Assessment of transcript reconstruction methods for RNA-seq

Tamara Steijger[1], Josep F Abril[2,11], Pär G Engström[1,10,11], Felix Kokocinski[3,11], The RGASP Consortium[4], Tim J Hubbard[3], Roderic Guigó[5,6], Jennifer Harrow[3] & Paul Bertone[1,7–9]

We evaluated 25 protocol variants of 14 independent computational methods for exon identification, transcript reconstruction and expression-level quantification from RNA-seq data. Our results show that most algorithms are able to identify discrete transcript components with high success rates but that assembly of complete isoform structures poses a major challenge even when all constituent elements are identified. Expression-level estimates also varied widely across methods, even when based on similar transcript models. Consequently, the complexity of higher eukaryotic genomes imposes severe limitations on transcript recall and splice product discrimination that are likely to remain limiting factors for the analysis of current-generation RNA-seq data.

High-throughput sequencing instruments necessitate a shotgun approach for all but the shortest target molecules. Full-length representation of most cellular RNAs from sequencing data requires computational reconstruction of transcript structures. The majority of such programs infer transcript models from the accumulation of read alignments to the genome[1–4]; some take the alternative approach of *de novo* reconstruction, in which contiguous transcript sequences are assembled without the use of a reference genome[5–7].

Here we present a detailed evaluation of computational methods for transcript reconstruction and quantification from RNA-seq data, in a framework based on the Encyclopedia of DNA Elements (ENCODE) Genome Annotation Assessment Project (EGASP)[8]. Developers of leading software programs were invited to participate in a consortium effort, the RNA-seq Genome Annotation Assessment Project (RGASP), to benchmark methods to predict and quantify expressed transcripts from RNA-seq data. Results were evaluated from methods based on genome alignments (Augustus[9], Cufflinks[3], Exonerate[10], GSTRUCT, iReckon[2], mGene[11], mTim, NextGeneid[12], SLIDE[4], Transomics, Trembly and Tromer[13]) as well as *de novo* assembly (Oases[5] and Velvet[14]). Our results identify aspects of RNA-seq analysis in which current approaches are relatively adept, along with more challenging areas for future improvement.

RESULTS

We evaluated a total of 25 transcript reconstruction protocols, basing our analysis on alternate parameter usage of 14 software packages on RNA-seq data sets for three species (**Supplementary Fig. 1**, **Supplementary Table 1** and **Supplementary Note**). Programs were run by the original developers, with the exception of Cufflinks, iReckon and SLIDE. So that we could assess the ability of each method to interpret transcript expression from RNA-seq data without prior knowledge of gene content, programs were run without genome annotation, aside from iReckon and SLIDE, which require such information. Performance was benchmarked relative to the subset of annotated exons to which RNA-seq reads mapped (coverage of ≥1 read pair per 100 bp) and their corresponding transcripts (Online Methods).

Identification of annotated features

We first assessed the degree to which gene components reported by each algorithm matched the reference annotation at the nucleotide level. From the *Caenorhabditis elegans* data, the methods Augustus, mGene and Transomics displayed excellent performance in detecting exonic bases but also reported the expression of substantial proportions of genomic sequence outside of reference exons (**Fig. 1** and **Supplementary Table 2**). Recall (sensitivity) was generally lower for *Drosophila melanogaster*, although most protocols exceeded 75% for both model organisms. Performance decreased for *Homo sapiens* data, for which trade-offs between precision and recall were more apparent. SLIDE and iReckon must be provided with gene annotation and therefore outperformed most other methods. Even so, iReckon attained low precision at the nucleotide level, primarily owing to the prediction of transcript isoforms with retained introns. Augustus, Exonerate, GSTRUCT, NextGeneid, Trembly and Velvet attained both precision and recall above

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK. [2]Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain. [3]Wellcome Trust Sanger Institute, Cambridge, UK. [4]Full lists of members and affiliations appear at the end of the paper. [5]Centre for Genomic Regulation, Barcelona, Spain. [6]Universitat Pompeu Fabra, Barcelona, Spain. [7]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [8]Developmental Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [9]Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. [10]Present address: Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden. [11]These authors contributed equally to this work. Correspondence should be addressed to P.B. (bertone@ebi.ac.uk).

RECEIVED 31 MARCH; ACCEPTED 23 SEPTEMBER; PUBLISHED ONLINE 3 NOVEMBER 2013; DOI:10.1038/NMETH.2714

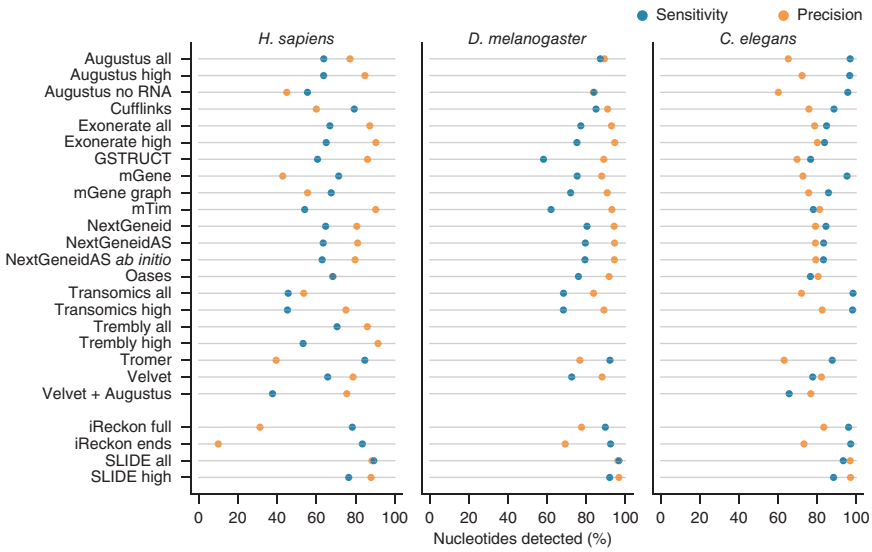


Figure 1 | Summary of nucleotide-level performance for the methods evaluated. The plots show performance at detecting exonic nucleotides. Sensitivity (blue) indicates the proportion of known exon sequence in each genome covered by assembled transcripts, and precision (orange) indicates the proportion of reported expressed sequence confined to known exons. Some protocol variants considered all expressed transcripts (all) or excluded those of low abundance (high). Programs run with gene annotation are grouped separately. iReckon was run with complete reference annotation (full) and with transcript boundaries only (ends). Transcript reconstruction methods are described in the **Supplementary Note**.

60% on the human data. The highest recall for methods without annotation was observed for Tromer and Cufflinks, albeit at the cost of low precision. These programs consistently displayed high sensitivity across the three species, but the precision rates for Tromer in particular indicate a tendency for overprediction.

Exon identification from RNA-seq data

We assessed the ability of each method to identify individual exons from RNA-seq data relative to the reference annotation (**Fig. 2**). Inaccurate determination of transcription start and end sites is a known shortcoming of RNA-seq and, together with biological variation, impairs the identification of transcript boundaries[15–19]. To mitigate this, we allowed the 5′ ends of first exons and 3′ ends of terminal exons to differ from the reference coordinates (Online Methods and **Supplementary Fig. 2**). Without these relaxed criteria, agreement in transcription start site and polyadenylation site positioning between predicted and annotated exons was extremely rare (**Supplementary Fig. 3**). Similarly, prediction accuracy for translation start and stop sites was lower than for internal exon boundaries, which can be inferred from spliced alignments (**Supplementary Fig. 3** and **Supplementary Table 3**). Allowing for variable transcript boundaries led to substantial improvements (**Supplementary Tables 4** and **5**). Although most protocols exhibited the lowest precision for the human RNA-seq data, for all three species, performance approached that of iReckon and SLIDE, despite the latter two benefiting from the use of high-quality gene annotation.

Coding exons can be identified directly from genomic sequence by the presence of translation start and stop sites and of splice acceptors and donors. Programs such as Augustus, Exonerate, mGene, NextGeneid, Tromer and Transomics use these features to improve exon discovery. Of these programs, Augustus, mGene and Transomics identified a greater proportion of annotated coding exons than did Exonerate, mTim, NextGeneid and Tromer (**Supplementary Fig. 4**). These methods augment data-driven transcript reconstruction with *ab initio* gene prediction, leading us to conclude that higher sensitivity measures are due to more extensive utilization of the underlying genomic sequence, which reduces the need for support from RNA-seq data.

We investigated the impact of sequencing depth on exon detection rates (**Fig. 3a**). Through the use of *ab initio* prediction, Augustus, mGene and Transomics were able to detect exons from protein-coding transcripts present at very low abundance. All other methods required a minimum average read depth to identify exons. Exon detection increased with sampling coverage at a roughly linear rate until reaching a plateau. One exception was Tromer, which often reported short exon fragments of 50–75 bp flanking introns without extending them to full exons (**Supplementary Fig. 5**). With increasing coverage, Tromer showed a tendency to predict very long exons spanning multiple annotated features. To a lesser extent, Oases and Velvet also showed

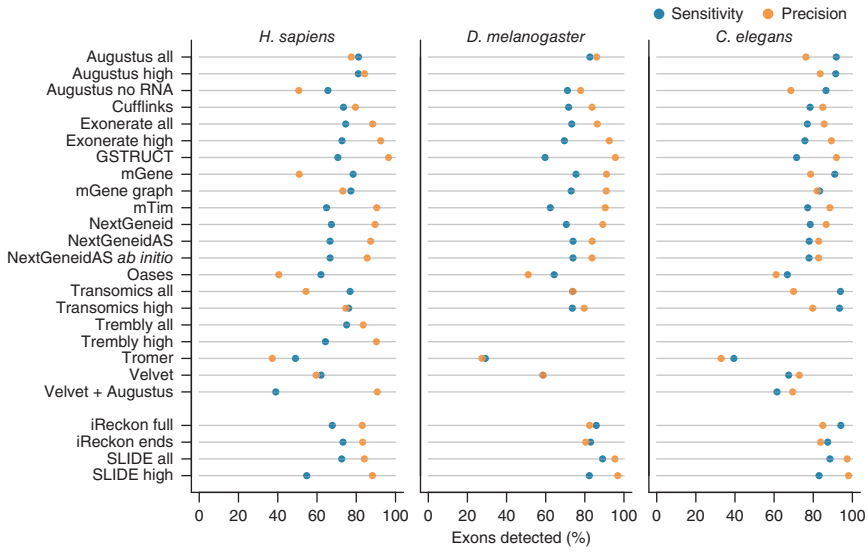


Figure 2 | Summary of exon-level performance for the methods evaluated. The plots show performance at detecting individual exons as the percentage of reference exons with a matching feature in the submission (sensitivity, blue) and the proportion of reported exons that agree with annotation (precision, orange).

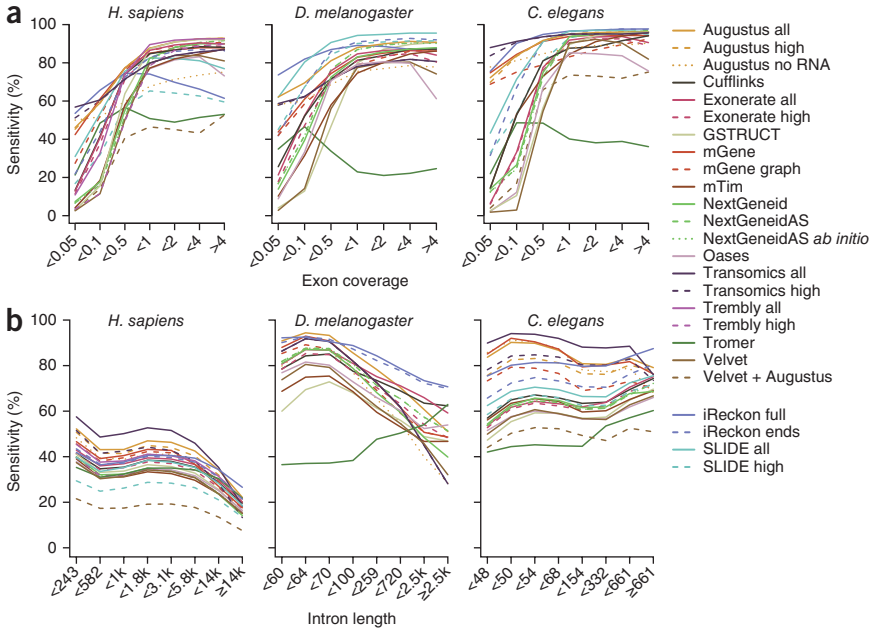


Figure 3 | Influence of read depth and intron length on detection performance. (**a**) Sensitivity for detection of annotated exons stratified by read depth. (**b**) Annotated introns were binned on length, and sensitivity was calculated separately for each bin.

reduced performance for high-coverage exons (**Supplementary Fig. 6**).

The *ab initio* prediction advantage of Augustus, mGene and Transomics was lost for noncoding transcripts, which lack the sequence features exploited by these methods (**Supplementary Fig. 7**). Nevertheless, detection rates were similar to those of other protocols (**Supplementary Fig. 8**). Noncoding RNAs tend to be expressed at lower levels than protein-coding genes (**Supplementary Fig. 9**) and were detected with lower sensitivity even when we controlled for differences in sequencing coverage (**Fig. 3a** and **Supplementary Fig. 7**). Exons of long intergenic noncoding RNAs were usually identified with lower frequency than those from pseudogenes and unclassified processed transcripts (**Supplementary Fig. 10**).

Intron detection from RNA-seq data

The relative number and size of introns differ markedly between the three species used for this study (**Supplementary Table 6**). Overall Augustus, mGene and Transomics showed the highest intron detection rates (**Fig. 3b**). However, Transomics exhibited a sharper decline with increased intron length. This trend was apparent for all methods except Tromer, for which a markedly lower detection rate was observed for introns shorter than 300 bp. To better characterize the differences in intron detection between methods, we classified reported introns on the basis of overlap with known splice sites (**Fig. 4**). Most protocols predominantly detected known introns; several, however, also predicted a substantial number of introns with one or two novel splice sites. The highest frequencies of novel junctions were predicted by mGene, Transomics, Tromer, Velvet and the Augustus protocol that used only genomic sequence.

To explain this trend, we note that intron detection is highly dependent on the underlying read alignments and that some aligners are more conservative than others[20]. For example, PALMapper[21] was used as the alignment component in the mGene and mTim protocols. This aligner places more reads across unannotated splice sites than do GEM[22], GSNAP[23] and TopHat[24,25]; the latter programs form part of the NextGeneid, GSTRUCT and Cufflinks protocols, respectively.

Assembly of exons into transcript isoforms

We next evaluated the performance of each method in linking exons into defined splice products. We initially determined the gene loci for which any expression was reported, regardless of whether a

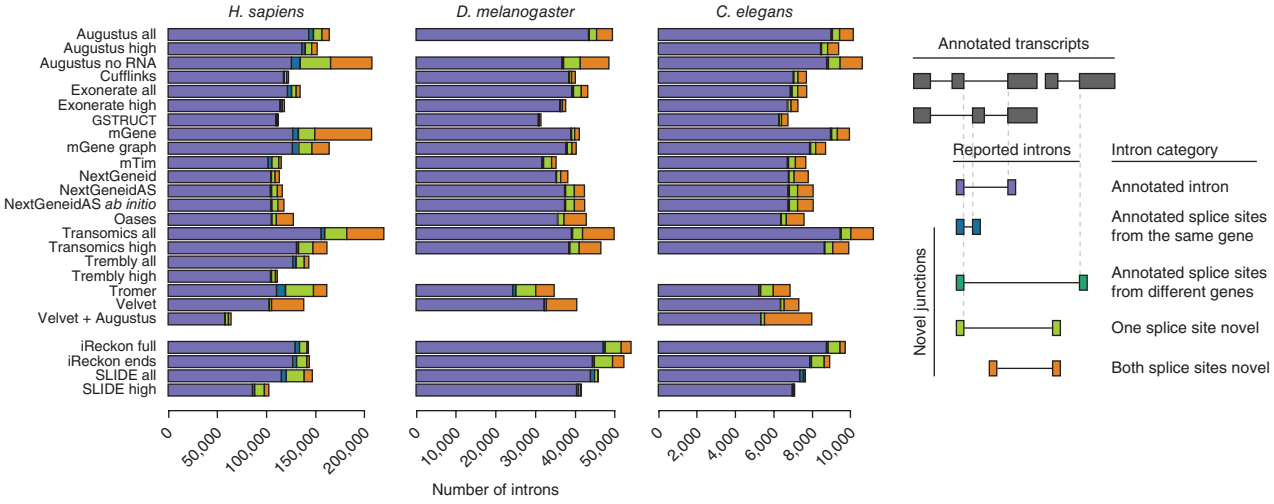


Figure 4 | Intron classification. Reported introns were classified by overlap with splice sites annotated in the reference gene sets.

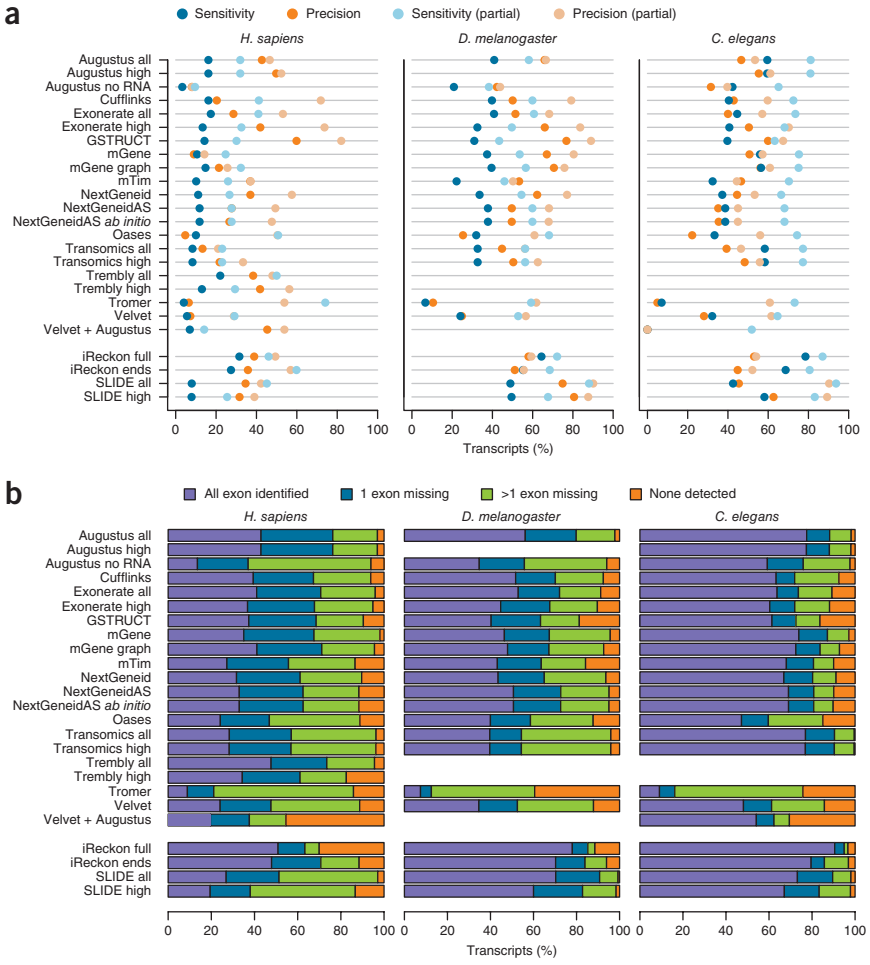


Figure 5 | Transcript assembly performance. (**a**) Reference transcripts with a matching submission entry (transcript-level sensitivity, blue) and reported transcripts that match the reference (transcript-level precision, orange). (**b**) Transcripts for which various subsets of constituent exons have been reported.

valid transcript was identified, followed by those consistent with at least one annotated isoform (**Supplementary Fig. 11**). Most algorithms detect transcription at over 80% of gene loci where expression is supported by RNA-seq reads. However, performance decreased substantially when genes were considered for which at least one annotated transcript had been identified. For unguided transcript reconstruction, valid isoforms were assembled for roughly half of expressed genes on average (*H. sapiens* mean 41%, maximum 61%; *D. melanogaster* mean 55%, maximum 73%; *C. elegans* mean 50%, maximum 73%), and for those only one isoform was typically identified (**Supplementary Fig. 12**).

A substantial reduction in sensitivity was also observed from the gene to transcript level, even when the flexible evaluation mode was used for first and terminal exons (**Fig. 5a** and **Supplementary Table 5**). The best-performing methods identified at most 56–59% of spliced protein-coding transcripts from *C. elegans* (Augustus, mGene and Transomics), 43% from *D. melanogaster* (Augustus) and merely 21% from *H. sapiens* (Trembly). Sensitivity increased by roughly 10% when partial isoform matches were considered, as did precision when partial predictions consistent with annotated isoforms were included (**Fig. 5a**).

Greater sequencing depth improved transcript assembly for *D. melanogaster* and *C. elegans* (**Supplementary Fig. 13a**), whereas in *H. sapiens* transcript detection remained low despite sequencing coverage in excess of 4,000 read pairs per kilobase in exonic regions. Generally, at least one consistent isoform was identified for highly expressed genes: >50% in *D. melanogaster* and *C. elegans* and >35% in *H. sapiens* (**Supplementary Fig. 13b**). Detection rates were even lower for noncoding RNAs (**Supplementary Fig. 14**). Pseudogenes were reported with similar frequency to that of protein-coding genes by Augustus, mGene, NextGeneid and Transomics, as pseudogenes retain partially intact coding sequences that can be identified by these methods (**Supplementary Fig. 15**).

