Event Analysis (100% events detected, APN>0) (all samples) | 9,272 | 5,115 | 0 | 343 | 728 | 96.4% | 92.7% |
| Event Analysis (≥75% events detected, APN>0) (all samples) | 9,937 | 24,969 | 0 | 399 | 63 | 96.1% | 99.4% |
| Event Analysis (≥75% events detected, APN≥5) (all samples) | 9,881 | 24,067 | 0 | 134 | 119 | 98.7% | 98.8% |
| iReckon (at least 1 sample) | 8,058 | 2,591 | 5,857 | 1,127 | 1,942 | 87.7% | 80.6% |
| iReckon (all samples) | 6,637 | 1,245 | 502 | 228 | 3,363 | 96.7% | 66.4% |

**Table S1.5. Summary of annotated junctions detected by Event Analysis and STAR for the 59-gene simulation**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Sample** | **Total** | **Events total** | **STAR total** | **Events ∪ STAR** | **Events ∩ Star** | **Sensitivity (Events)** | **Sensitivity (STAR)** |
| Sample 1 | 1,075 | 1,062 | 1,045 | 1,063 | 1,044 | 98.8% | 97.2% |
| Sample 2 | 1,075 | 1,062 | 1,045 | 1,063 | 1,044 | 98.8% | 97.2% |
| Sample 3 | 1,075 | 1,063 | 1,045 | 1,063 | 1,045 | 98.9% | 97.2% |
| Sample 4 | 1,075 | 1,063 | 1,045 | 1,063 | 1,045 | 98.9% | 97.2% |
| Sample 5 | 1,075 | 1,062 | 1,045 | 1,062 | 1,045 | 98.8% | 97.2% |
| Sample 6 | 1,075 | 1,063 | 1,045 | 1,063 | 1,045 | 98.9% | 97.2% |
| All samples (intersection) | 1,075 | 1,061 | 1,043 | 1,062 | 1,042 | 98.7% | 97.0% |
| Any sample (union) | 1,075 | 1,063 | 1,046 | 1,063 | 1,046 | 98.9% | 97.3% |

**Table S1.6. Summary of the simulated unannotated junctions detected by Event Analysis and STAR for the 59-gene simulation**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Sample** | **Total** | **Events total** | **STAR total** | **Events ∪ STAR** | **Events ∩ Star** | **Sensitivity (Events)** | **Sensitivity (STAR)** |
| Sample 1 | 9 | 9 | 6 | 9 | 6 | 100.0% | 66.7% |
| Sample 2 | 9 | 9 | 6 | 9 | 6 | 100.0% | 66.7% |
| Sample 3 | 9 | 9 | 6 | 9 | 6 | 100.0% | 66.7% |
| Sample 4 | 9 | 9 | 6 | 9 | 6 | 100.0% | 66.7% |
| Sample 5 | 9 | 9 | 6 | 9 | 6 | 100.0% | 66.7% |
| Sample 6 | 9 | 9 | 6 | 9 | 6 | 100.0% | 66.7% |
| All samples (intersection) | 9 | 9 | 6 | 9 | 6 | 100.0% | 66.7% |
| Any sample (union) | 9 | 9 | 6 | 9 | 6 | 100.0% | 66.7% |

**Table S1.7. Summary of Event Analysis and iReckon results for the 59-gene simulation**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Transcriptome** | **Transcripts correctly identified** | **Related non-Refseq transcripts** | **Unrelated transcripts** | **Missing transcripts** | **Precision** | **Sensitivity** |
| Event Analysis (100% events detected, APN>0) | 428 | 0 | 3 | 39 | 99.3% | 91.6% |
| Event Analysis (≥75% events detected, APN>0) | 466 | 0 | 3 | 1 | 99.4% | 99.8% |
| Event Analysis (≥75% events detected, APN≥5) | 466 | 0 | 2 | 1 | 99.6% | 99.8% |
| iReckon (union of all samples) | 391 | 546 | 19 | 76 | 95.4% | 83.7% |
| iReckon (intersection of all samples) | 180 | 12 | 0 | 287 | 100% | 38.5% |

**References**

1. Frazee AC, Jaffe AE, Langmead B, Leek JT: **Polyester: simulating RNA-seq datasets with differential transcript expression**. *Bioinformatics* 2015, **31**(17):2778-2784.

2. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.

3. Li H: **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM**. *arXic preprint arXiv:13033997* 2013.

4. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.

5. Mezlini AM, Smith EJM, Fiume M, Buske O, Savich GL, Shah S, Aparicio S, Chiang DY, Goldenberg A, Brudno M: **iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data**. *Genome Research* 2013, **23**(3):519-529.