**Summary:**

**Validation of iReckon-assembled transcripts with PacBio sequenced transcriptomes**

**Method:**

To determine how well iReckon reassembles transcripts, the set of transcripts identified using PacBio sequencing [38] is used for validation. Sequences for each iReckon transcript for each NPC replicate were extracted using BEDtools (version 2.17.0; [1]) from the output GTF files and compared to the complete set of PacBio transcript sequences using the MegaBLAST algorithm [55]. BLAST alignments were classified based on the length of the BLAST hit (100% of the iReckon transcript sequence matches a PacBio transcript, no more than 10bp different in length between iReckon and PacBio transcripts, at least 90% of the iReckon transcript sequence matches a PacBio transcript, at least 75% of the iReckon transcript sequence matches a PacBio transcript, at least 50% of the iReckon transcript sequence matches a PacBio transcript), and whether there are no mismatches or no more than 5 mismatches nucleotides. iReckon transcripts with gaps or multiple, fragmented hits to the same PacBio transcript were excluded.

**Results summary:**

We find that relatively few of the 5,880 PacBio transcripts (401-562 transcripts, 7-10%) match to the iReckon-assembled transcriptome (Table S4.1), and many of those identified are present in only one of the two NPC replicates. In contrast, Event Analysis is able to identify 5,686 of the 5,880 (97%) PacBio transcripts (at least 75% events detected at an event-detection threshold of APN > 0). Even if only RefSeq transcripts with 100% of their events detected (APN > 0) are considered, 77% of the PacBio transcriptome is identified using Event Analysis, substantially higher than that assembled with iReckon (Table S4.1).

When comparing the number of iReckon transcripts with PacBio or RefSeq hits, we observed that relatively few iReckon transcripts matched to PacBio transcripts (13%) or RefSeq transcripts (10%) (Table S4.2). The slightly higher proportion of BLAST hits to the PacBio transcriptome likely reflects 5'- and 3'-UTR length differences between PacBio transcripts and RefSeq transcripts in these data, as previously reported [2]. We also find that relatively few iReckon transcripts (~5%) are also represented by Event Analysis reduced references (Table S4.3).

We used the 23,438 distinct iReckon transcripts identified from either NPC replicate and classified them on the basis of their identifier assigned by iReckon: “Known” (i.e. has a RefSeq identifier), “Intron Retention” (at least one intron is retained), “Novel” (novel transcript not matching to an annotated transcript), and “Unspliced” (transcripts where all intervening intron sequences are not spliced out). We found that the majority of iReckon transcripts are novel transcripts (60%), novel intron-retaining transcripts (11%), or novel unspliced transcripts (22%), with the remainder matching to known RefSeq transcripts, with or without intron retention (Table S4.4). As the set of transcripts present in these samples is known (PacBio transcript), this suggests that iReckon has difficulty in accurately reconstructing the transcripts present in these data.

**Table S4.1. PacBio transcripts identified by iReckon 5,880 PacBio transcripts that are in RefSeq**

Note: here everything is both “known” and has no multigene exonic regions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Match type** | **Total** | **NPC1** | **NPC2** | **NSC1 ∪ NSC2** | **NSC1 ∩ NSC2** | **Event Analysis (\*)** |
| Complete match | 5,880 | 274 | 227 | 401 | 100 | 5,686 / 4,538 |
| Match length is no more than 10bp diff | 5,880 | 291 | 243 | 426 | 108 | 5,686 / 4,538 |
| At least 90% match | 5,880 | 353 | 302 | 531 | 124 | 5,686 / 4,538 |
| At least 75% match | 5,880 | 368 | 310 | 549 | 129 | 5,686 / 4,538 |
| At least 50% match | 5,880 | 376 | 316 | 562 | 130 | 5,686 / 4,538 |

\* Number of PacBio transcripts retained with Event Analysis, at detection APN>0 and at least 75% events detected / 100% events detected

**Table S4.2: iReckon transcripts present in PacBio / Refseq:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **NPC1** | **NPC2** | **NSC1 ∪ NSC2** | **NSC1 ∩ NSC2** |
| No matches | 11,630 | 9,675 | 19,164 | 2,108 |
| PacBio-only | 1,278 | 825 | 1,948 | 160 |
| Refseq-only | 701 | 551 | 1,166 | 98 |
| PacBio and Refseq | 709 | 531 | 1,160 | 96 |
| Total | 14,318 | 11,582 | 23,438 | 2,462 |

**Table S4.3: iReckon compared to Event Analysis and PacBio, including 2 reduced references:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **NPC1** | **NPC2** | **NSC1 ∪ NSC2** | **NSC1 ∩ NSC2** |
| No matches | 11,630 | 9,675 | 19,164 | 2,108 |
| PacBio-only | 1,278 | 825 | 1,948 | 160 |
| Refseq match, not in RR1 or RR2 | 355 | 264 | 586 | 38 |
| Refseq match, in RR1 | 124 | 68 | 160 | 32 |
| Refseq match, in RR2 | 50 | 52 | 97 | 6 |
| Refseq match, in RR1 and RR2 | 172 | 167 | 323 | 22 |
| Refseq and PacBio match, not in RR1 or RR2 | 256 | 181 | 411 | 34 |
| Refseq and PacBio match, in RR1 | 36 | 14 | 48 | 2 |
| Refseq and PacBio match, in RR2 | 62 | 70 | 126 | 6 |
| Refseq and PacBio match, in RR1 and RR2 | 355 | 266 | 575 | 54 |

RR1= 100% events detected, APN>0

RR2= >=75% events detected, APN>=5

**Table S4.4. 23,438 distinct iReckon transcripts from the mouse NPC data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Transcript type** | **NPC1** | **NPC2** | **NSC1 ∪ NSC2** | **NSC1 ∩ NSC2** |
| Known transcript (NM/NR/XM/XR) | 837 | 664 | 1,361 | 140 |
| Known transcript with intron retention | 148 | 92 | 230 | 10 |
| Novel isoform | 8,080 | 6,235 | 14,131 | 184 |
| Novel isoform with intron retention | 1,543 | 1,063 | 2,583 | 23 |
| Novel unspliced isoform | 3,710 | 3,528 | 5,133 | 2,105 |
| Total | 14,318 | 11,582 | 23,438 | 2,462 |

23438 distinct iReckon transcripts From the mouse NPC data:

“Transcript type” is called on the basis of the transcript ID assigned by iReckon.

Transcripts with a RefSeq accession ID in their iReckon ID are the “known” transcript types

Transcripts with “IntronRetention” in their iReckon ID are the “intron retention” transcript tpyes

Transcripts with “Novel” in their iReckon ID are the “novel” transcript types

Transcripts with “unspliced” in their iReckon ID are the “unspliced” transcripts

**References**

1. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-842.

2. Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K *et al*: **SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification**. *Genome Research* 2018, **28**(3):396-411.