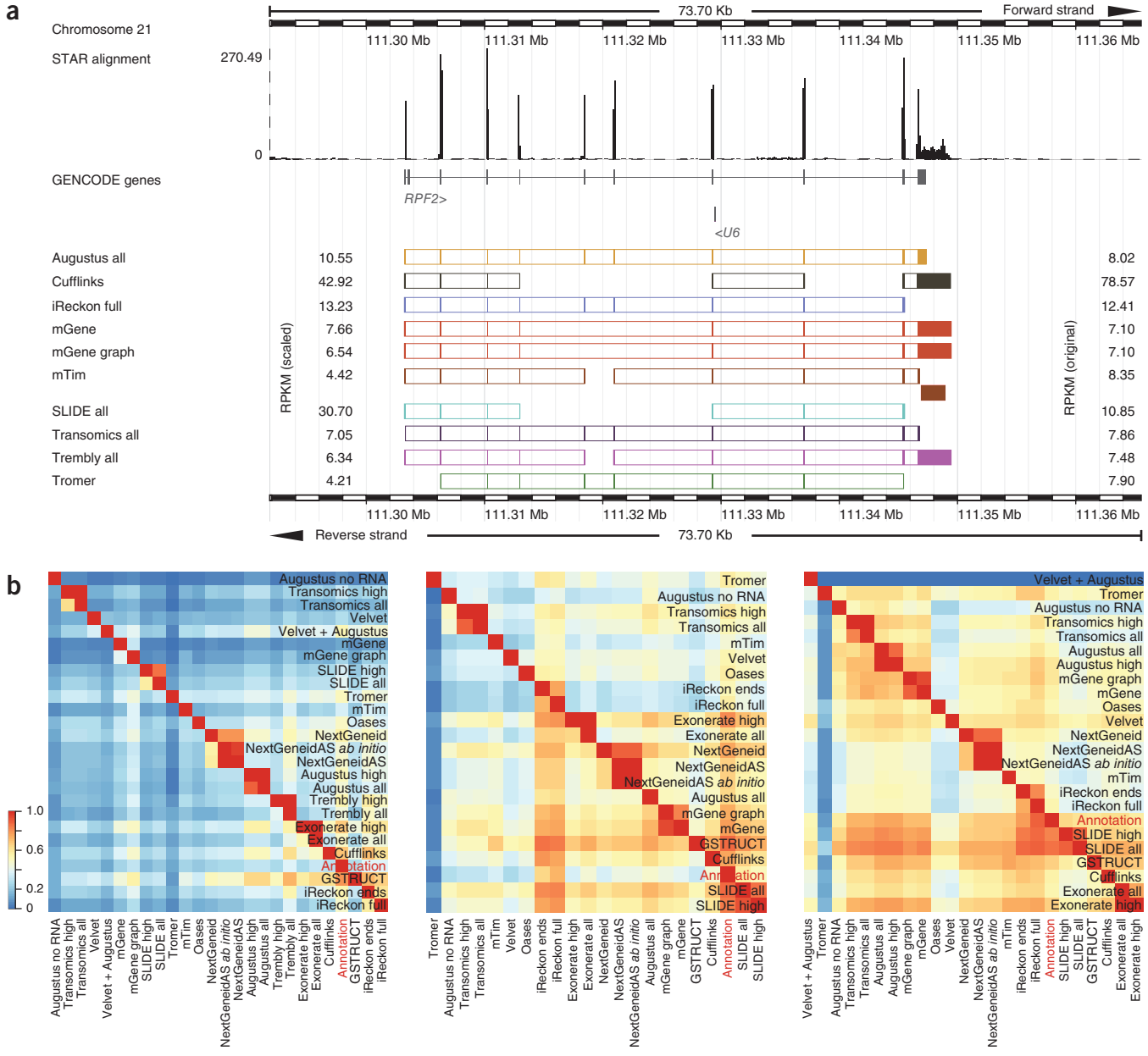
The dramatic differences between species is further due to the tendency of methods to assign one splice product per gene (**Supplementary Table 1**). Whereas fewer than 25% of genes in *C. elegans* and *D. melanogaster* give rise to more than two transcript isoforms, human genes are annotated with an average of five, and it is unclear how many are simultaneously expressed. Assigning a single transcript model per gene will therefore impede the detection of multiple isoforms expressed in a given sample.

To identify the limiting factors in this process, for each method we calculated the number of known transcripts for which (i) all exons were identified, (ii) exactly one exon was missing, (iii) more than one exon was missing and (iv) no exons were detected at all (**Fig. 5b**). The results clearly show that missing exons severely compromised transcript identification. For a substantial percentage of transcripts, not all exons were identified, ranging from 30% in *C. elegans* to greater than 60% in *H. sapiens*. Interestingly, although Trembly did not perform as well as Augustus, mGene and Transomics at the exon level, this method reported the highest number of transcripts for which all exons were represented from *H. sapiens* data. In contrast, Augustus, mGene and Transomics detected at least one exon for most transcripts. The remaining methods failed to identify any exons for nearly 20% of all transcripts expressed in the RNA-seq data. SLIDE exhibited the same trend despite the provision of annotated exon coordinates.

We then examined the topology of transcript structures to determine how well each method was able to link exons into complete isoforms. Even in cases in which all exons of an annotated transcript had been identified, full isoforms were often not assembled (**Supplementary Fig. 16**). For *C. elegans* and *D. melanogaster*, most methods were able to reconstruct 60% of transcripts from the RNA-seq data. However, from the *H. sapiens* data, less than 40% of known transcripts were assembled. Tromer stands out as an exception: the program identified all exons for relatively few genes; but once accounted for, these were frequently linked into annotated transcript structures. Further inspection showed that these tended

to be short isoforms comprising 2–3 exons on average and thus represent a more tractable subset of the transcriptome.

Provision of transcript start and end sites gave iReckon an advantage for the more complex human transcriptome, as evidenced by increased accuracy in assembling partial transcripts. In contrast, SLIDE consults exon coordinates but ignores their connectivity, performing at a level similar to methods without any prior transcript-level information. Reported transcript structures



c

	Augustus	Cufflinks	iReckon	mGene	mGene graph	mTim	SLIDE	Transomics	Trembly	Tromer
r_c	0.63 (62%)	0.68 (66%)	0.70 (79%)	0.59 (55%)	0.65 (61%)	0.53 (44%)	0.57 (61%)	0.34 (41%)	0.68 (65%)	0.61 (58%)
r_a	0.73 (87%)	0.74 (78%)	0.72 (86%)	0.74 (89%)	0.73 (82%)	0.68 (65%)	0.64 (89%)	0.74 (92%)	0.71 (74%)	0.74 (82%)

Figure 6 | Examples of transcript calls and expression-level estimates. (**a**) The upper tracks show RNA-seq read coverage (from STAR alignments; see Online Methods) and annotated genes. Exon predictions from the ten methods that quantified transcripts are illustrated below the annotated gene by colored boxes. Exons predicted to belong to the same transcript isoform are connected. Original and median-scaled RPKM values are presented to the right and left, respectively, of the transcript models. For the gene *RPF2*, all methods reported different isoforms and expression levels. Where multiple overlapping isoforms were identified, that with the higher RPKM was selected for visualization, and spliced isoforms were prioritized over unspliced ones. The noncoding RNA *U6* is not expressed. (**b**) Heat maps illustrate pairwise agreement between reported transcript isoforms for *H. sapiens* (left), *D. melanogaster* (center) and *C. elegans* (right). (**c**) Correlation between reported RPKM values and NanoString counts (Pearson *r* of log-transformed values). NanoString counts were compared to the highest RPKM value reported for transcript isoforms consistent with the probe design (correlation r_c) or for any isoform from the locus (correlation r_a).

often differed substantially (**Fig. 6a**), and few were consistent across all methods (**Supplementary Fig. 17**). Pairwise agreement (**Fig. 6b**) was markedly higher for the model organisms than for human (median 25%), reflecting the number of partial isoforms identified as a function of transcriptome complexity.

Quantification of expression levels from RNA-seq data

A common feature of transcript reconstruction software is the estimation of expression levels from transcribed genes. These are given as digital read counts normalized by transcript length and sequencing depth (reads per kilobase of exon model per million mapped reads, RPKM)[26]. RPKM values were reported at the transcript level from a subset of methods. A range of expression-level distributions was evident (**Supplementary Fig. 18**), but generally there was strong agreement among Augustus, iReckon, mGene and Trembly for all three RNA-seq data sets (**Supplementary Figs. 19–22**). One source of variation arises from gene loci where divergent or incomplete transcript models have been computed (**Fig. 5a** and **Supplementary Figs. 23** and **24**). However, expression-level estimates can vary considerably even where concordant transcript structures are reported (**Supplementary Fig. 17**). Such differences were also apparent after we scaled the RPKM distributions to equalize median expression values.

To establish independent expression-level quantification, we assayed a set of human genes using the NanoString nCounter amplification-free detection system[27] (**Supplementary Tables 7–9**). Correlation between NanoString counts and RNA-seq RPKMs ranged from 0.34 for Transomics to 0.68 for Cufflinks (**Fig. 6c** and **Supplementary Fig. 25**). Many methods failed to report numerous targeted exons or junctions that were expressed according to NanoString counts. Read support at those loci was typically sparse, with 19 probes having no corresponding alignments from the RNA-seq data. These were, however, represented by low NanoString counts, which indicated that the nCounter assay exhibits higher sensitivity for low-abundance transcripts than RNA-seq (**Supplementary Fig. 26**). For ten of the unsupported NanoString probes, consistent isoforms were still reported by either Augustus, iReckon, mGene, SLIDE or Transomics. Thus, although the expression levels of these genes reflect the lower limits of detection for both technologies, sequencing reads dispersed over the gene body can allow for adequate transcript identification where *ab initio* methods or gene annotation were applied.

In general, all methods displayed higher identification rates for exons and junctions with higher NanoString counts, and reliable detection from RNA-seq data was dependent on read depth (**Supplementary Fig. 27**). Nonetheless, each failed to report a subset of exons and junctions despite the availability of adequate RNA-seq alignments (**Supplementary Fig. 27b**). Comparing NanoString counts with RPKM values of the predominant isoform reported for each gene (irrespective of whether the targeted exon or junction was identified) improved correlation for most methods and did so substantially for mTim and Transomics (**Fig. 6c** and **Supplementary Fig. 28**).

DISCUSSION

Technical limitations imposed by short-read sequencing lead to a number of computational challenges in transcript reconstruction and quantification. Methods that combine *ab initio* prediction with experimental data were more effective at detecting genes expressed at low abundance or genes from samples with low sequencing coverage. Even so, the benefits of this approach lessened with increased transcriptome complexity.

These results underscore the difficulty of transcript assembly. For most transcripts, automated methods failed to identify all constituent exons, and in cases in which all exons were reported, the protocols tested often failed to assemble the exons into complete isoforms. Whereas methods using *ab initio* prediction retained an advantage in detecting individual exons, others performed better at linking them together. No single protocol excelled at all metrics. Comparing the performance of Augustus with and without RNA-seq data as input revealed that using experimental evidence only slightly improved exon-level detection but increased transcript-level precision. Transomics featured enhanced precision for high-abundance transcripts, but expression-level differences had little impact on detection sensitivity. Precision was a consistent strength of GSTRUCT, whereas mGene exhibited diminished performance on human RNA-seq data, a result underscoring that choice of method can depend on the organism under study.

Considerable variation was observed in the range of expression-level estimates reported for transcripts arising from the same gene loci. This was exacerbated by nonuniform exon detection and linkage between methods but was also apparent when similar or identical transcript structures were reported. Thus, it may be unreliable to directly compare gene-based RPKM values from sample data processed independently with different software tools. RNA-seq data to be compared from disparate sources should be treated in an identical manner from the initial processing steps. When this is not possible, care should be taken to ensure that similar gene models have been identified, and RPKM distributions should be inspected before expression-level thresholds are applied in downstream analyses. Alternatively, uniform quantification of predicted transcripts can be performed with dedicated software[28–30].

The potential for noncoding RNA discovery and characterization is a distinct advantage of RNA-seq over gene-based expression profiling. However, this remains a challenging area for automated analysis methods. Performance is often impaired by lower expression levels of noncoding transcripts relative to that of many protein-coding genes, coupled with the inherent lack of translational features at the sequence level. The presence of open reading frames and translation start and stop signals allowed some methods to identify protein-coding transcripts even at very low expression levels, whereas the detection of noncoding RNAs at high confidence required much greater read depth. Sequencing coverage thus appears to be crucial for accurate noncoding RNA analysis.

The methods evaluated here can be applied to a range of analysis strategies, largely dependent on the state of the reference genome assembly and associated gene annotation for the target species (**Supplementary Table 10** and **Supplementary Note**). To improve the accuracy of existing annotation using RNA-seq data, both Cufflinks and iReckon consult known gene structures during the transcript assembly process and may be useful in refining the coordinates of exon and transcript boundaries. Where a finished genome and high-quality annotation are available, Cufflinks and rQuant (part of the mGene protocol) can be applied solely for transcript quantification, which can be further improved by correcting for fragment bias. Gene prediction algorithms such

as Augustus and mGene can be used to automate the annotation of novel genomes, whereas RNA-seq experiments based on partial or low-quality genome builds can be approached with a *de novo* assembler such as Oases. This last application is expected to receive increasingly wider attention with the continued sequencing of new genomes.

RNA-seq offers the potential to refine existing gene annotation through the discovery of novel exons and junction sites. However, unannotated transcript isoforms assembled from RNA-seq data should be interpreted with care, and those critical to an experimental study should be subjected to independent validation. The expression of multiple transcript isoforms and novel splice variants presents a major obstacle to accurate transcriptome reconstruction. Both exon identification and novel RNA discovery can improve with increased read depth, but the benefits of additional sampling to transcript assembly are inherently limited by the library construction requirements of current high-throughput sequencing platforms. Ultimately, the evolution of RNA-seq will move toward single-pass determination of intact transcripts. Third-generation instruments will realize that potential and inspire new computing approaches to meet the next wave of innovation in transcriptome analysis.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by European Molecular Biology Laboratory, US National Institutes of Health/NHGRI grants U54HG004555 and U54HG004557, Wellcome Trust grant WT098051, and grants BIO2011-26205 and CSD2007-00050 from the Ministerio de Educación y Ciencia.

AUTHOR CONTRIBUTIONS

J.H., R.G. and T.J.H. conceived of and organized the study. Consortium members provided transcript models for evaluation. J.H. and P.B. coordinated the analysis, which was carried out by T.S., J.F.A., P.G.E. and F.K. T.S., P.B. and P.G.E. wrote the manuscript with input from the other authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
2. Mezlini, A.M. *et al.* iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res.* **23**, 519–529 (2013).
3. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**, 2325–2329 (2011).
4. Li, J.J., Jiang, C.-R., Brown, J.B., Huang, H. & Bickel, P.J. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc. Natl. Acad. Sci. USA* **108**, 19867–19872 (2011).
5. Schulz, M.H., Zerbino, D.R., Vingron, M. & Birney, E. Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086–1092 (2012).
6. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
7. Robertson, G. *et al.* *De novo* assembly and analysis of RNA-seq data. *Nat. Methods* **7**, 909–912 (2010).
8. Guigó, R. *et al.* EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.* **7** (suppl. 1), S2 (2006).
9. Stanke, M. *et al.* AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
10. Slater, G.S.C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
11. Schweikert, G. *et al.* mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Res.* **19**, 2133–2143 (2009).
12. Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Curr. Protoc. Bioinformatics* **18**, 4.3 (2007).
13. Sperisen, P. *et al.* trome, trEST and trGEN: databases of predicted protein sequences. *Nucleic Acids Res.* **32**, D509–D511 (2004).
14. Zerbino, D.R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
15. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* **13**, 233–245 (2012).
16. Di Giammartino, D.C., Nishida, K. & Manley, J.L. Mechanisms and consequences of alternative polyadenylation. *Mol. Cell* **43**, 853–866 (2011).
17. Tian, B., Hu, J., Zhang, H. & Lutz, C.S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33**, 201–212 (2005).
18. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T.R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* **23**, 169–180 (2013).
19. Shepard, P.J. *et al.* Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**, 761–772 (2011).
20. Engström, P.G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* doi:10.1038/nmeth.2722 (3 November 2013).
21. Jean, G., Kahles, A., Sreedharan, V.T., De Bona, F. & Rätsch, G. RNA-Seq read alignments with PALMapper. *Curr. Protoc. Bioinformatics* **32**, 11.6 (2010).
22. Marco-Sola, S., Sammeth, M., Guigo, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).
23. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
24. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
25. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
26. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
27. Kulkarni, M.M. Digital multiplexed gene expression analysis using the NanoString nCounter system. *Curr. Protoc. Mol. Biol.* **94**, 25B.10 (2011).
28. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
29. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
30. Bohnert, R. & Rätsch, G. rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res.* **38**, W348–W351 (2010).

RNA-seq Genome Annotation Assessment Project (RGASP) Consortium

Josep F Abril[2,11], Martin Akerman[12], Tyler Alioto[13], Giovanna Ambrosini[14,15], Stylianos E Antonarakis[16], Jonas Behr[17,18], Paul Bertone[1,7–9], Regina Bohnert[18], Philipp Bucher[14,15], Nicole Cloonan[19], Thomas Derrien[5], Sarah Djebali[6], Jiang Du[20], Sandrine Dudoit[21], Pär G Engström[1,10,11], Mark Gerstein[20,22,23], Thomas R Gingeras[12], David Gonzalez[5], Sean M Grimmond[19], Roderic Guigó[5,6], Lukas Habegger[23], Jennifer Harrow[3], Tim J Hubbard[3], Christian Iseli[15,24], Géraldine Jean[18], André Kahles[17,18], Felix Kokocinski[3,11], Julien Lagarde[5], Jing Leng[23], Gregory Lefebvre[14,15], Suzanna Lewis[25], Ali Mortazavi[26], Peter Niermann[18], Gunnar Rätsch[17,18], Alexandre Reymond[27], Paolo Ribeca[13], Hugues Richard[28], Jacques Rougemont[14,15], Joel Rozowsky[22], Michael Sammeth[5], Andrea Sboner[22], Marcel H Schulz[28], Steven M J Searle[3], Naryttza Diaz Solorzano[15,24], Victor Solovyev[29], Mario Stanke[30], Tamara Steijger[1], Brian J Stevenson[15,24], Heinz Stockinger[15], Armand Valsesia[15,24], David Weese[31], Simon White[3], Barbara J Wold[32], Jie Wu[12,33], Thomas D Wu[34], Georg Zeller[18], Daniel Zerbino[1] & Michael Q Zhang[12]

[12]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [13]Centre Nacional d'Analisi Genomica, Barcelona, Spain. [14]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [15]Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland. [16]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [17]Computational Biology Center, Sloan-Kettering Institute, New York, New York, USA. [18]Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany. [19]Queensland Centre for Medical Genomics, The University of Queensland, St. Lucia, Australia. [20]Department of Computer Science, Yale University, New Haven, Connecticut, USA. [21]Division of Biostatistics, School of Public Health, University of California, Berkeley, Berkeley, California, USA. [22]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, USA. [23]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA. [24]Ludwig Institute for Cancer Research, Lausanne, Switzerland. [25]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. [26]Department of Developmental and Cell Biology, University of California, Irvine, Irvine, California, USA. [27]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. [28]Max Planck Institute for Molecular Genetics, Berlin, Germany. [29]Department of Computer Science, Royal Holloway, University of London, London, UK. [30]Institute for Microbiology and Genetics, Göttingen, Germany. [31]Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany. [32]Biology Division, California Institute of Technology, Pasadena, California, USA. [33]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York, USA. [34]Department of Bioinformatics and Computational Biology, Genentech, San Francisco, California, USA.

ONLINE METHODS

RNA-seq data. RNA-seq data were generated as part of the ENCODE[31] and modENCODE projects[32], along with a third data set of compatible sequencing format and read depth, and represent three widely studied species: *H. sapiens* (liver hepatocellular carcinoma cell line HepG2)[33], *D. melanogaster* (L3 stage larvae)[34], and *C. elegans* (L3 stage larvae)[35]. These were chosen to reflect realistic examples of varying transcriptome complexity and where high-quality annotated reference genomes are available. Libraries were prepared for the Illumina platform and sequenced in 76-nt paired-end format to obtain approximately 100 million read pairs per sample.

H. sapiens RNA-seq data correspond to ENCODE[33] HepG2 whole-cell long poly(A)$^+$ RNA CALTECH replicate 2, available from http://www.encodeproject.org/. The *D. melanogaster* data set comprised a total of five sequencing runs from the modENCODE project[34] for three L3 stage larval samples and can be obtained from the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession numbers SRR023546, SRR023608, SRR023505, SRR027108 and SRR026433. The *C. elegans* RNA-seq data have previously been described[35] and are available under accession SRR065719. All of the data used in this study have been consolidated as a single experimental record in the ArrayExpress repository (http://www.ebi.ac.uk/arrayexpress/) under accession E-MTAB-1730.

Reference gene annotation. As not all genes are expressed in the samples used in the study, benchmarking methods against the entire set of annotated genes would underestimate transcript detection sensitivity. Therefore, we processed the genome annotations (*H. sapiens*: GENCODE[31] v.15 (Ensembl release 70), *D. melanogaster*: FB2013_01, *C. elegans*: WS200) to include only exons and transcripts with sufficient support in the RNA-seq data. Reads were mapped to the reference genomes using STAR version 2.2.0c, an independent RNA-seq aligner that is not a component in any of the evaluated transcript assembly methods[36]. To improve spliced alignment, STAR was provided with exon junction coordinates from the reference annotations. Default alignment parameters were used for the human data. For *D. melanogaster* and *C. elegans*, the intron size limit was reduced to 100,000 and 15,000 respectively (using options -alignIntronMax and -alignMatesGapMax). For each annotated exon, the read coverage (number of uniquely mapped read pairs divided by exon length) was computed, and exons with a value below 0.01 fragments per base pair were excluded from further analysis. Only transcripts for which all exons satisfied this criterion were included in transcript-level assessments. The threshold was determined by examining the exonic read coverage distribution, which consisted of three main features: a small peak at the low end (coverage < 0.01 fragments per base pair), a dominant peak (coverage > 0.1) and a shoulder in between. Inspection of read alignments suggested that spurious reads are overrepresented in the minor peak, whereas the shoulder region comprises low-abundance transcripts and was therefore included in the analysis. To rule out potential bias imparted by the choice of alignment program, we calculated sensitivity and precision metrics for expressed genes using several different spliced aligners (GSNAP, STAR and TopHat2), with no substantial change to the results (**Supplementary Figs. 29–31**).

Transcript prediction and assembly. Developer teams were provided with RNA-seq data and reference genome sequences for each species. So that we avoided potential biases, teams were not informed of the final evaluation criteria and were not provided with gene annotation unless otherwise noted (for example, iReckon and SLIDE). Developers providing transcript models for evaluation could not access submissions from other teams and were prohibited from participating in the analysis phase as part of the study design. Details of transcript reconstruction protocols are provided in the **Supplementary Note**.

Data processing for Cufflinks, iReckon and SLIDE. RNA-seq reads were aligned with TopHat version 2.0.3 using parameters suited to each species. The genomes of *D. melanogaster* and *C. elegans* contain a high percentage of small introns (**Supplementary Table 2**); examining their size distributions led us to set the parameters -i, -min-coverage-intron and -min-segment-intron to 30 for *C. elegans*, 40 for *D. melanogaster* and 50 for *H. sapiens*.

Cufflinks was run with default settings except for the parameter -min-intron-length, which was set to 30 for *C. elegans*, 40 for *D. melanogaster* and 50 for *H. sapiens*, consistent with the TopHat alignments. So that we maintained the greatest compatibility with submitted results that were computed without annotation. The protocol iReckon ends was run with the minimum annotation requirements, i.e., start and end sites of all annotated transcripts (not filtered by read coverage), whereas iReckon full was provided with the complete reference annotation. SLIDE was run in discovery mode and provided with the full unfiltered annotation for each genome.

Evaluation of prediction sets. Feature predictions were evaluated against the filtered reference annotation sets at four structural levels: nucleotide, exon, transcript and gene. The nucleotide-level metrics measure the ability of methods to identify exonic regions, ignoring the strand and exact boundaries of features. Nucleotide-level precision was computed as the number of genomic base pairs within both annotated and predicted exons, divided by the number of genomic base pairs within predicted exons. Similarly, nucleotide recall was computed as the number of genomic base pairs shared between annotated and predicted exons, divided by the number of genomic base pairs within annotated exons.

The exon-level metrics measure the ability of the different algorithms to identify the correct strand and boundaries of exons. Precision was calculated as the percentage of reported exons with an annotated counterpart, and recall denotes the percentage of annotated exons that were correctly assembled. Annotated exons were classified as first, internal, terminal and those comprising unspliced transcripts (single exons). Unless stated otherwise, a flexible evaluation mode was employed for first, terminal and single exons. Specifically, first and terminal exons were required to have correctly predicted internal borders only, and exons constituting unspliced transcripts were scored as correct if covered to at least 60% by a predicted transcript. Exons shared between different transcript isoforms were counted once. For comparison, certain analyses were also carried out using a fixed evaluation mode, where annotated and predicted exons were required to match exactly.

Transcript-level precision was computed as the percentage of reported spliced transcripts matching an annotated transcript,

and recall as the percentage of annotated spliced transcripts with a counterpart in the transcript reconstruction output. Consistent with the flexible evaluation mode for exons (see above), transcript start and end sites were allowed to differ between reference and prediction, but splice sites were required to match exactly. Genes were scored as correctly predicted if at least one annotated transcript isoform in a given gene locus was correct. To estimate the degree of similarity between transcript predictions, we calculated a pairwise agreement score. The score $a[i,j]$ denotes the fraction of transcription products predicted by protocol i consistent with those from protocol j. Methods were ordered by hierarchical clustering based on the distance metric $1 - (a[i,j] + a[j,i])/2$.

Evaluation of transcript quantification. To compare transcript quantification results between methods, we identified for each annotated gene the corresponding predominant transcript reported; this was defined as the transcript with the highest reported RPKM value among those isoforms intersecting annotated exons of the gene. A subset of human transcripts was quantified independently by NanoString assays. Genes of at least 1 kb in length, for which annotated exon-intron structures have been manually curated, and having at least two transcripts satisfying these criteria were selected. A total of 109 genes were targeted by 141 distinct probes, designed against specific exons or splice junctions.

NanoString counts were compared to the highest RPKM value reported for transcript isoforms consistent with the probe design (correlation r_c) or for any isoform from the locus (correlation r_a). Predicted transcripts were required to contain the exon or junction targeted by the NanoString probe. Where multiple such transcripts were reported for the same gene, the highest RPKM value was used. Where no such transcript was reported, an RPKM of 0 was assigned. Percentages reflect the probes for which transcripts satisfying these criteria were reported. Pearson's r was calculated on the basis of the log-transformed NanoString counts and RNA-seq RPKM values. Expression values were incremented by 1 before transformation to avoid infinite numbers.

Software availability. Source code for the evaluations performed in this study can be obtained from https://github.com/RGASP-consortium/.

31. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
32. The modENCODE Consortium *et al.* Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
33. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
34. Graveley, B.R. *et al.* The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471**, 473–479 (2011).
35. Mortazavi, A. *et al.* Scaffolding a *Caenorhabditis* nematode genome with RNA-seq. *Genome Res.* **20**, 1740–1747 (2010).
36. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

doi:10.1038/nmeth.2714

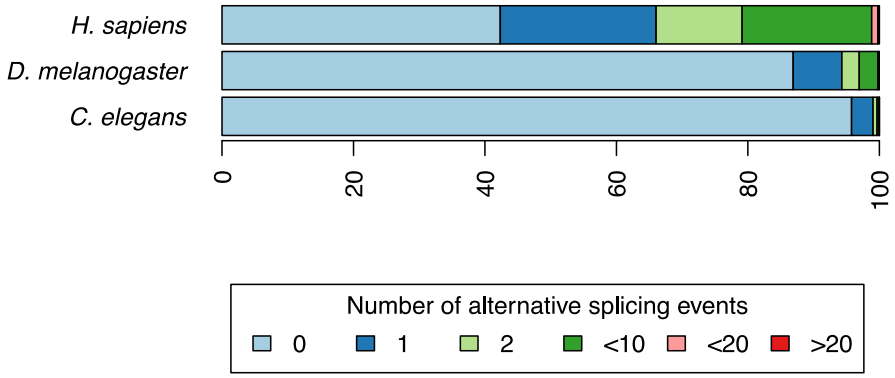
Assessment of transcript reconstruction methods for RNA-seq

Tamara Steijger, Josep F. Abril, Pär G. Engström, Felix Kokocinski, RGASP Consortium,
Tim J. Hubbard, Roderic Guigó, Jennifer Harrow and Paul Bertone

Supplement

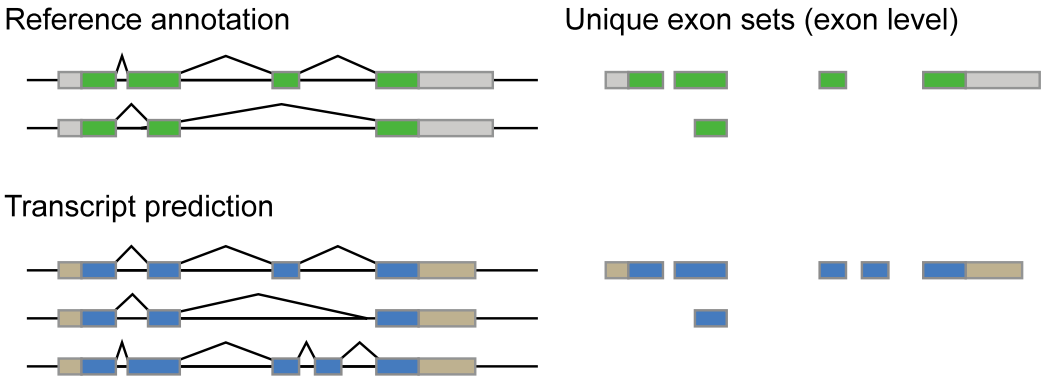
Contents

Nature Methods: doi:10.1038/nmeth.2714



Supplementary Figure 1. Frequency of alternative splicing events. Bars show the percentage of genes with the indicated number of alternative splicing events in the reference annotation. Events were counted by analysis of annotated transcripts to identify skipped exons, retained introns, and alternative donor and acceptor sites.

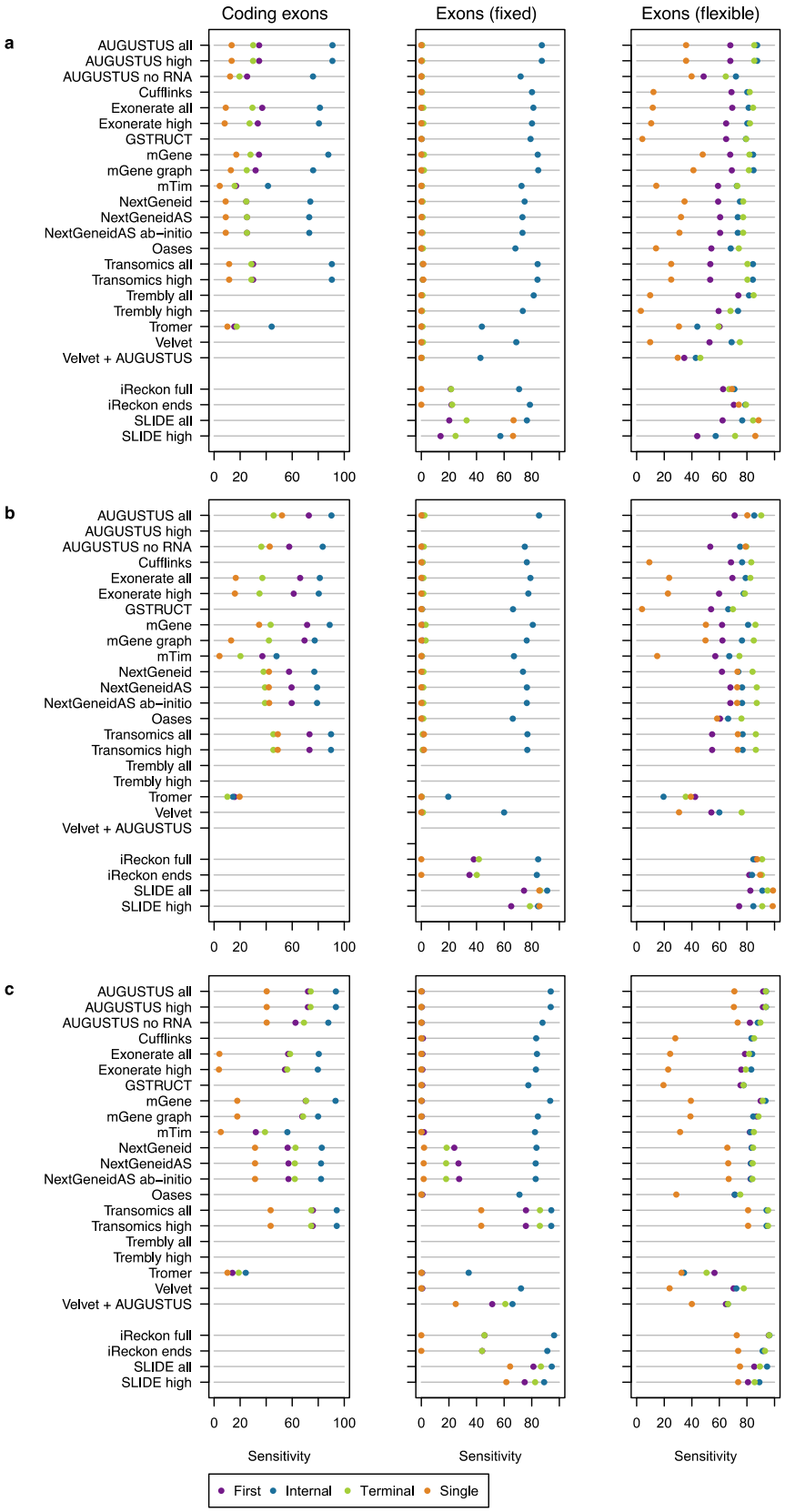
Nature Methods: doi:10.1038/nmeth.2714



Level		fixed mode		flexible mode	
	Feature	Sensitivity	Precision	Sensitivity	Precision
Exon	CDS	1.00	0.83	NA	NA
	Exon	0.80	0.67	1.00	0.83
Transcript	CDS	0.50	0.33	NA	NA
	Exon	0.00	0.00	0.50	0.33
Gene	CDS	1.00	1.00	NA	NA
	Exon	0.00	0.00	1.00	1.00

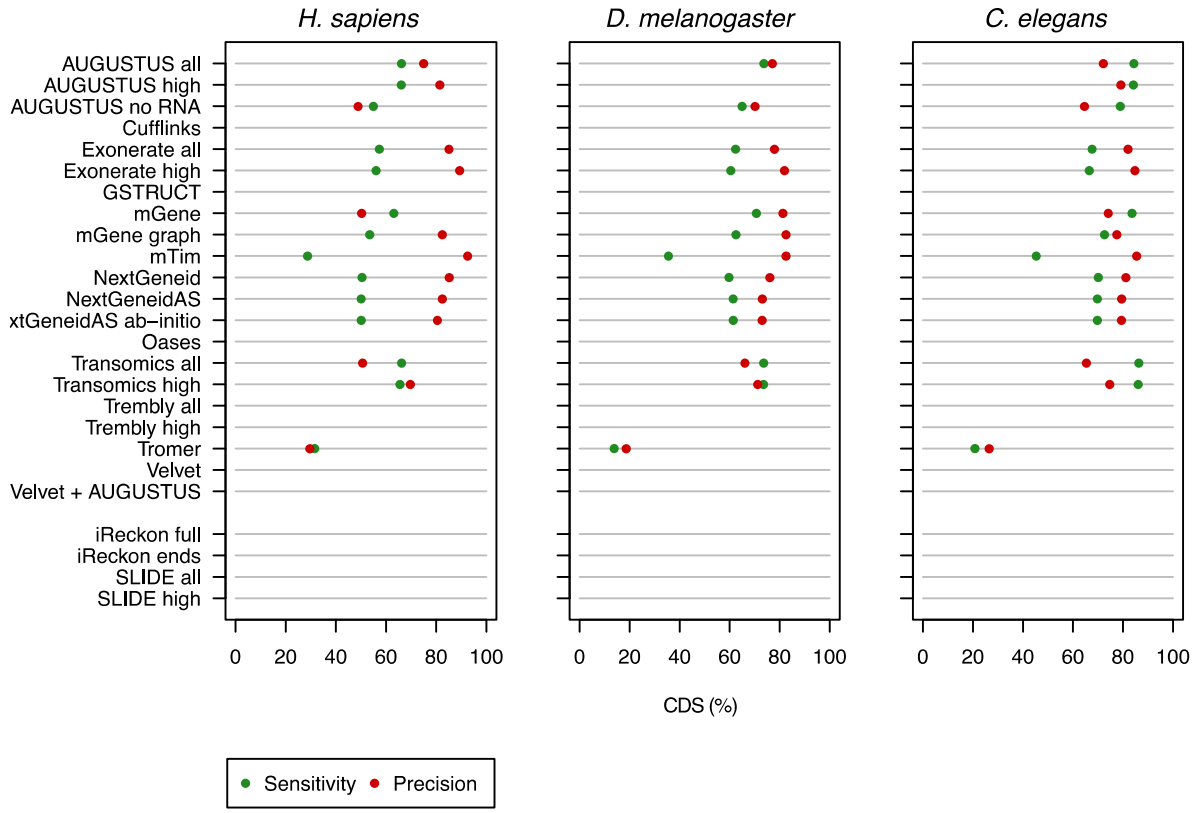
Supplementary Figure 2. Structural validation strategy. Transcript models were validated against annotated isoforms. For exon-level evaluation, transcripts were collapsed into unique exon sets, i.e. exons shared between transcript isoforms are counted once. Sensitivity (a.k.a. recall) was calculated as the proportion of reference features (exons, transcripts, or genes) matched by a reported feature. Precision was calculated as the proportion of reported features matching a reference feature. We primarily used a flexible evaluation strategy where exact agreement between transcript boundaries was not required. For comparison, certain analyses were also carried out using a fixed evaluation mode, where annotated and predicted exons were required to match exactly. See Methods for further details.

Nature Methods: doi:10.1038/nmeth.2714

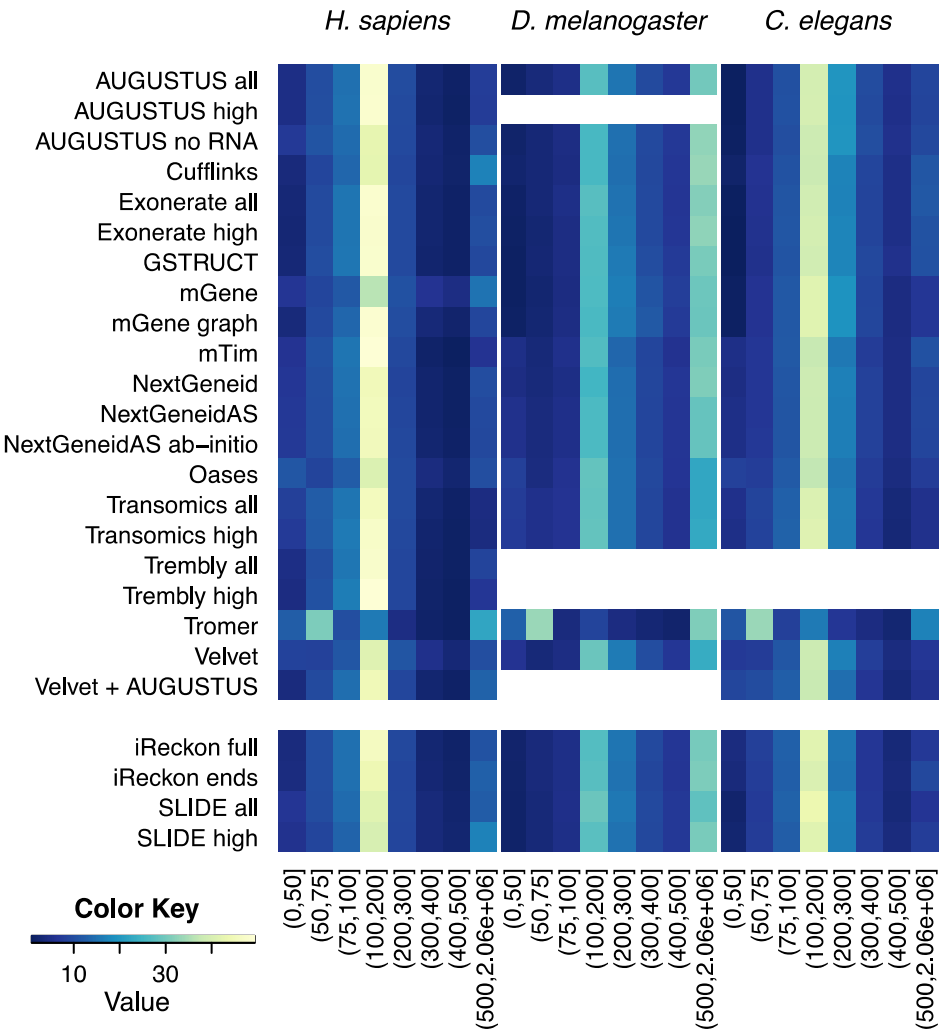


Supplementary Figure 3. Influence of exon rank on detection performance. Results are shown for *D. melanogaster* (a) and *C. elegans* (b). Annotated exons were classified as first, internal, terminal or single (i.e., those comprising an entire transcript) and sensitivity calculated separately for each class. Exon boundaries were required to be predicted exactly as annotated (left, center) or according to relaxed criteria for the external transcript boundaries (right). Programs run with reference annotation are grouped separately (lower tracks). As SLIDE is provided with full gene annotation as a requirement, those protocols do not display a strong preference for internal exons. Several methods were unable to accurately determine the strand orientation for unspliced transcripts, resulting in low sensitivity for constituent exons.

Nature Methods: doi:10.1038/nmeth.2714

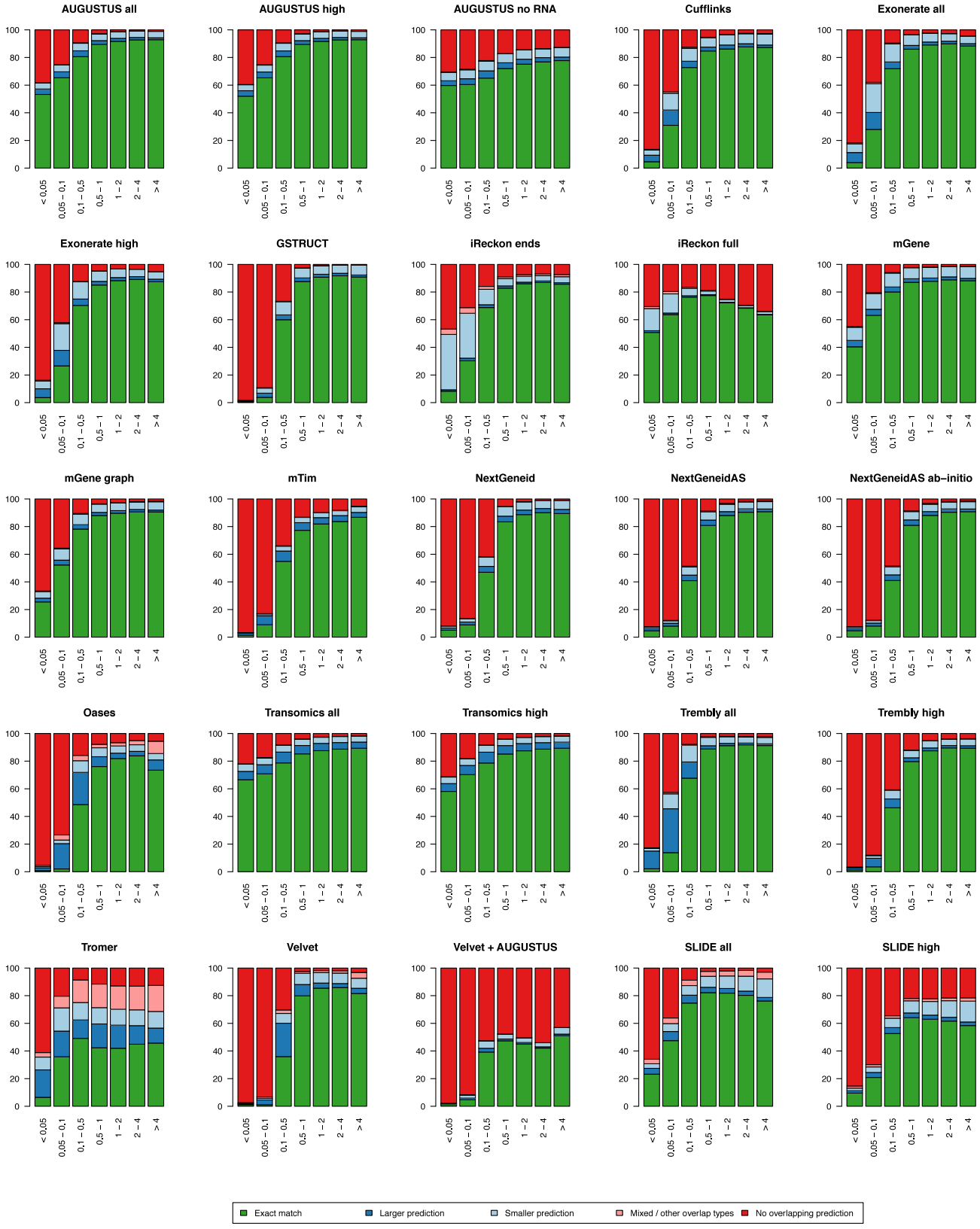


Supplementary Figure 4. Performance at detecting individual coding exons. Points indicate the percentage of reference coding exons with a matching feature in the submitted transcript models (recall, green), and the proportion of reported coding exons that agree with annotation (precision, red).



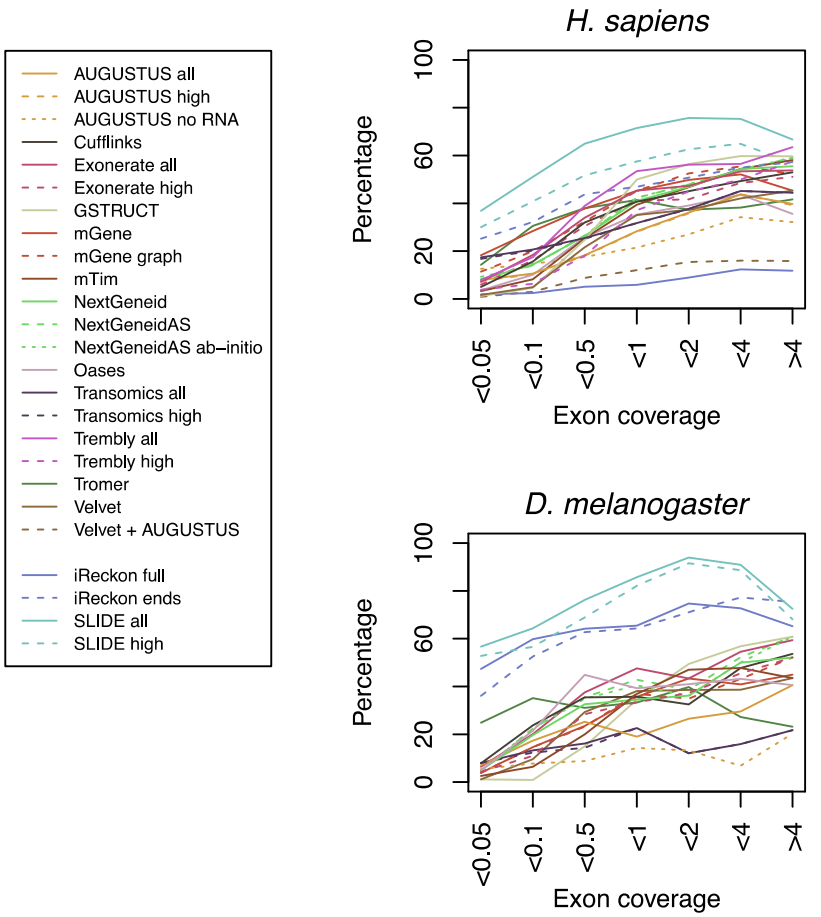
Supplementary Figure 5. Exon length distributions in transcriptome assembly results. Colors indicate percentage of exons within the indicated length intervals.

Nature Methods: doi:10.1038/nmeth.2714



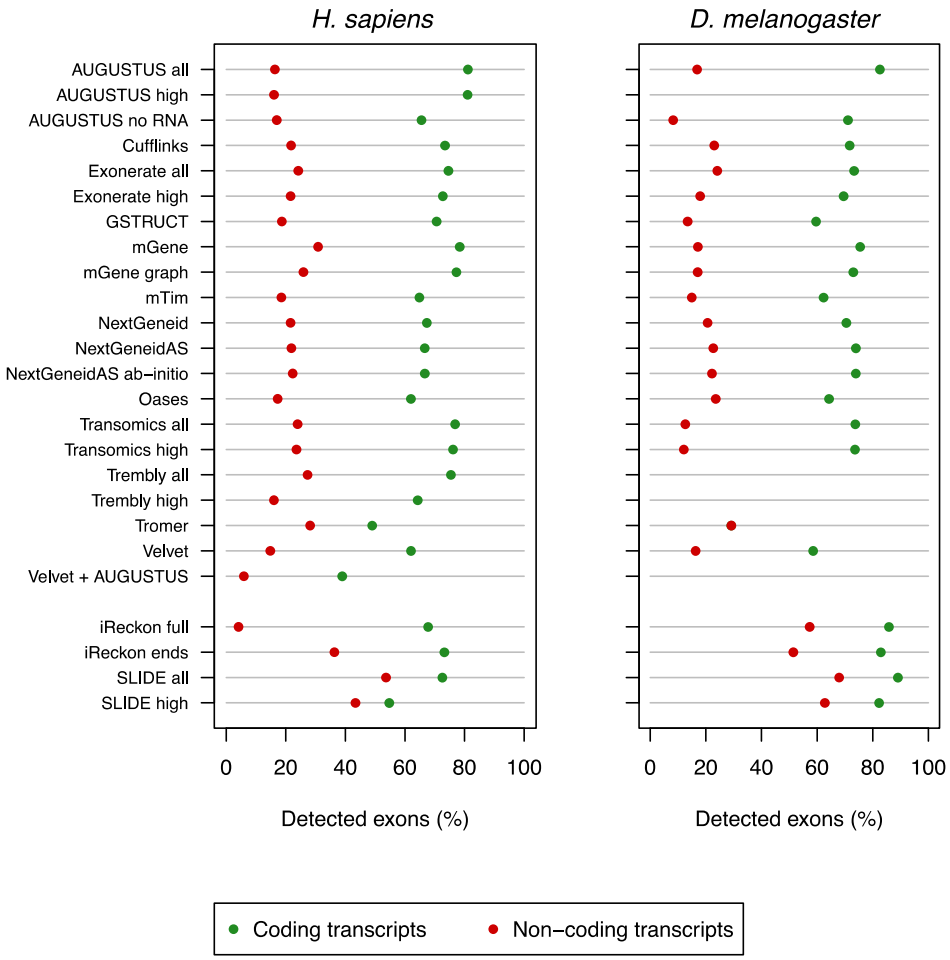
Supplementary Figure 6. Internal exon detection rate stratified by read coverage. Bars indicate the percentage of annotated internal exons (of human protein-coding genes) that overlap with reported exons. Reference exons were binned by read coverage (*x* axis) and further classified based on overlap with predicted exons (inset legend). Specifically, the classes represent exons with a perfectly matching prediction (green); exons for which all overlapping predictions span a larger region, including the entire reference exon (dark blue); exons for which all overlapping predictions are contained within the reference exon (light blue); and exons with other or multiple overlap types (pink). Note the decrease in detection performance at high read coverage for Oases, Velvet and SLIDE, as well as the high frequency of imperfect overlaps for Tromer. See also Figure 2b.

Nature Methods: doi:10.1038/nmeth.2714



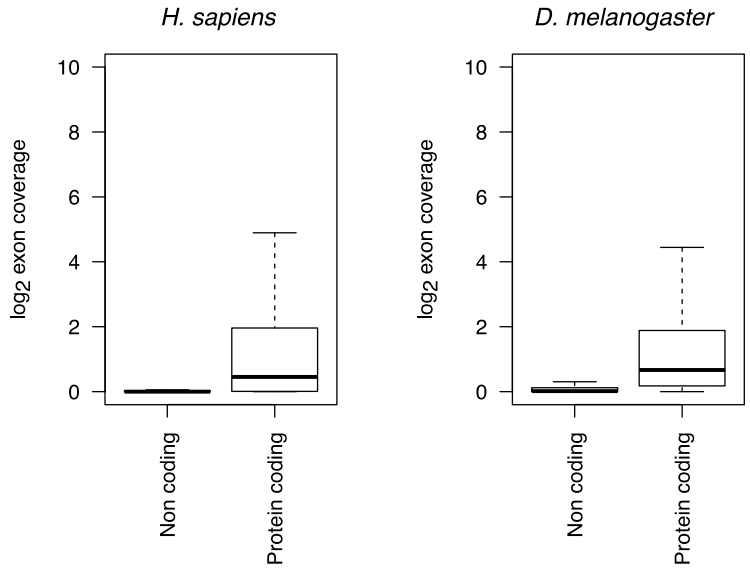
Supplementary Figure 7. Influence of sequencing coverage on non-coding exon-level sensitivity. Annotated exons of non-coding transcripts were binned according to RNA-seq read coverage and method sensitivities were calculated for each bin separately.

Nature Methods: doi:10.1038/nmeth.2714



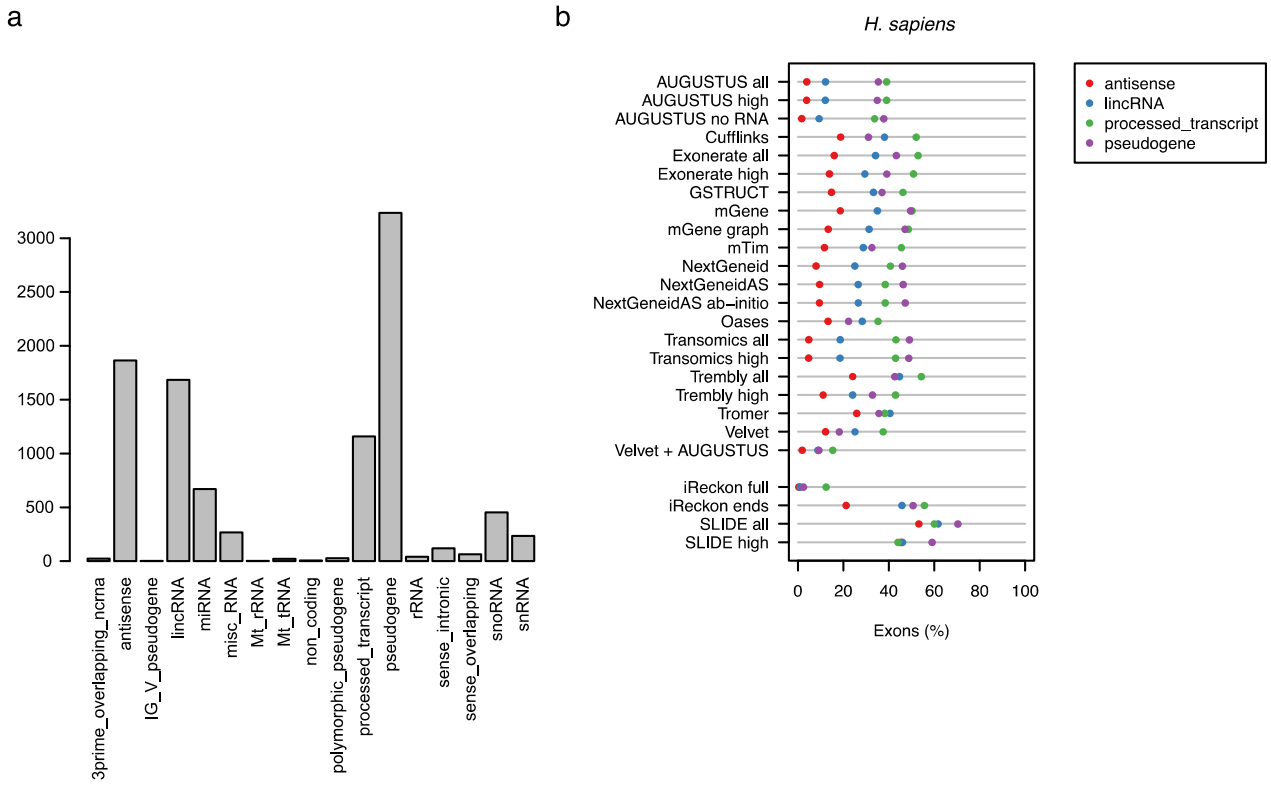
Supplementary Figure 8. Exon detection sensitivity relative to coding potential. Percentage of detected exons belonging to coding (green) and non-coding (red) transcripts in *H. sapiens* and *D. melanogaster*.

Nature Methods: doi:10.1038/nmeth.2714



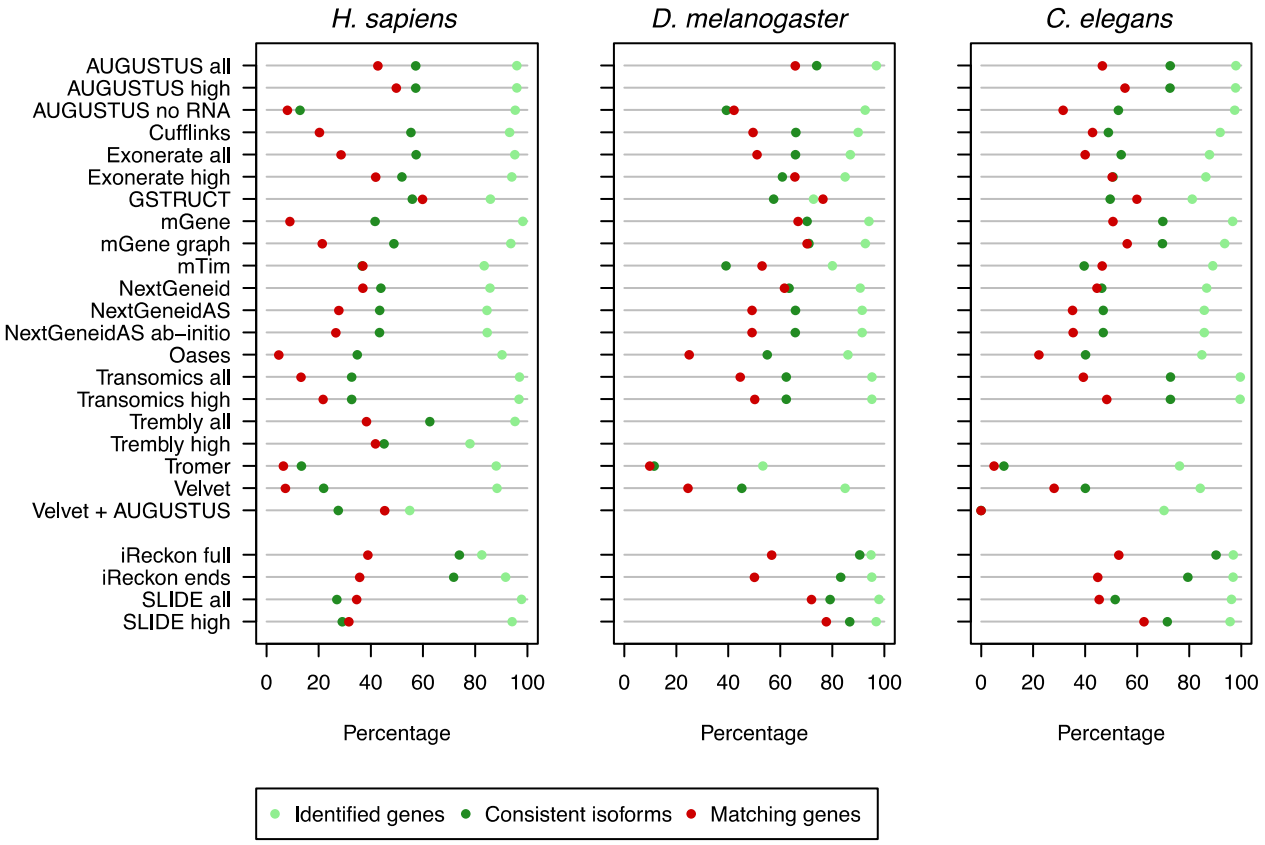
Supplementary Figure 9. RNA-seq read coverage for exons of coding and non-coding transcripts.

Nature Methods: doi:10.1038/nmeth.2714



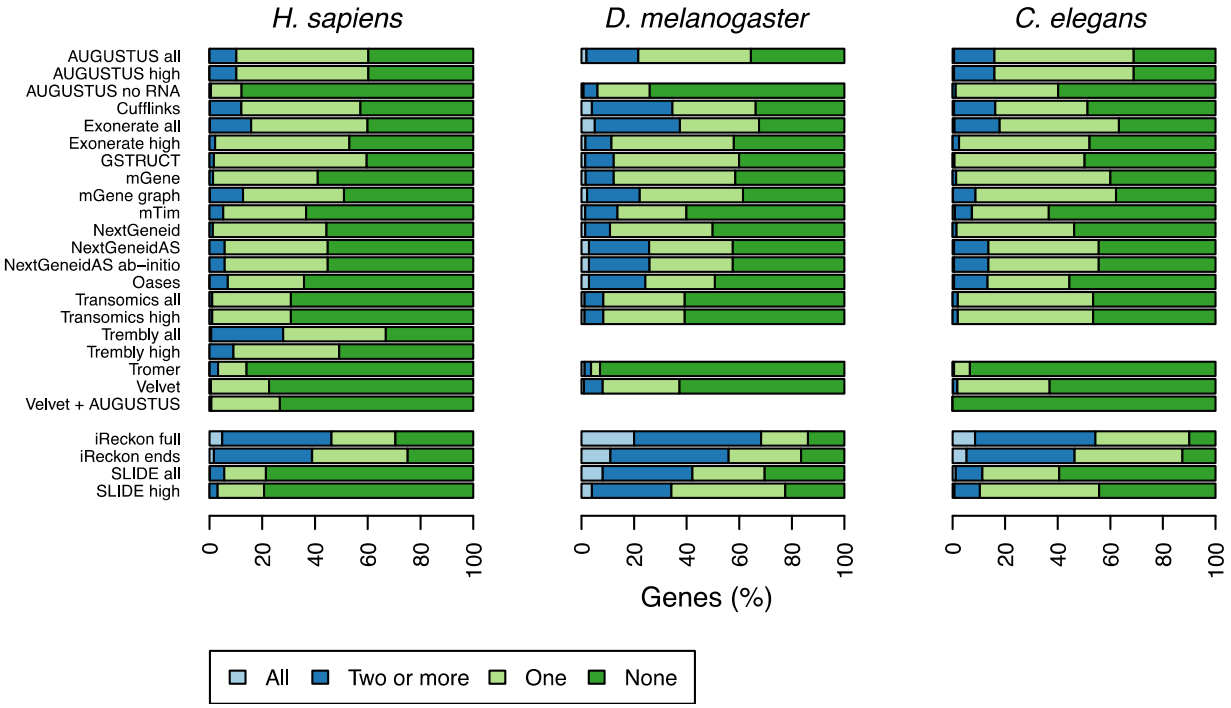
Supplementary Figure 10. Exon detection sensitivity for non-coding genes. (a) Non-coding transcripts expressed in the human RNA-seq data set, annotated by gene biotype. (b) Detected exons from transcripts of non-coding gene biotypes. Small RNAs were excluded, as they are underrepresented in data sets derived from standard mRNA-seq protocols that incorporate poly(A) selection. Alternative library construction protocols are required to specifically interrogate small RNA populations.

Nature Methods: doi:10.1038/nmeth.2714



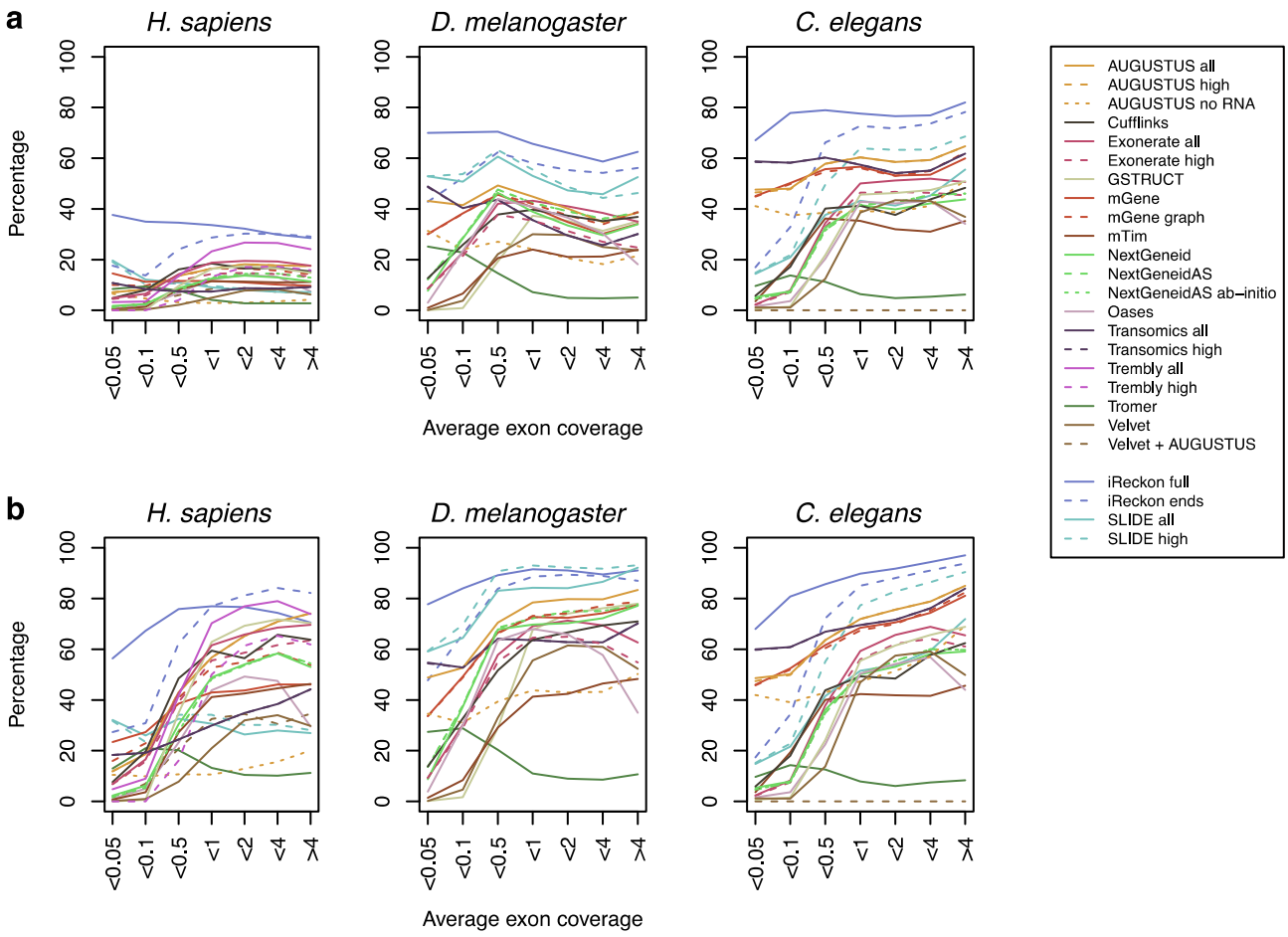
Supplementary Figure 11. Gene detection performance. Points indicate the percentage of reference genes with a matching assembled transcript (recall, green) and reported genes with at least one transcript matching the reference (precision, red).

Nature Methods: doi:10.1038/nmeth.2714



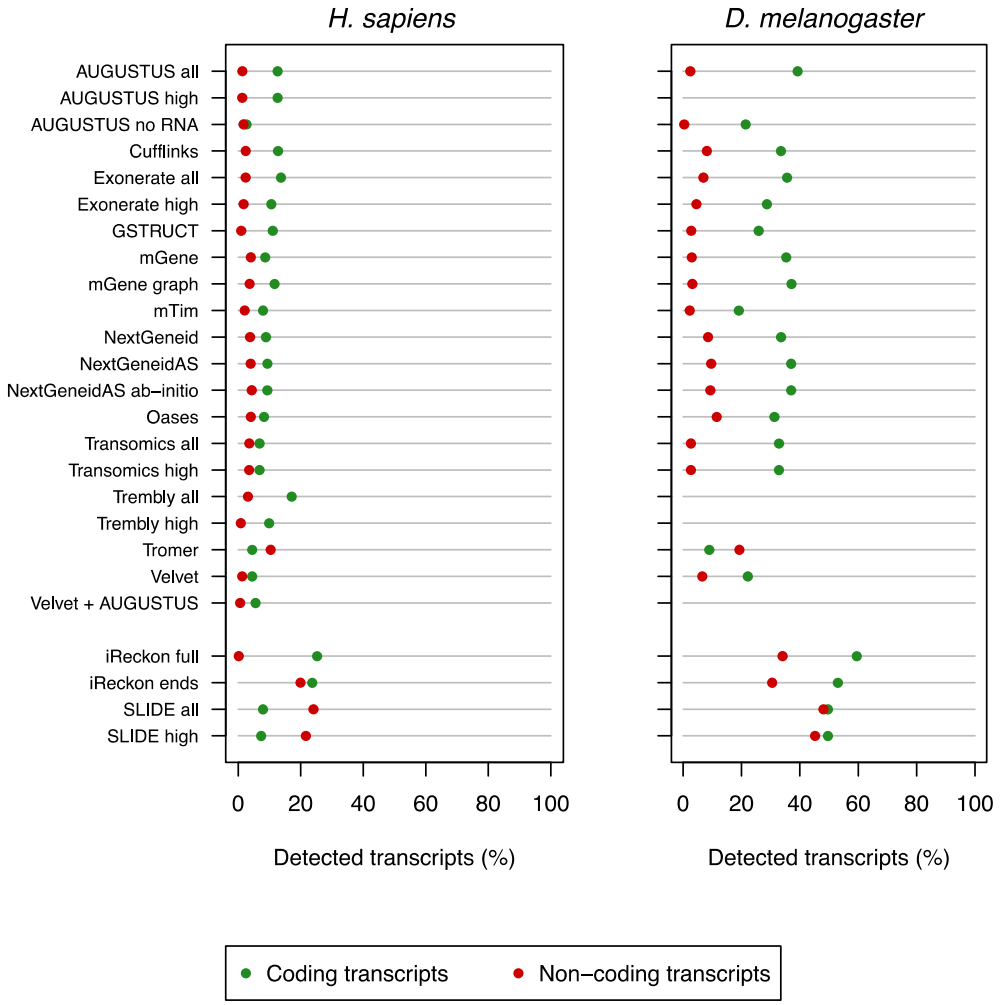
Supplementary Figure 12. Number of isoforms detected per gene. Genes with at least three annotated splice products for which various subsets have been reported.

Nature Methods: doi:10.1038/nmeth.2714



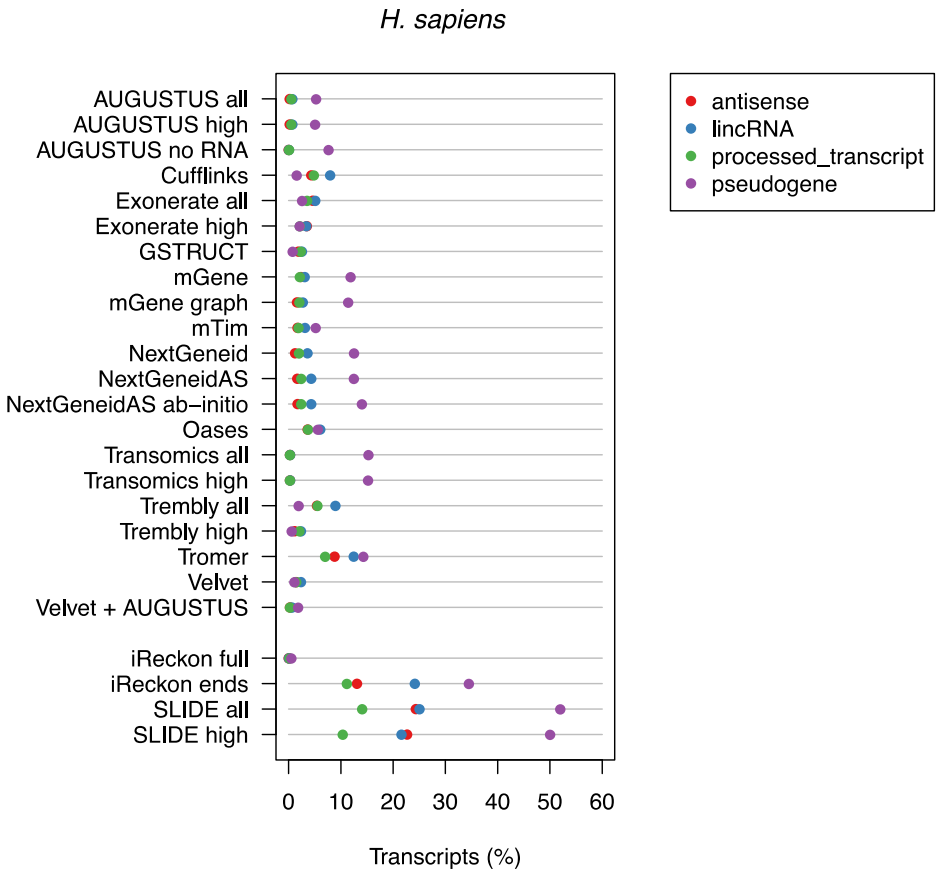
Supplementary Figure 13. Influence of read depth on transcript-level (a) and gene-level (b) sensitivity. Annotated transcripts and genes were binned according to RNA-seq read coverage and method sensitivities were calculated for each bin separately.

Nature Methods: doi:10.1038/nmeth.2714



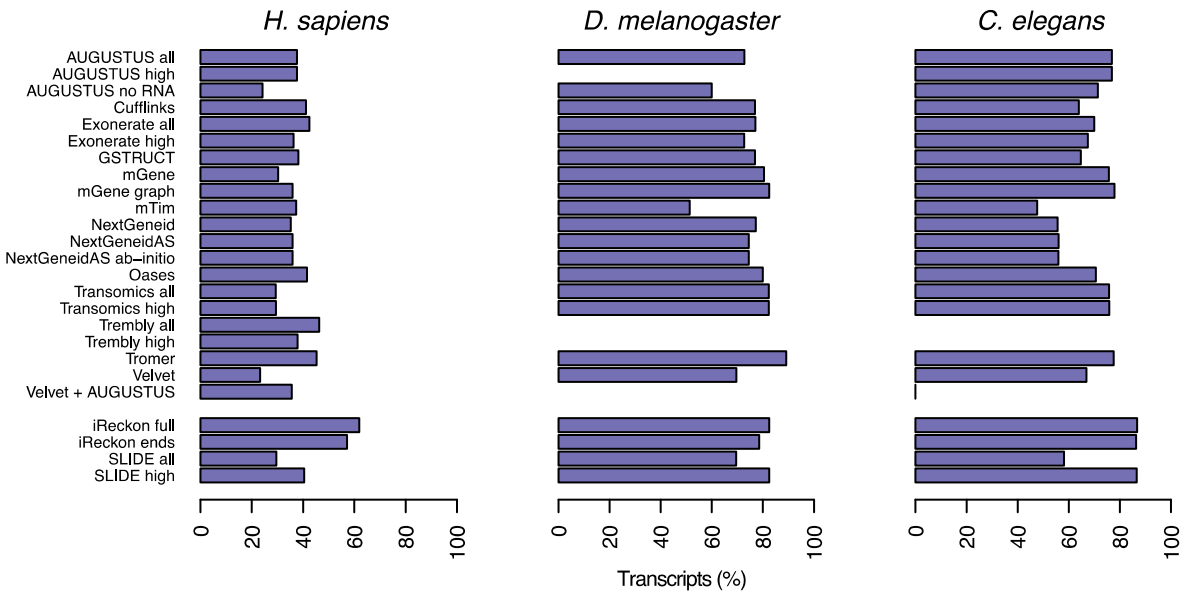
Supplementary Figure 14. Transcript level performance for coding and non-coding transcripts. Percentage of transcripts annotated in *H. sapiens* and *D. melanogaster* matching a reported transcript. Note, SLIDE shows higher sensitivity for non-coding transcripts as these tend to be shorter than protein coding transcripts. For transcripts with a given number of exons SLIDE identifies more protein coding than non-coding transcripts.

Nature Methods: doi:10.1038/nmeth.2714

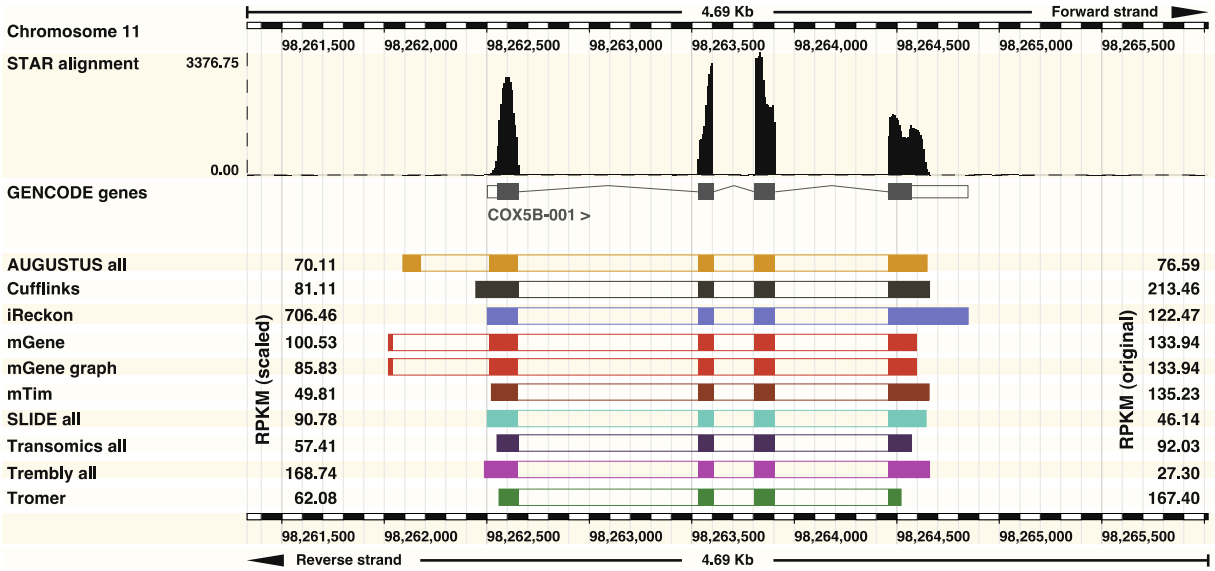


Supplementary Figure 15. Transcript detection sensitivity for non-coding genes. Transcripts identified in the human data set annotated by non-coding gene biotype.

Nature Methods: doi:10.1038/nmeth.2714

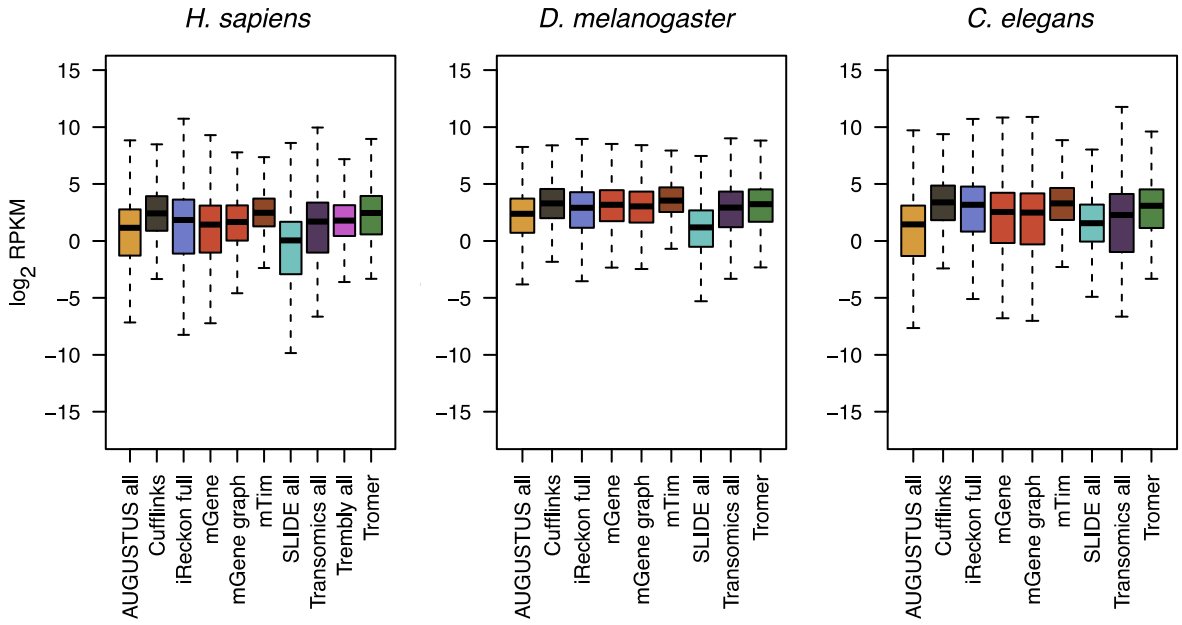


Supplementary Figure 16. Transcript assembly performance. Percentage of transcripts, for which all exons have been identified, that were correctly assembled to a full-length annotated splice variant.



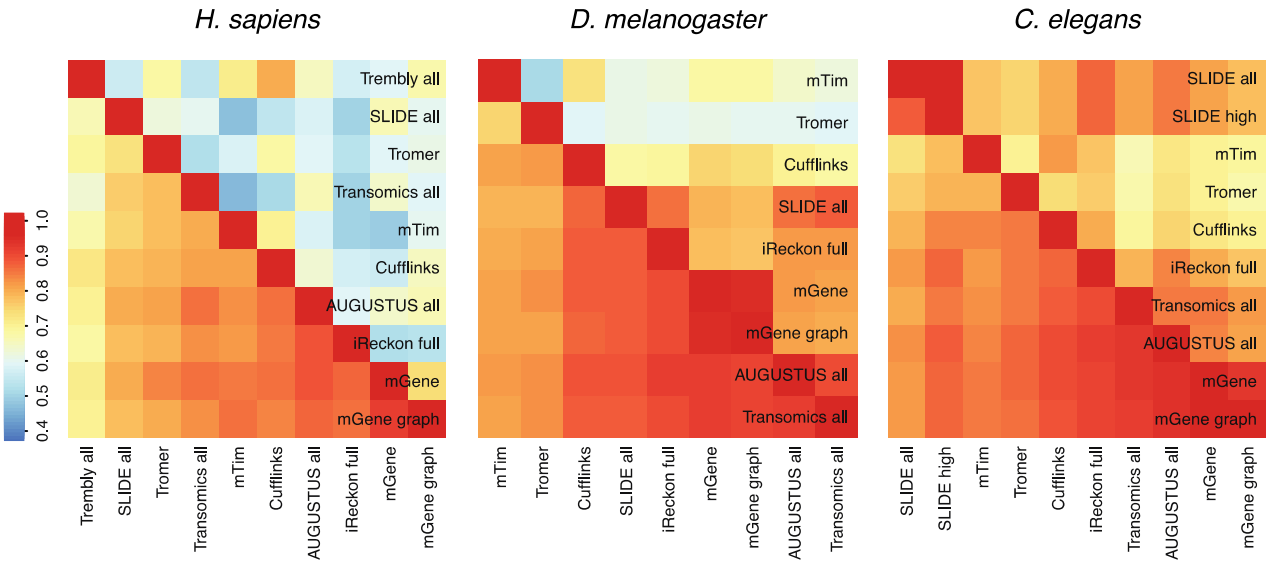
Supplementary Figure 17. Transcript predictions and expression level estimates (in RPKM) at the *COX5B* locus. Upper tracks depict RNA-seq read coverage (from STAR alignments; see Methods) and annotated genes. Exon predictions from the 10 methods that provided RPKM values are illustrated below the annotated gene by colored boxes. Exons reported as part the same transcript isoform are connected. iReckon full does not predict retained introns for this gene. Original and median-scaled RPKMs are presented to the right and left, respectively.

Nature Methods: doi:10.1038/nmeth.2714



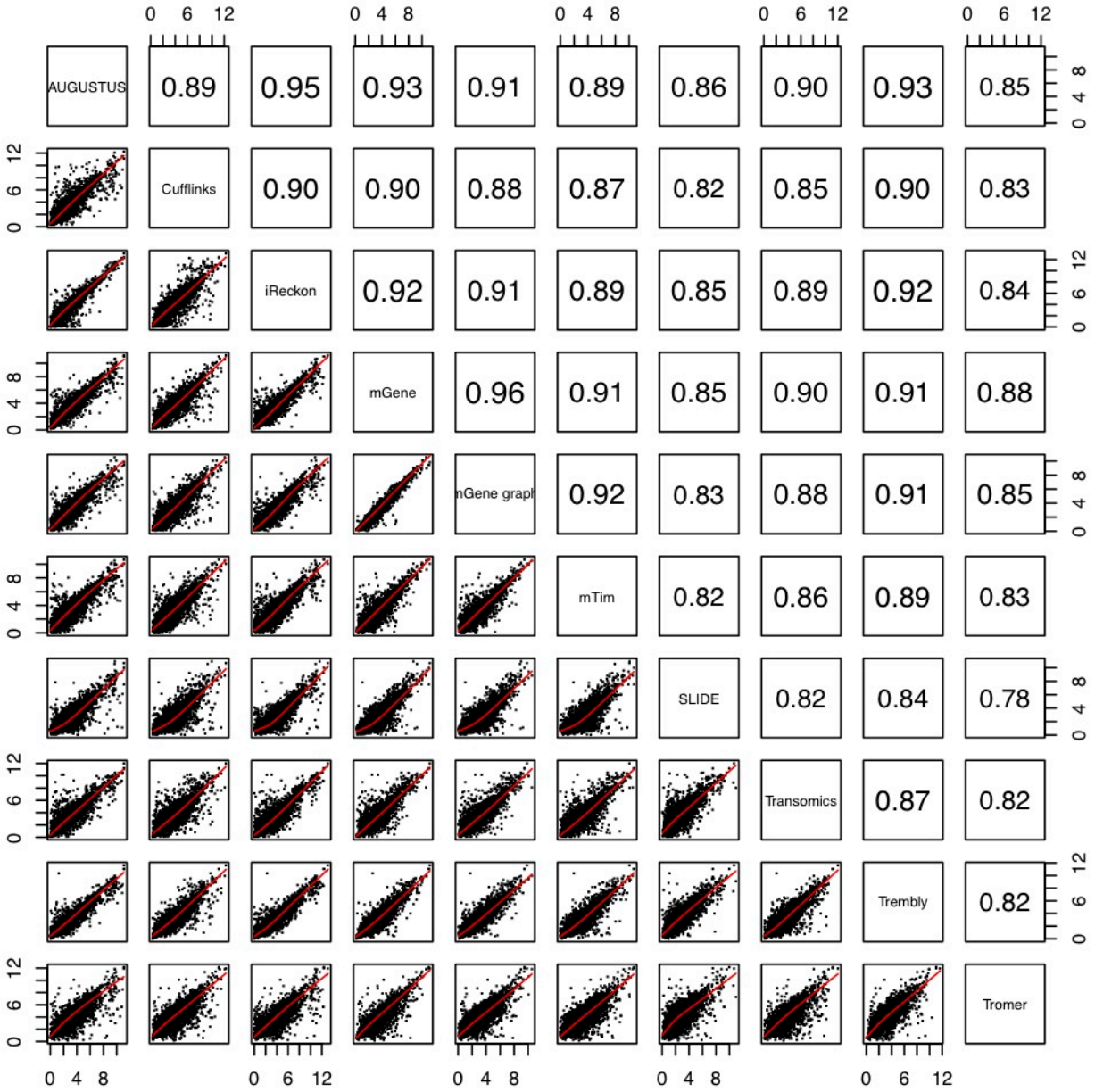
Supplementary Figure 18. Distribution of gene expression values (RPKM) for each method. Results are shown for annotated genes only. Where multiple transcripts were reported for the same gene, the highest RPKM value was used, corresponding to the predominant transcript identified by each method.

Nature Methods: doi:10.1038/nmeth.2714

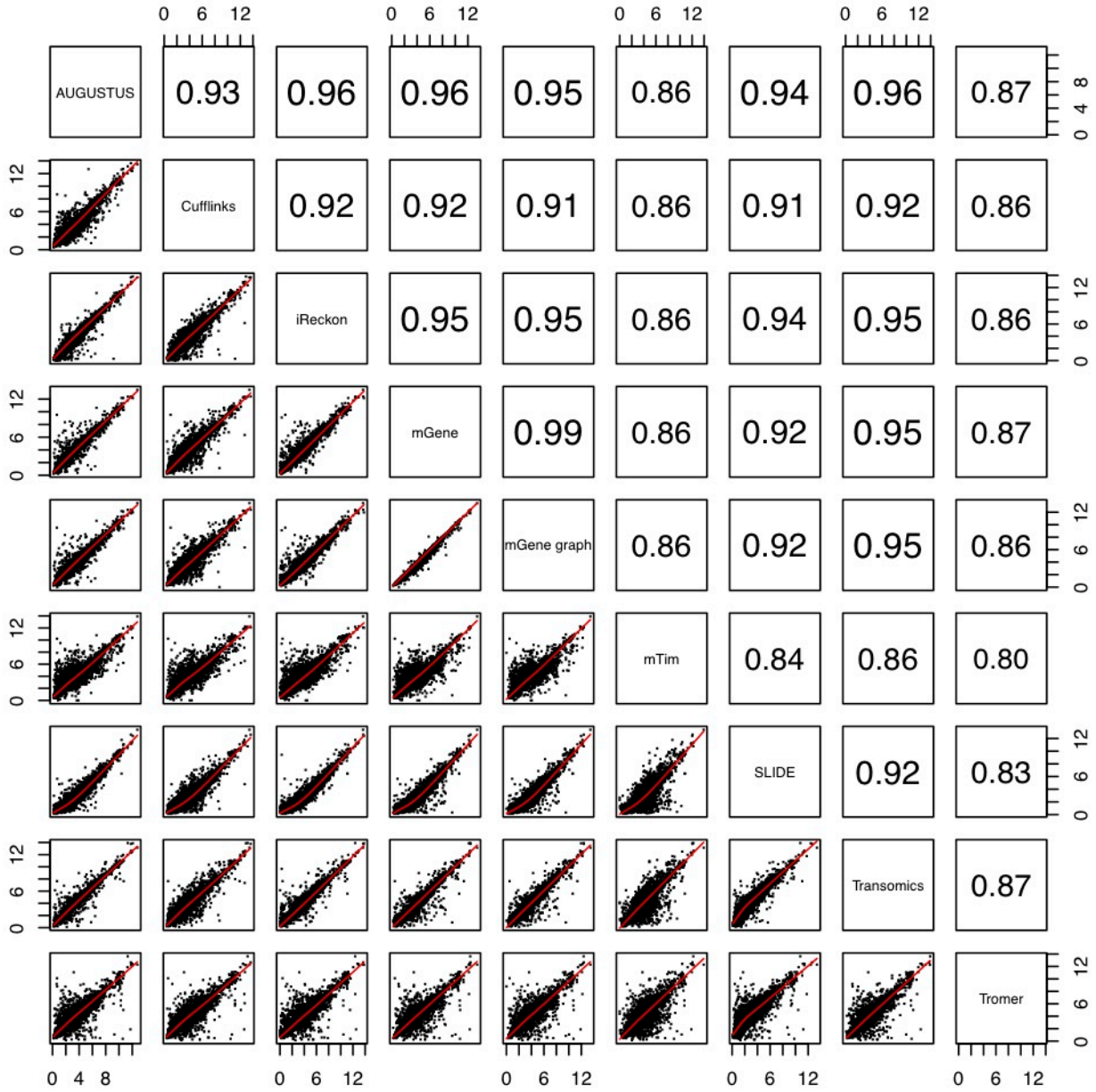


Supplementary Figure 19. Pairwise agreement between methods. Lower triangles show expression correlation (Pearson *r* of log RPKM) for the set of genes identified by all methods. Upper triangles depict the proportion of genes shared between each method pair, i.e. the number of genes identified as expressed in both divided by number of genes identified as expressed in either. Methods were ordered by hierarchical clustering using 1–*r* as the distance metric.

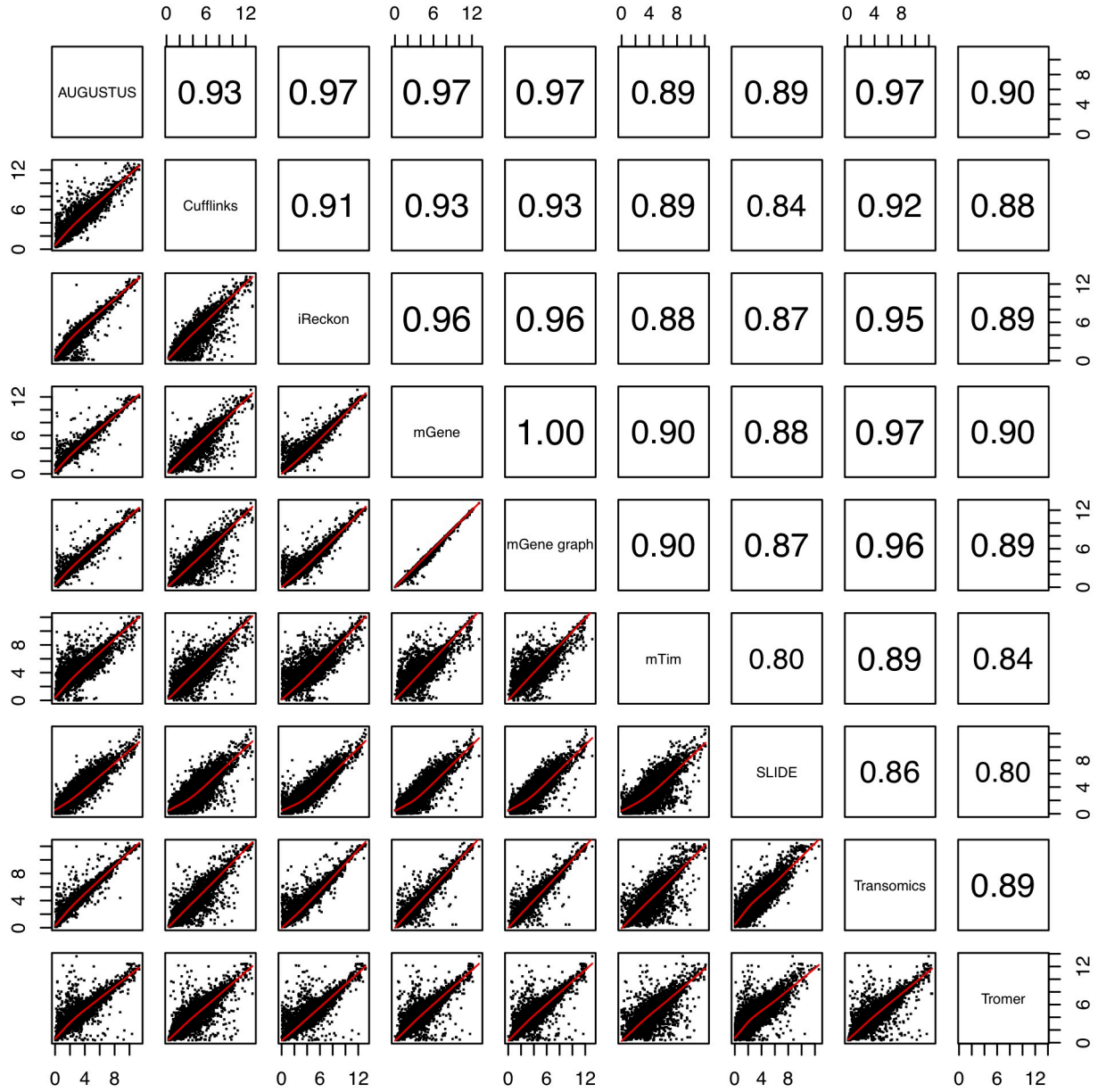
Nature Methods: doi:10.1038/nmeth.2714



Supplementary Figure 20. Comparison of quantification methods for *H. sapiens*. For each pair of methods, scatter plots relate $\log_2$ RPKM values for the genes identified by all methods. The corresponding correlation coefficients (Pearson *r*) are shown opposite. Where multiple transcripts were reported for the same gene, the highest RPKM value was used, corresponding to the predominant transcript identified by each method. RPKM values for AUGUSTUS, iReckon, SLIDE, Transomics and Trembly correspond to the values reported by their 'all', and 'full' protocols.

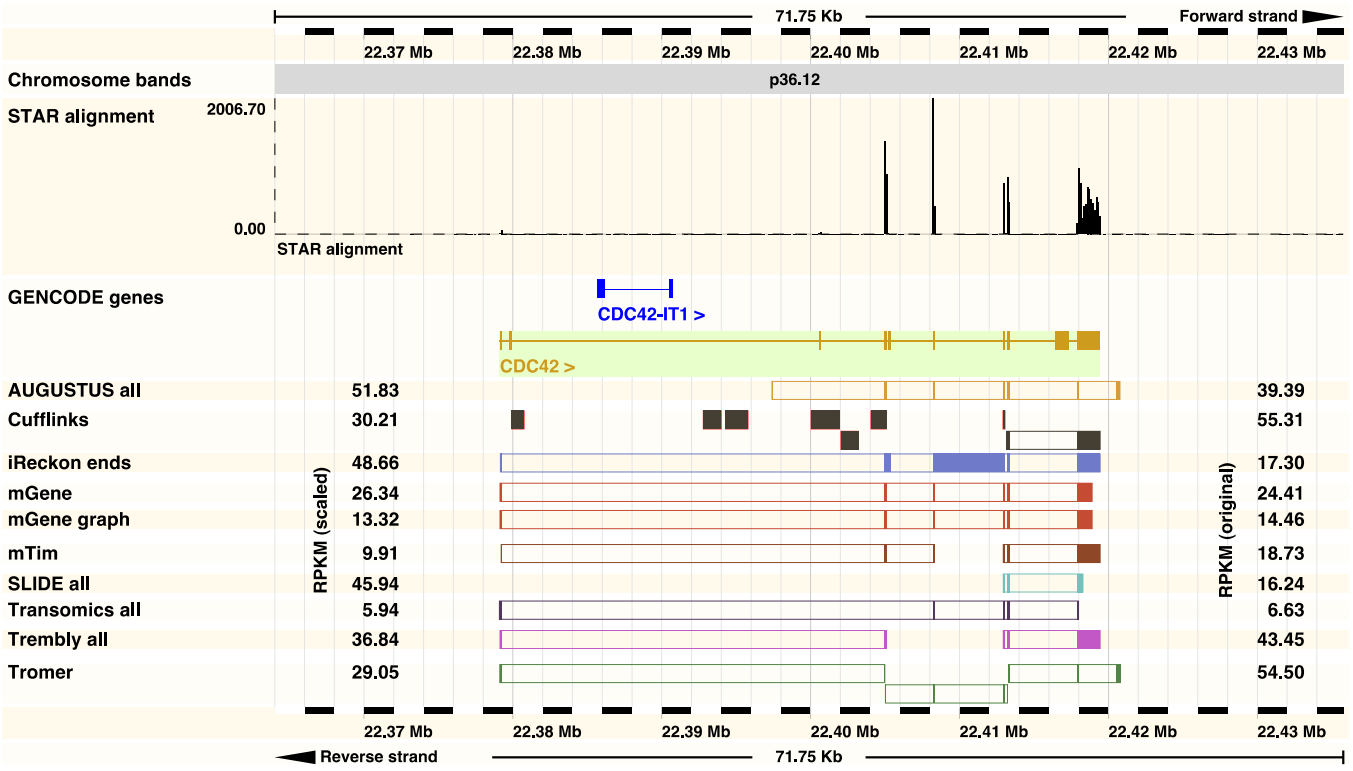


Supplementary Figure 21. Comparison of quantification methods for *D. melanogaster*. See Supplementary Figure 20 for details.



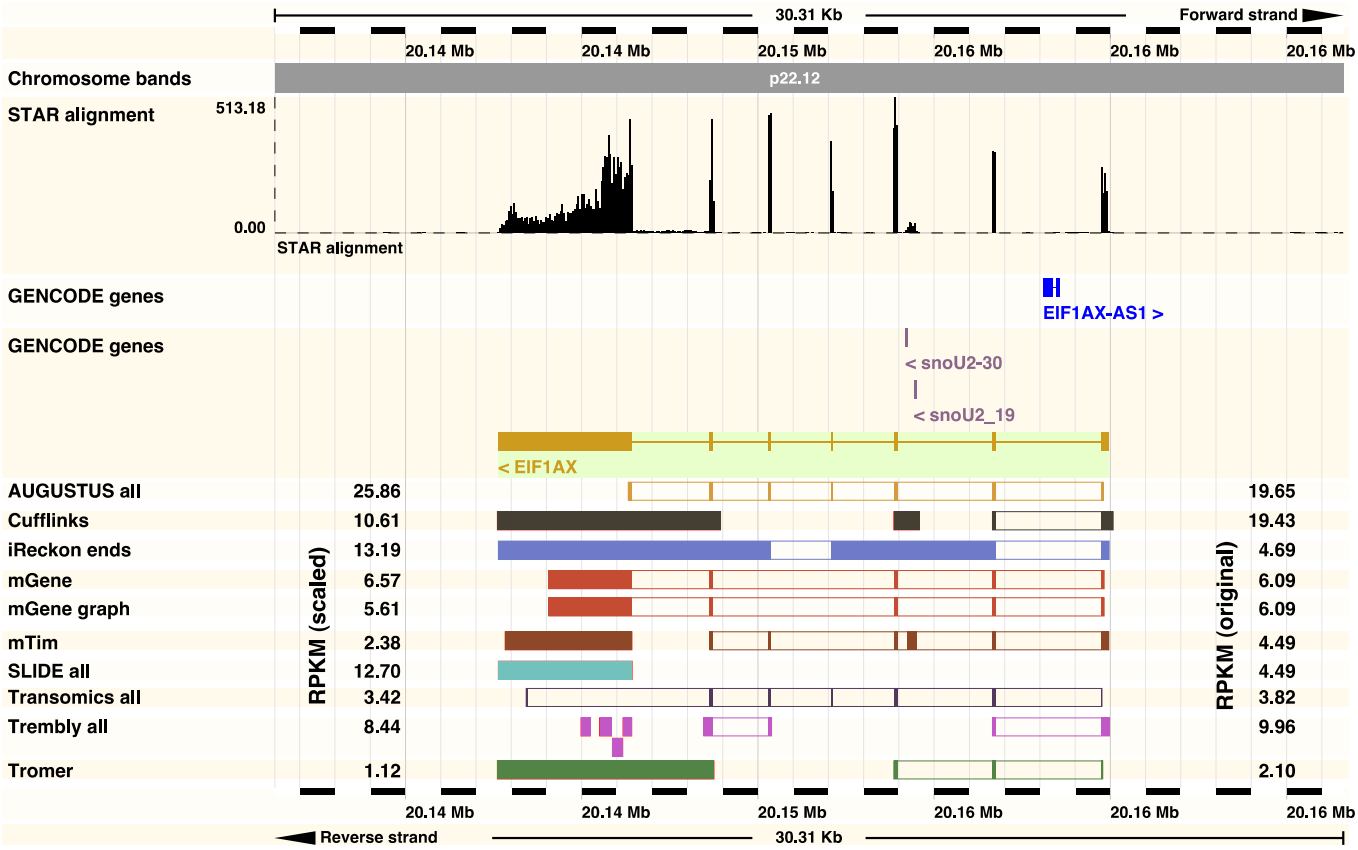
Supplementary Figure 22. Comparison of quantification methods for *C. elegans*. See Supplementary Figure 20 for details.

Nature Methods: doi:10.1038/nmeth.2714



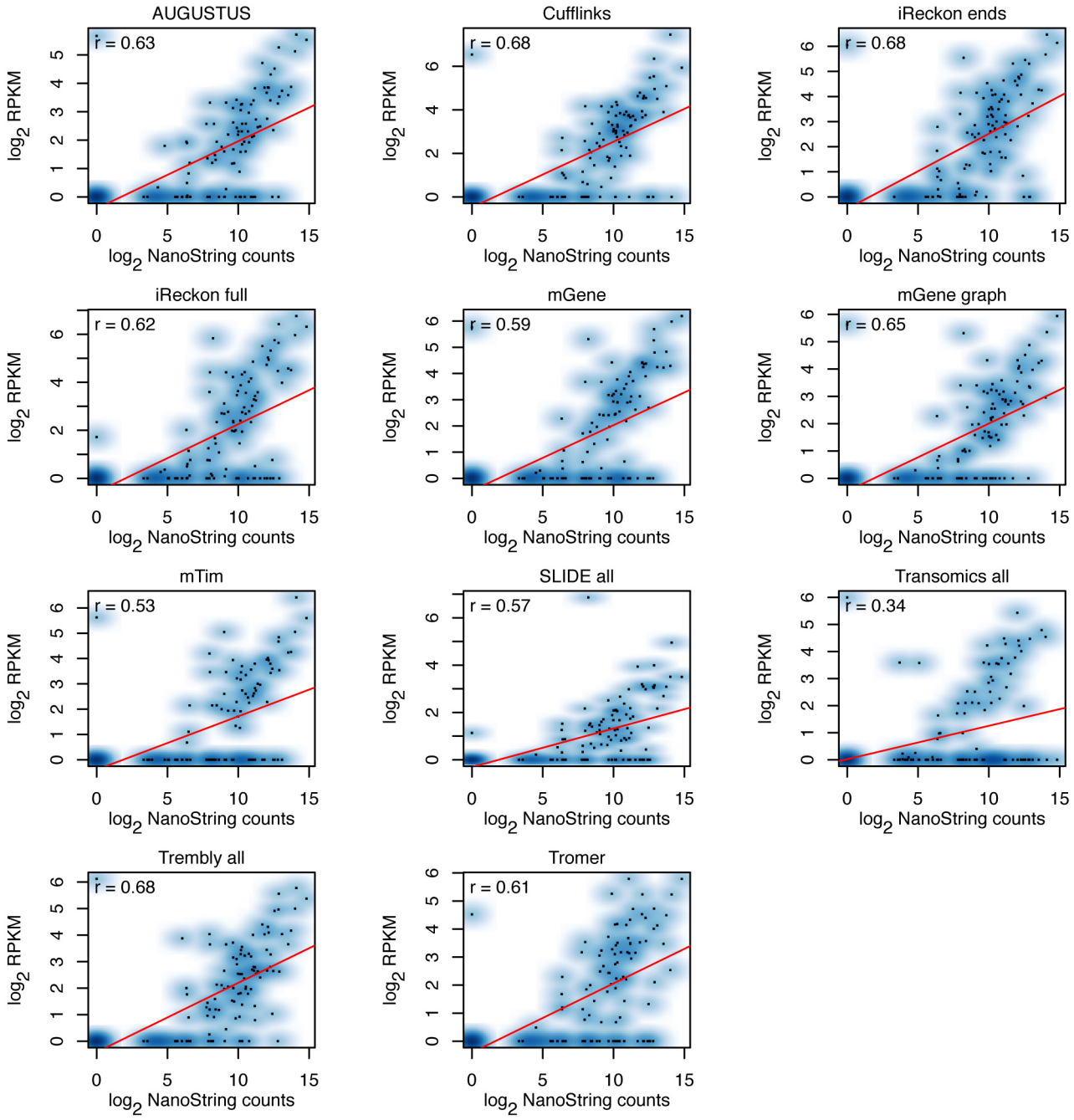
Supplementary Figure 23. Transcript predictions and expression level estimates (in RPKM) at the *CDC42* locus. Upper tracks depict RNA-seq read coverage (from STAR alignments; see Methods) and annotated genes. Exon predictions from the 10 methods that provided RPKM values are illustrated below the annotated gene by colored boxes. Exons reported as part the same transcript isoform are connected. iReckon full does not predict retained introns for this gene. Original and median-scaled RPKMs are presented to the right and left, respectively.

Nature Methods: doi:10.1038/nmeth.2714



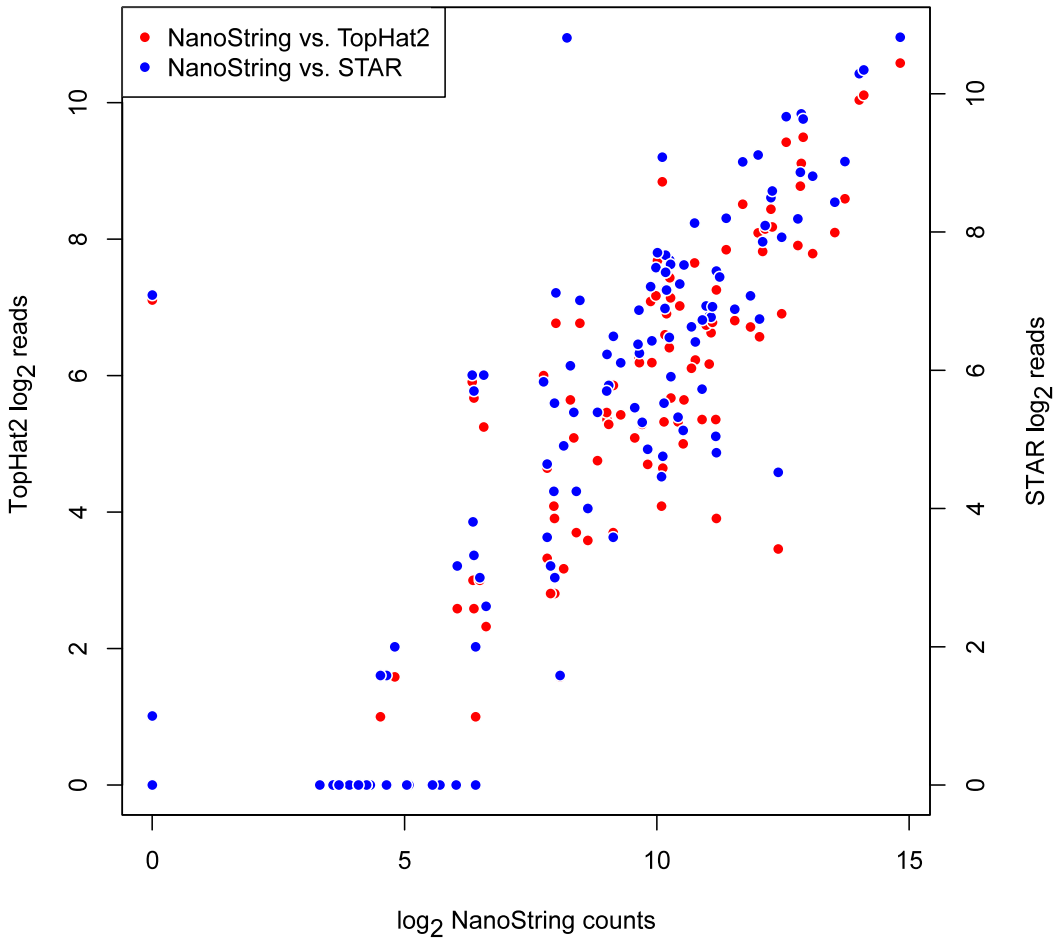
Supplementary Figure 24. Transcript predictions and expression level estimates (in RPKM) at the *EIF1AX* locus. See Supplementary Fig. 23 for details. iReckon full does not predict retained introns for this gene.

Nature Methods: doi:10.1038/nmeth.2714



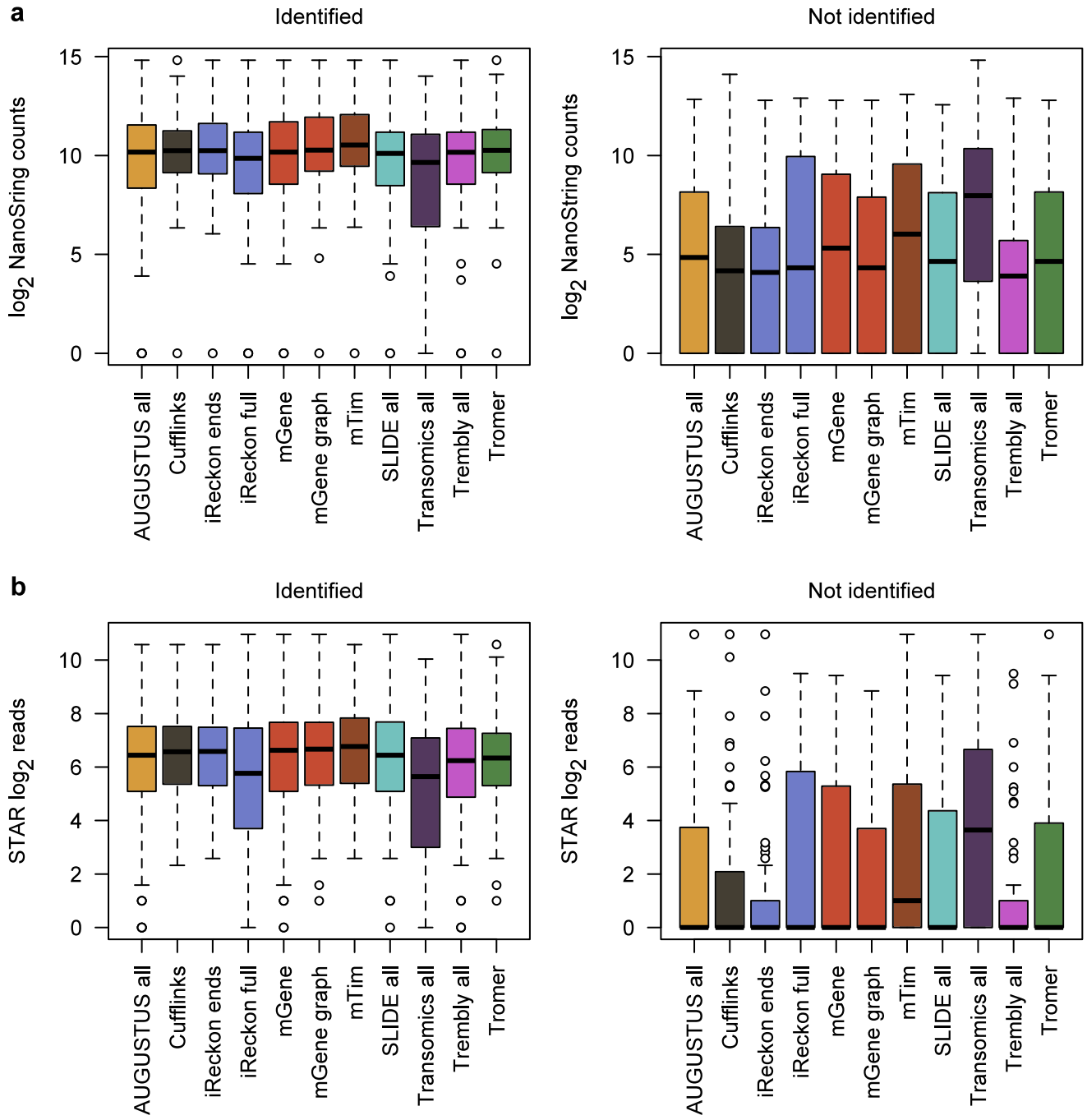
Supplementary Figure 25. Correlation between NanoString counts and transcript RPKMs. Scatter plots show individual data points in black, with color intensity indicating the density of data points. Predicted transcripts were required to contain the exon or junction targeted by the NanoString probe. Where multiple such transcripts were reported for the same gene, the highest RPKM value was used. Where no such transcript was reported, an RPKM of zero was assigned. Correlation coefficients (Pearson *r*) are given for each comparison. Expression values were incremented by 1 prior to log transformation to avoid infinite numbers. Notably, the protocol iReckon ends identifies more genes than iReckon full. When provided with complete gene annotation, iReckon often fails to resolve transcripts in complex loci with many annotated isoforms. This occurs less frequently when the program is given only transcript boundaries.

Nature Methods: doi:10.1038/nmeth.2714



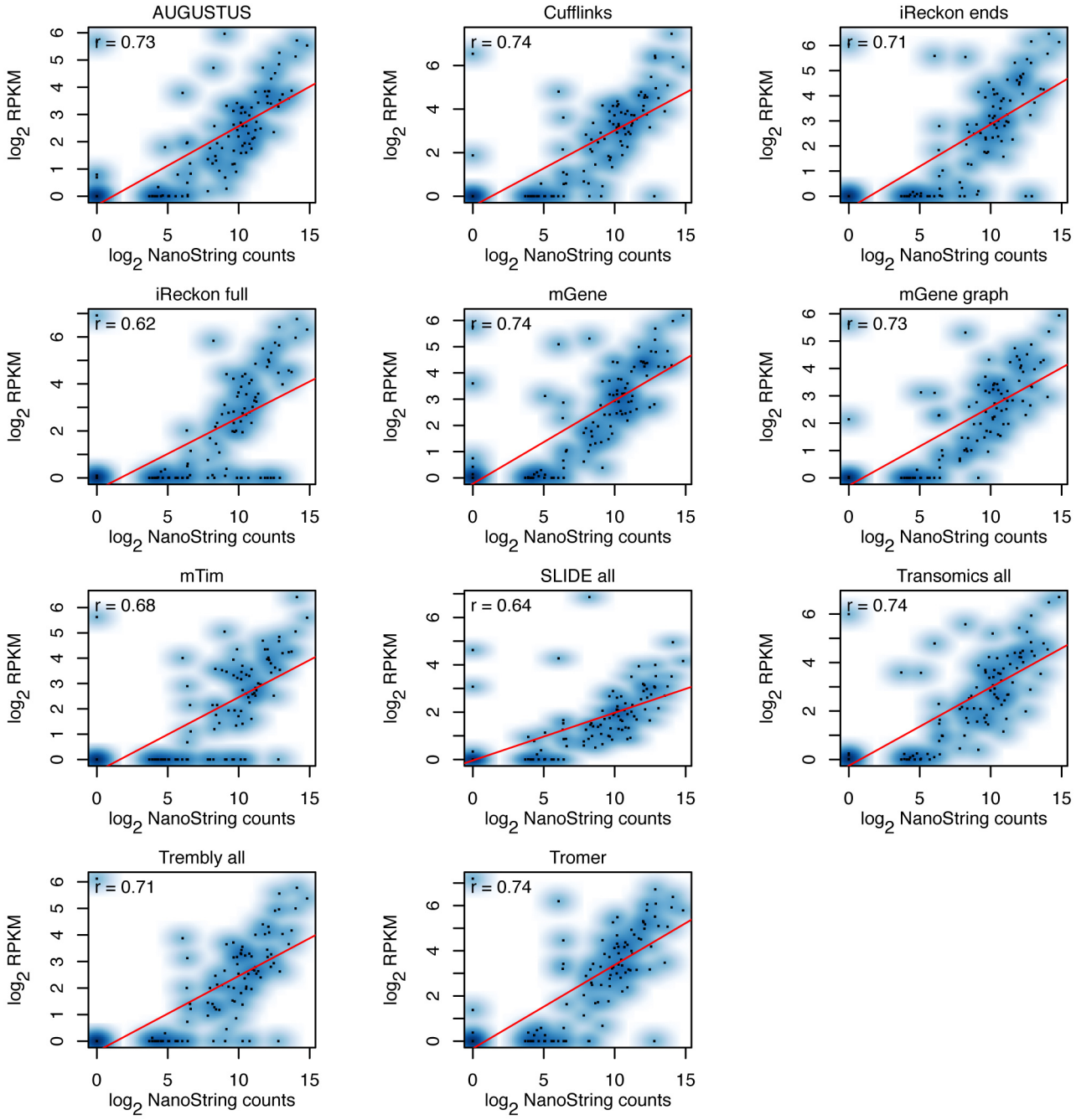
Supplementary Figure 26. Correlation between NanoString counts and numbers of mapped reads for targeted exons and junctions. Scatter plots show individual data points in red (Tophat) and blue (STAR). Count values were incremented by 1 prior to log transformation to avoid infinite numbers.

Nature Methods: doi:10.1038/nmeth.2714



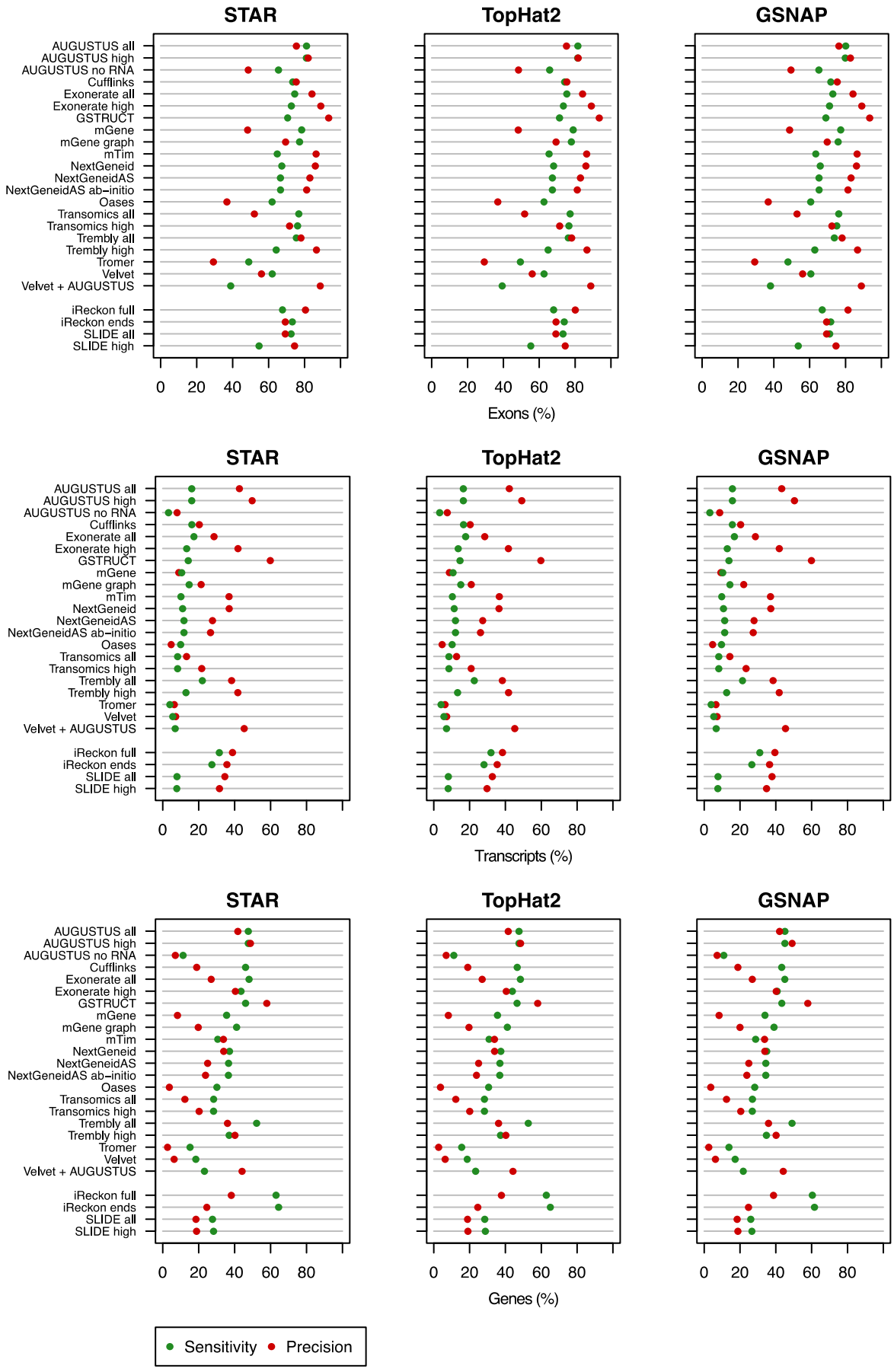
Supplementary Figure 27. Distribution of NanoString counts (a) and mapped reads by the STAR aligner (b) for probes depending on whether a method identified an isoform consistent with a probe (left) or not (right). Both mTim and Transomics failed to identify many exons or junctions targeted by NanoString probes with RNA-seq read support. Count values were incremented by 1 prior to log transformation to avoid infinite numbers.

Nature Methods: doi:10.1038/nmeth.2714



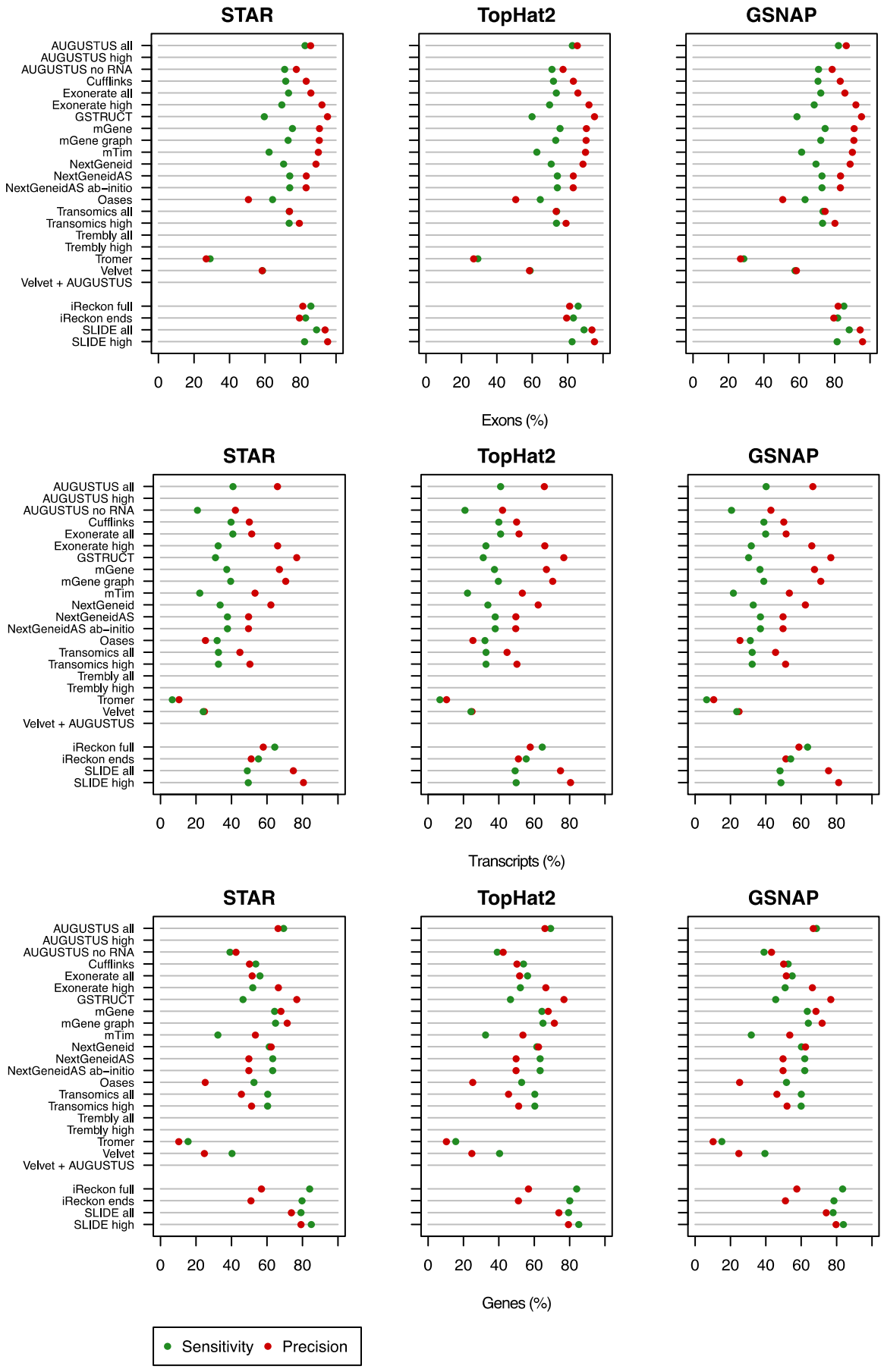
Supplementary Figure 28. Correlation between NanoString counts and gene RPKMs. Scatter plots show individual data points in black, with color intensity indicating the density of data points. Where multiple transcripts were reported for the same gene, the highest RPKM value was used (irrespective of whether that transcript contained the exon or junction targeted by the NanoString probe). Correlation coefficients (Pearson *r*) are given for each comparison.

Nature Methods: doi:10.1038/nmeth.2714



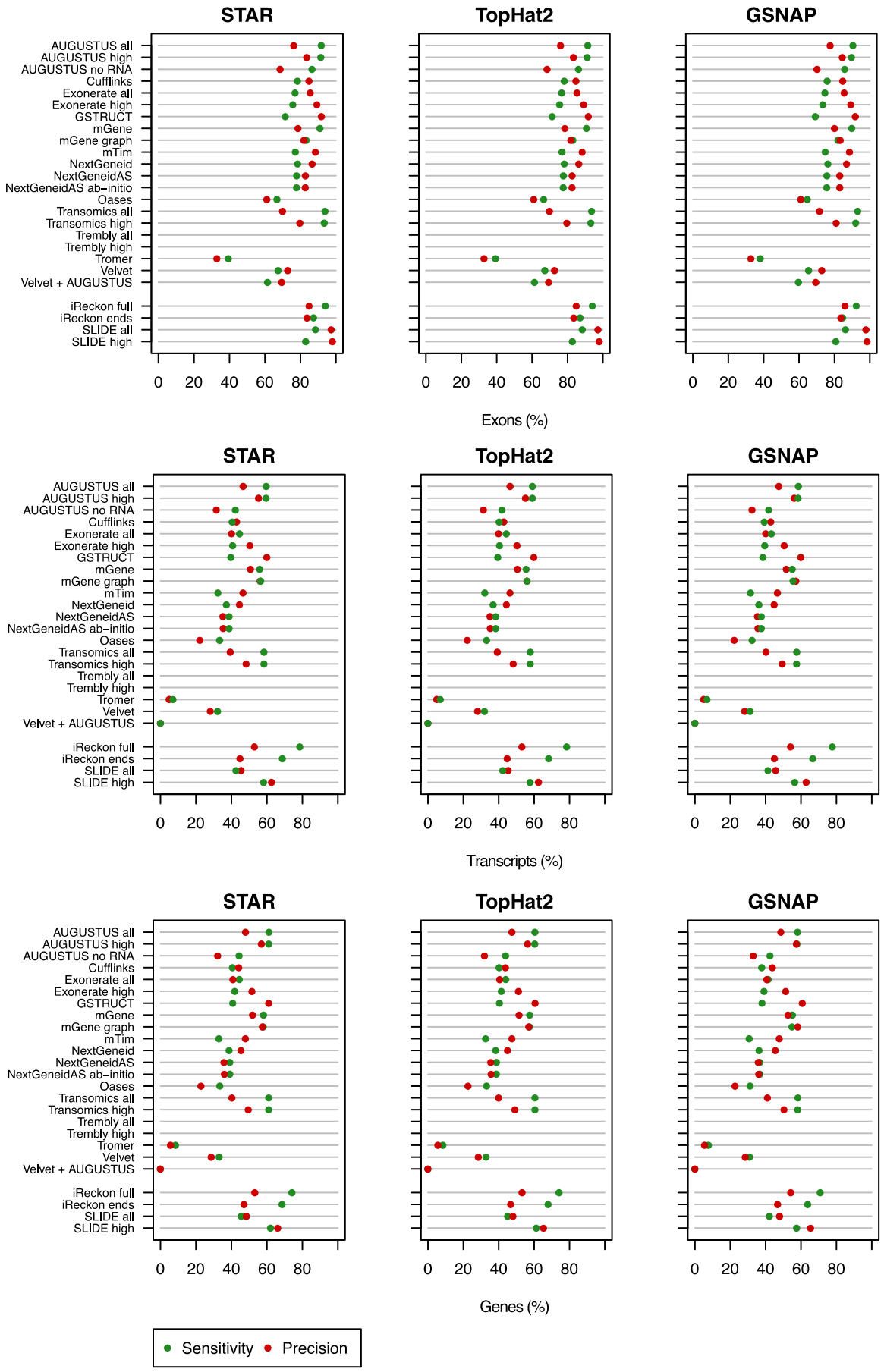
Supplementary Figure 29. Influence of different aligners on annotation usage (*H. sapiens*). Exon, transcript and gene level performance relative to filtered annotation based on the aligners STAR, TopHat2 and GSNAP.

Nature Methods: doi:10.1038/nmeth.2714



Supplementary Figure 30. Influence of different aligners on annotattion usage (*D. melanogaster*). See Supplementary Fig. 29 for details.

Nature Methods: doi:10.1038/nmeth.2714



Supplementary Figure 31. Influence of different aligners on annotation usage (*C. elegans*). See Supplementary Fig. 29 for details.

Nature Methods: doi:10.1038/nmeth.2714

Supplementary Table 1. Developer team submission details

Developer team	Protocol designation	Underlying alignment programs	Coding sequence predicted	Use of reference annotation	Quantified features	Multiple transcripts reported per gene
H. sapiens						
Iseli	Tromer	fetchGWI, megablast, SIBsim4	yes	no	transcript	yes
Gerstein	Trembly all	TopHat	no	no	transcript	yes
	Trembly high	TopHat	no	no	transcript	yes
Rätsch	mGene	PALMapper	yes	no	transcript	yes
	mGene graph	PALMapper	yes	no	transcript	yes
	mTim	PALMapper	no	no	transcript	yes
Richard	Oases	BLAT	no	no	no	yes
n.a.	Cufflinks	TopHat	no	no	transcript	yes
Stanke	AUGUSTUS high	BLAT	yes	no	transcript	yes
	AUGUSTUS all	BLAT	yes	no	transcript	yes
	AUGUSTUS de-novo	n.a	yes	no	transcript	no
Searle	Exonerate SM all	Exonerate	yes	no	no	no
	Exonerate SM high	Exonerate	yes	no	no	no
Wu	GSTRUCT	GSNAP	no	no	no	no
Guigo	Nextgeneid	GEM	yes	no	no	no
	NextgeneidAS	GEM	yes	no	no	no
	NextgeneidAS de-novo	GEM	yes	no	no	no
Solovyev	Transomics all		yes	no	transcript	no
	Transomics high		yes	no	transcript	no
Wold	Velvet	BLAT	no	no	exon	yes
	Velvet AUGUSTUS	BLAT	no	no	exon	yes
n.a	iReckon full	TopHat	no	yes	transcript	yes
	iReckon ends	TopHat	no	yes	transcript	yes
n.a	SLIDE all	TopHat	no	yes	transcript	yes
	SLIDE high	TopHat	no	yes	transcript	yes
D. melanogaster						
Iseli	Tromer	fetchGWI, megablast, SIBsim4	yes	no	transcript	yes
Rätsch	mGene	PALMapper	yes	no	transcript	yes
	mGene graph	PALMapper	yes	no	transcript	yes
	mTim	PALMapper	no	no	transcript	yes
Richard	Oases	BLAT	no	no	no	yes
n.a.	Cufflinks	TopHat	no	no	transcript	yes
Stanke	AUGUSTUS all	BLAT	yes	no	transcript	yes
	AUGUSTUS de-novo	n.a.	yes	no	transcript	no
Wu	GSTRUCT	GSNAP	no	no	no	no
Guigo	Nextgeneid	GEM	yes	no	no	no
	NextgeneidAS	GEM	yes	no	no	no
	NextgeneidAS de-novo	GEM	yes	no	no	no
Solovyev	Transomics all		yes	no	transcript	no
	Transomics high		yes	no	transcript	no
Wold	Velvet	BLAT	no	no	exon	yes
n.a.	iReckon full	TopHat	no	yes	transcript	yes
	iReckon ends	TopHat	no	yes	transcript	yes
n.a.	SLIDE all	TopHat	no	yes	transcript	yes
	SLIDE high	TopHat	no	yes	transcript	yes
C. elegans						
Iseli	Tromer	fetchGWI, megablast, SIBsim4	yes	no	transcript	yes
Rätsch	mGene	PALMapper	yes	no	transcript	yes
	mGene graph	PALMapper	yes	no	transcript	yes
	mTim	PALMapper	no	no	transcript	yes
Richard	Oases	BLAT	no	no	no	yes
n.a.	Cufflinks	TopHat	no	no	transcript	yes
Stanke	AUGUSTUS high	BLAT	yes	no	transcript	yes
	AUGUSTUS all	BLAT	yes	no	transcript	yes
	AUGUSTUS de-novo	n.a.	yes	no	transcript	no
Searle	Exonerate SM all	Exonerate	yes	no	no	no
	Exonerate SM high	Exonerate	yes	no	no	no
Wu	GSTRUCT	GSNAP	no	no	no	no
Guigo	Nextgeneid	GEM	yes	no	no	no
	NextgeneidAS	GEM	yes	no	no	no
	NextgeneidAS de-novo	GEM	yes	no	no	no
Solovyev	Transomics all		yes	no	transcript	no
	Transomics high		yes	no	transcript	no
Wold	Velvet	BLAT	no	no	exon	yes
	Velvet AUGUSTUS	BLAT	no	no	exon	yes
n.a.	iReckon full	TopHat	no	yes	transcript	yes
	iReckon ends	TopHat	no	yes	transcript	yes
n.a.	SLIDE all	TopHat	no	yes	transcript	yes
	SLIDE high	TopHat	no	yes	transcript	yes

Nature Methods: doi:10.1038/nmeth.2714

Supplementary Table 2. Nucleotide-level performance

	H. sapiens		*D. melanogaster*		*C. elegans*	
	Sensitivity	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**
AUGUSTUS all	63.77%	77.14%	87.17%	89.26%	96.88%	65.28%
AUGUSTUS high	63.70%	84.75%			96.65%	72.32%
AUGUSTUS no RNA	55.46%	44.91%	83.93%	83.63%	95.68%	60.17%
Cufflinks	79.31%	59.98%	84.98%	90.87%	88.61%	75.81%
Exonerate all	66.89%	87.21%	77.19%	92.87%	84.82%	78.73%
Exonerate high	65.01%	90.33%	75.22%	94.54%	83.78%	80.12%
GSTRUCT	60.59%	86.08%	58.13%	88.91%	76.67%	69.75%
iReckon full	78.33%	31.29%	89.72%	77.56%	96.08%	83.42%
iReckon ends	83.44%	10.01%	92.40%	69.20%	97.17%	73.29%
mGene	71.38%	42.86%	75.33%	87.87%	95.28%	72.73%
mGene graph	67.56%	55.51%	71.98%	90.68%	85.81%	75.70%
mTim	54.06%	90.24%	61.97%	93.06%	78.08%	81.31%
NextGeneid	64.73%	80.59%	80.36%	94.22%	84.51%	79.16%
NextGeneidAS	63.48%	81.01%	79.51%	94.44%	83.32%	79.13%
NextGeneidAS ab-initio	62.95%	79.72%	79.32%	94.37%	83.26%	79.32%
Oases	68.28%	68.53%	75.98%	91.65%	76.58%	80.52%
SLIDE all	89.28%	88.33%	96.67%	96.15%	93.36%	96.90%
SLIDE high	76.48%	87.87%	91.94%	96.64%	88.41%	97.07%
Transomics all	45.65%	53.57%	68.35%	83.70%	98.36%	72.00%
Transomics high	45.29%	75.06%	68.30%	88.99%	98.12%	82.61%
Trembly all	70.54%	85.99%				
Trembly high	53.24%	91.41%				
Tromer	84.67%	39.50%	92.06%	76.70%	87.74%	63.15%
Velvet	65.79%	78.69%	72.54%	88.07%	77.73%	82.18%
Velvet + AUGUSTUS	37.66%	75.51%			65.72%	76.78%

Nature Methods: doi:10.1038/nmeth.2714

Supplementary Table 3. Exon-, transcript- and gene-level performance for CDS reconstruction

	Exon		Transcript		Gene	
	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
H. sapiens						
AUGUSTUS all	66.18%	75.03%	19.51%	43.70%	61.47%	45.64%
AUGUSTUS high	66.09%	81.46%	19.50%	49.45%	61.46%	53.23%
AUGUSTUS no RNA	54.96%	48.88%	5.34%	9.28%	17.61%	9.28%
Exonerate all	57.36%	85.11%	19.77%	31.88%	58.12%	31.88%
Exonerate high	56.04%	89.39%	16.24%	42.65%	54.29%	42.65%
mGene	63.13%	50.32%	14.62%	10.01%	50.01%	10.02%
mGene graph	53.49%	82.44%	16.03%	34.44%	49.33%	46.01%
mTim	28.76%	92.55%	8.82%	46.66%	27.52%	52.53%
NextGeneid	50.47%	85.22%	11.29%	38.01%	40.96%	38.01%
NextGeneidAS	50.11%	82.48%	11.77%	31.47%	39.84%	31.47%
NextGeneidAS ab-initio	50.14%	80.49%	11.76%	29.20%	39.82%	29.20%
Transomics all	66.23%	50.68%	11.10%	14.59%	39.52%	14.59%
Transomics high	65.58%	69.73%	11.10%	23.89%	39.51%	23.89%
Tromer	31.58%	29.65%	2.23%	0.93%	6.30%	1.66%
D. melanogaster						
AUGUSTUS all	73.74%	77.11%	24.47%	39.36%	48.53%	44.03%
AUGUSTUS no RNA	64.97%	70.15%	16.60%	34.09%	33.18%	34.09%
Exonerate all	62.39%	77.96%	17.31%	28.01%	33.88%	28.01%
Exonerate high	60.48%	81.96%	16.34%	39.36%	32.63%	39.36%
mGene	70.73%	81.30%	22.00%	44.02%	43.99%	44.02%
mGene graph	62.54%	82.58%	19.34%	41.26%	38.43%	47.14%
mTim	35.55%	82.59%	8.90%	34.06%	17.66%	40.08%
NextGeneid	59.77%	76.11%	18.69%	38.84%	37.37%	38.84%
NextGeneidAS	61.43%	73.08%	19.20%	32.29%	37.92%	32.29%
NextGeneidAS ab-initio	61.44%	72.99%	19.24%	32.21%	37.99%	32.21%
Transomics all	73.62%	66.12%	23.48%	33.54%	46.95%	33.54%
Transomics high	73.56%	71.22%	23.48%	37.72%	46.93%	37.72%
Tromer	13.85%	18.64%	3.26%	2.81%	6.46%	5.75%
C. elegans						
AUGUSTUS all	84.38%	72.19%	48.20%	36.02%	60.15%	38.76%
AUGUSTUS high	84.21%	79.22%	48.13%	42.44%	60.06%	45.98%
AUGUSTUS no RNA	78.94%	64.66%	36.28%	27.20%	45.52%	27.20%
Exonerate all	67.67%	82.06%	34.18%	32.66%	42.46%	32.66%
Exonerate high	66.55%	84.82%	32.55%	41.25%	40.84%	41.25%
mGene	83.62%	74.15%	45.48%	41.94%	57.05%	41.94%
mGene graph	72.62%	77.61%	45.02%	45.46%	56.38%	47.01%
mTim	45.34%	85.51%	20.24%	36.66%	25.24%	43.02%
NextGeneid	70.15%	81.18%	30.28%	39.80%	37.98%	39.80%
NextGeneidAS	69.78%	79.51%	30.39%	33.37%	37.96%	33.37%
NextGeneidAS ab-initio	69.78%	79.43%	30.39%	33.26%	37.97%	33.26%
Transomics all	86.37%	65.43%	48.30%	32.82%	60.56%	32.82%
Transomics high	86.10%	74.75%	48.28%	40.33%	60.54%	40.33%
Tromer	20.85%	26.55%	1.20%	0.51%	1.50%	1.11%

Nature Methods: doi:10.1038/nmeth.2714

Supplementary Table 4. Exon-, transcript- and gene-level performance (fixed evaluation mode)

	Exon		Transcript		Gene	
	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
H. sapiens						
AUGUSTUS all	49.07%	64.29%	0.00%	0.00%	0.01%	0.01%
AUGUSTUS high	49.00%	69.85%	0.00%	0.01%	0.01%	0.01%
AUGUSTUS no RNA	41.27%	42.41%	0.00%	0.00%	0.00%	0.00%
Cufflinks	42.63%	62.90%	0.00%	0.00%	0.01%	0.00%
Exonerate all	43.32%	71.18%	0.00%	0.00%	0.01%	0.00%
Exonerate high	42.75%	75.85%	0.00%	0.01%	0.01%	0.01%
GSTRUCT	41.58%	79.67%	0.00%	0.00%	0.00%	0.00%
iReckon full	48.48%	68.40%	0.06%	0.06%	0.24%	0.06%
iReckon ends	50.26%	64.32%	0.05%	0.03%	0.18%	0.03%
mGene	47.14%	41.63%	0.00%	0.00%	0.01%	0.00%
mGene graph	46.52%	59.98%	0.00%	0.01%	0.02%	0.01%
mTim	38.28%	73.72%	0.00%	0.00%	0.01%	0.01%
NextGeneid	39.79%	73.26%	0.01%	0.12%	0.02%	0.12%
NextGeneidAS	38.94%	69.02%	0.00%	0.09%	0.02%	0.09%
NextGeneidAS ab-initio	38.97%	67.58%	0.00%	0.14%	0.02%	0.14%
Oases	36.00%	29.91%	0.00%	0.00%	0.01%	0.00%
SLIDE all	52.62%	79.66%	3.13%	4.75%	12.03%	18.18%
SLIDE high	38.56%	84.06%	3.82%	11.53%	14.75%	18.76%
Transomics all	48.52%	45.72%	0.05%	0.63%	0.19%	0.63%
Transomics high	48.01%	62.85%	0.05%	1.02%	0.19%	1.02%
Trembly all	43.14%	65.35%	0.01%	0.01%	0.02%	0.01%
Trembly high	38.76%	74.96%	0.00%	0.00%	0.01%	0.01%
Tromer	23.87%	21.04%	0.00%	0.00%	0.00%	0.00%
Velvet	36.28%	46.45%	0.00%	0.00%	0.00%	0.00%
Velvet + AUGUSTUS	22.57%	73.75%	0.00%	0.00%	0.00%	0.00%
D. melanogaster						
AUGUSTUS all	45.45%	51.16%	0.01%	0.02%	0.03%	0.02%
AUGUSTUS no RNA	40.06%	47.22%	0.00%	0.01%	0.01%	0.01%
Cufflinks	39.06%	49.65%	0.00%	0.01%	0.01%	0.01%
Exonerate all	40.32%	51.70%	0.02%	0.03%	0.04%	0.03%
Exonerate high	39.52%	57.66%	0.02%	0.04%	0.03%	0.04%
GSTRUCT	33.34%	59.03%	0.00%	0.01%	0.01%	0.01%
iReckon full	61.82%	61.71%	0.03%	0.03%	0.05%	0.03%
iReckon ends	58.56%	59.42%	0.02%	0.02%	0.04%	0.02%
mGene	42.61%	55.60%	0.03%	0.05%	0.05%	0.05%
mGene graph	40.36%	54.37%	0.03%	0.05%	0.05%	0.05%
mTim	33.87%	55.67%	0.01%	0.02%	0.02%	0.03%
NextGeneid	37.49%	51.53%	0.01%	0.02%	0.03%	0.02%
NextGeneidAS	38.88%	47.37%	0.02%	0.02%	0.03%	0.02%
NextGeneidAS ab-initio	38.88%	47.33%	0.03%	0.03%	0.04%	0.03%
Oases	33.61%	28.54%	0.01%	0.01%	0.03%	0.01%
SLIDE all	84.78%	94.14%	44.19%	25.52%	76.99%	70.36%
SLIDE high	77.01%	95.78%	44.75%	50.81%	82.73%	75.38%
Transomics all	41.16%	44.40%	0.44%	0.65%	0.88%	0.65%
Transomics high	41.09%	47.77%	0.44%	0.73%	0.88%	0.73%
Tromer	10.46%	10.56%	0.00%	0.00%	0.00%	0.00%
Velvet	30.27%	32.91%	0.02%	0.02%	0.03%	0.02%
C. elegans						
AUGUSTUS all	62.92%	53.19%	0.00%	0.00%	0.00%	0.00%
AUGUSTUS high	62.80%	58.34%	0.00%	0.00%	0.00%	0.00%
AUGUSTUS no RNA	59.53%	47.94%	0.00%	0.00%	0.00%	0.00%
Cufflinks	53.55%	59.41%	0.00%	0.00%	0.00%	0.00%
Exonerate all	53.51%	60.84%	0.00%	0.00%	0.00%	0.00%
Exonerate high	53.01%	63.98%	0.00%	0.00%	0.00%	0.00%
GSTRUCT	49.19%	64.98%	0.00%	0.00%	0.00%	0.00%
iReckon full	78.70%	71.81%	0.00%	0.00%	0.00%	0.00%
iReckon ends	71.85%	69.63%	0.00%	0.00%	0.00%	0.00%
mGene	62.93%	55.35%	0.00%	0.00%	0.00%	0.00%
mGene graph	56.22%	56.33%	0.00%	0.00%	0.00%	0.00%
mTim	52.80%	62.77%	0.00%	0.00%	0.00%	0.00%
NextGeneid	59.65%	67.18%	4.16%	5.03%	5.22%	5.03%
NextGeneidAS	59.67%	64.82%	4.58%	4.21%	5.66%	4.21%
NextGeneidAS ab-initio	59.74%	64.92%	4.63%	4.28%	5.73%	4.28%
Oases	45.33%	42.17%	0.00%	0.00%	0.00%	0.00%
SLIDE all	87.50%	97.24%	40.85%	12.66%	50.26%	43.98%
SLIDE high	81.65%	97.94%	56.51%	50.52%	70.19%	61.42%
Transomics all	88.87%	67.32%	54.36%	36.94%	68.16%	36.94%
Transomics high	88.45%	76.78%	54.35%	45.40%	68.14%	45.40%
Tromer	22.72%	19.22%	0.00%	0.00%	0.00%	0.00%
Velvet	45.87%	50.48%	0.00%	0.00%	0.00%	0.00%
Velvet + AUGUSTUS	58.49%	67.24%	0.00%	0.00%	0.00%	0.00%

Nature Methods: doi:10.1038/nmeth.2714

Supplementary Table 5. Exon-, transcript- and gene-level performance (flexible evalutation mode)

	Exon		Transcript		Gene	
	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
H. sapiens						
AUGUSTUS all	81.16%	77.47%	16.27%	39.26%	56.16%	42.71%
AUGUSTUS high	81.05%	84.15%	16.27%	44.41%	56.16%	49.78%
AUGUSTUS no RNA	65.55%	50.73%	3.52%	8.03%	13.38%	8.03%
Cufflinks	73.45%	79.57%	16.03%	19.29%	53.55%	20.36%
Exonerate all	74.59%	88.34%	17.22%	28.60%	55.53%	28.60%
Exonerate high	72.72%	92.46%	13.23%	41.92%	50.28%	41.92%
GSTRUCT	70.63%	96.42%	14.10%	59.85%	54.02%	59.85%
iReckon full	67.78%	83.05%	31.80%	38.87%	73.63%	38.87%
iReckon ends	73.24%	83.26%	27.78%	35.77%	71.61%	35.77%
mGene	78.40%	50.96%	10.83%	8.98%	41.35%	9.00%
mGene graph	77.27%	73.14%	14.90%	14.89%	48.17%	21.42%
mTim	64.86%	90.42%	10.14%	22.66%	35.78%	36.95%
NextGeneid	67.36%	89.58%	11.30%	36.99%	43.27%	36.99%
NextGeneidAS	66.65%	87.32%	11.96%	27.73%	42.73%	27.73%
NextGeneidAS ab-initio	66.68%	85.56%	11.94%	26.60%	42.63%	26.60%
Oases	62.02%	40.55%	10.04%	5.11%	34.05%	4.74%
SLIDE all	72.58%	84.10%	8.78%	12.68%	29.08%	34.60%
SLIDE high	54.75%	88.22%	8.65%	22.66%	31.09%	31.62%
Transomics all	76.84%	54.35%	8.41%	13.28%	32.16%	13.28%
Transomics high	76.14%	74.64%	8.41%	21.74%	32.16%	21.74%
Trembly all	75.45%	83.44%	21.82%	22.73%	60.53%	38.37%
Trembly high	64.30%	90.24%	12.84%	30.28%	43.60%	41.80%
Tromer	49.02%	37.22%	4.27%	3.70%	13.92%	6.48%
Velvet	62.05%	59.56%	5.57%	6.81%	21.35%	7.26%
Velvet + AUGUSTUS	38.97%	90.68%	7.17%	45.22%	27.44%	45.35%
D. melanogaster						
AUGUSTUS all	82.61%	86.03%	42.66%	61.14%	73.03%	65.92%
AUGUSTUS no RNA	71.12%	77.83%	23.30%	42.22%	41.21%	42.22%
Cufflinks	71.73%	83.70%	36.18%	46.32%	56.37%	50.06%
Exonerate all	73.32%	86.35%	38.51%	51.44%	58.83%	51.44%
Exonerate high	69.54%	92.50%	30.98%	66.02%	54.54%	66.02%
GSTRUCT	59.66%	95.60%	28.21%	76.77%	49.10%	76.77%
iReckon full	85.85%	82.43%	65.22%	57.94%	89.07%	57.94%
iReckon ends	82.93%	80.44%	57.05%	51.13%	83.09%	51.13%
mGene	75.49%	91.06%	38.02%	67.02%	67.17%	67.05%
mGene graph	73.03%	90.92%	40.00%	64.82%	67.78%	70.56%
mTim	62.35%	90.40%	20.59%	40.70%	34.09%	53.27%
NextGeneid	70.53%	89.19%	36.30%	62.23%	64.27%	62.23%
NextGeneidAS	73.97%	83.76%	40.05%	49.64%	66.30%	49.64%
NextGeneidAS ab-initio	73.95%	83.66%	40.05%	49.61%	66.25%	49.61%
Oases	64.31%	51.05%	33.99%	18.61%	55.47%	25.39%
SLIDE all	89.08%	95.36%	52.80%	31.26%	82.02%	74.88%
SLIDE high	82.31%	96.83%	53.22%	58.60%	88.50%	80.58%
Transomics all	73.74%	74.03%	35.11%	44.78%	62.71%	44.78%
Transomics high	73.63%	79.67%	35.11%	50.38%	62.69%	50.38%
Tromer	29.15%	27.41%	9.37%	4.79%	15.69%	10.47%
Velvet	58.59%	58.75%	24.09%	22.63%	42.39%	24.81%
C. elegans						
AUGUSTUS all	91.76%	76.26%	59.28%	44.09%	72.35%	46.62%
AUGUSTUS high	91.48%	83.55%	59.21%	51.96%	72.26%	55.33%
AUGUSTUS no RNA	86.51%	68.56%	42.25%	31.53%	52.77%	31.53%
Cufflinks	78.38%	84.84%	39.75%	37.53%	48.04%	42.88%
Exonerate all	76.99%	85.55%	43.82%	40.02%	52.90%	40.02%
Exonerate high	75.75%	89.27%	39.98%	50.42%	49.83%	50.42%
GSTRUCT	71.48%	91.85%	39.03%	59.92%	48.72%	59.92%
iReckon full	94.09%	84.88%	78.10%	52.98%	89.87%	52.98%
iReckon ends	87.39%	83.71%	68.45%	44.85%	79.25%	44.85%
mGene	90.99%	78.64%	55.27%	50.73%	69.01%	50.73%
mGene graph	83.26%	81.96%	55.78%	54.68%	68.88%	56.18%
mTim	77.16%	88.48%	31.86%	37.72%	38.96%	46.54%
NextGeneid	78.45%	86.60%	37.04%	44.55%	46.19%	44.55%
NextGeneidAS	77.91%	82.80%	38.45%	35.17%	46.71%	35.17%
NextGeneidAS ab-initio	77.85%	82.76%	38.45%	35.41%	46.71%	35.41%
Oases	66.77%	61.04%	32.98%	17.39%	39.70%	22.24%
SLIDE all	88.52%	97.34%	42.85%	13.34%	51.92%	45.44%
SLIDE high	82.97%	98.04%	58.13%	51.78%	71.61%	62.66%
Transomics all	93.90%	69.92%	58.10%	39.32%	72.52%	39.32%
Transomics high	93.43%	79.73%	58.08%	48.32%	72.50%	48.32%
Tromer	39.43%	32.92%	7.24%	1.96%	8.97%	4.96%
Velvet	67.44%	72.86%	31.77%	27.15%	39.54%	28.09%
Velvet + AUGUSTUS	61.49%	69.46%	0.00%	0.00%	0.00%	0.00%

Nature Methods: doi:10.1038/nmeth.2714

Supplementary Table 6. Alternative splicing and transcript diversity

	Minimum	First quartile	Median	Mean	Third quartile	Maximum
H, sapiens						
CDS length (bp)	1	79	115	143.7	159	17330
Exon length (bp)	1	87	126	224.9	186	91670
Intron length (bp)	3	498	1569	6410	4481	4251000
Exons per transcript	1	3	5	6.646	8	118
Transcripts per gene	1	1	1	5.225	7	80
D. melanogaster						
CDS length (bp)	1	124	197	372.6	402	27710
Exon length (bp)	1	144	246	476.2	544	28070
Intron length (bp)	4	65	104	1540	733	139300
Exons per transcript	1	2	4	5.496	7	78
Transcripts per gene	1	1	1	1.943	2	31
C. elegans						
CDS length (bp)	1	99	146	205.3	234	14980
Exon length (bp)	1	99	146	205.3	234	14980
Intron length (bp)	3	50	71	330.6	345	21230
Exons per transcript	1	4	6	6.819	9	66
Transcripts per gene	1	1	1	1.255	1	15

Supplementary Table 7. NanoString probes and targeted transcript isoforms

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
adar1_sp1	ENST00000292205, ENST00000494866, ENST00000368471	ACTGGCAGTCTCCGGGTGTCCGGCCGTGTCCCGAGGAAGTGCAAGACCCGGGGTATTCC CTCAGCGGATACTACACCCATCCATTTCAAGGCTATGAGCA
adar1_sp2	ENST00000368474	CGGCGGGTCGGGCCGGGCAATGCCTCGCGGGCGCAATGAATCCGCGGCAGGGGTATTC CCTCAGCGGATACTACACCCATCCATTTCAAGGCTATGAGCA
atf2_common	ENST00000264110, ENST00000345739, ENST00000392543, ENST00000392544, ENST00000409499, ENST00000409833, ENST00000413123, ENST00000415955, ENST00000417080, ENST00000421438, ENST00000426833, ENST00000428760, ENST00000429579, ENST00000435231, ENST00000437522, ENST00000445349, ENST00000456655, ENST00000487334, ENST00000538946, ENST00000542046, ENST00000409635	GGTACTAGATGGAAACTTGAGAAAGGACTGCTTATTGATAACAGCTAAGGTATTCCTGG AAGCAGAGTAAATAAAGCTCATGGCCCACCAGCTAGAAAG
atf2_sp1	ENST00000345739, ENST00000437522, ENST00000435231, ENST00000415955, ENST00000409635, ENST00000456655	ATATGAGTGATGACAAACCCTTTCTATGTACTGCGCCTGGATGTGGCCAGATCAGACCCC AACACCAACAAGATTCTTGAAAAACTGTGAAGAAGTGGGT
atf2_sp2	ENST00000429579, ENST00000538946, ENST00000428760, ENST00000417080, ENST00000421438, ENST00000409437, ENST00000392544, ENST00000264110, ENST00000435004, ENST00000542046, ENST00000426833, ENST00000409833, ENST00000487334	ATATGAGTGATGACAAACCCTTTCTATGTACTGCGCCTGGATGTGGCCAGCGTTTTACCA ACGAGGATCATTTGGCTGTCCATAAACATAAACATGAGAT
ATP5J_common	ENST00000284971, ENST00000400087, ENST00000400090, ENST00000400093, ENST00000400094, ENST00000457143, ENST00000486002, ENST00000400099	CAGAGTATCAGCAAGAGCTGGAGAGGGAGCTTTTTAAGCTCAAGCAAATGTTTGGTAAT GCAGACATGAATACATTTCCCACCTTCAAATTTGAA
Bcl11a_sp1	ENST00000335712, ENST00000356842, ENST00000359629, ENST00000489516	TTTATCAACGTCATCTAGAGGAATTTGCCCCAAACAGGAACACATAGCAGATAAACTTCT GCACTGGAGGGGCCTCTCCTCCCCTCGTTCTGCACATGGA
Bcl11a_sp2	ENST00000409351	TATCAACGTCATCTAGAGGAATTTGCCCCAAACAGGAACACATAGCAGCTCAGACTGAAC TGGAGGATGTATTTGTGTACCTTATGGTGT
BCL3	ENST00000164227, ENST00000403534	CGGAGCCTTACTGCCTTTGTACCCCACTCGGGCCATGGGCTCCCCGTTTCCTCTGGTGAAC CTGCCTACACCCCTATACCCCATGATGTGCCCCATGGAA
BHLHB2	ENST00000256495	AAAAGCTTCAAAGTCTTGGTCTGTGAGTCACTCTTCAGTTTGGGAGCTGGGTCTGTGGCT TTGATCAGAAGGTACTTTCAAAAGAGGGCTTTCCAGGGCT
Blnk_sp1_T	ENST00000371176	TGAAAACTATATTCATCCCACAGAAAGCAGTTCACCTCCACCTGAAAAAGGTCGAAACAG TGGGGCCTGGGAAACCAAGTCACCTCCACCAGCTGCACCA
Blnk_sp2	ENST00000413476, ENST00000224337, ENST00000427367, ENST00000467799	TGAAAACTATATTCATCCCACAGAAAGCAGTTCACCTCCACCTGAAAAAGCTCCCATGGT GAATAGATCAACCAAGCCAAATTCCTCAACGCCCGCCTCT
CARM1_sp1_T	ENST00000592516, ENST00000344150	TGTTATTGCCAGTGGCTCCAGCGTGGGCCACAACAACCTGATTCCTTTAGGGTCCTCCGG CGCCCAGGGCAGTGGTGGTGGCAGCACGAGTGCCCACTAT
CD19	ENST00000324662, ENST00000538922, ENST00000565089, ENST00000567541	GAAGGTCTCAGCTGTGACTTTGGCTTATCTGATCTTCTGCCTGTGTTCCCTTGTGGGCATT CTTCATCTTCAAAGAGCCCTGGTCCTGAGGAGGAAAAGA
Cd79b_sp1_T	ENST00000349817	TGAGCCAGTACCAGCAGCCAGATCGGAGGACCGGTACCGGAATCCCAAAGGATTCAGCA CCTTGGCACAGCTGAAGCAGAGGAACACGCTGAAGGATGGT
Cd79b_sp2_T	ENST00000559358, ENST00000006750, ENST00000392795	CAACACCTCGGAGGTCTACCAGGGCTGCGGCACAGAGCTGCGAGTCATGGGATTCAGCA CCTTGGCACAGCTGAAGCAGAGGAACACGCTGAAGGATGGT
cdkn1a_sp1_T	ENST00000244741	GAGCCGGAGCTGGGCGCGGATTCGCCGAGGCACCGAGGCACTCAGAGGAGGCGCCATG TCAGAACCGGCTGGGGATGTCCGTCAGAACCCATGCGGCAGC
cdkn1a_sp2_T	ENST00000405375, ENST00000478800	GGATGCGTGTTCGCGGGTGTGTGCTGCGTTCACAGGTGTTTCTGCGGCAGGCGCCATGT CAGAACCGGCTGGGGATGTCCGTCAGAACCCATGCGGCAGC
CEBPA	ENST00000425420, ENST00000498907	CTAGTATTTAGGATAACCTTGTGCCTTGGAAATGCAAACTCACCGCTCCAATGCCTACTGA GTAGGGGGAGCAAATCGTGCCTTGTCATTTTATTTGGAG
CTCF_common	ENST00000264010, ENST00000401394	CCCAACGGAGACCTCACGCCCGAGATGATCCTCAGCATGATGGACCGGTGATGGCGGAG CCTTGTGCGTCGCCAGGACTTCTCTGGGCTGTGTTTAAACG
CTCF_sp1	ENST00000566078, ENST00000264010	GAGCTGGGTTCTATTTTCCCTCCTCAAACTGACTTTGCAGCCACGGAGAGGCAGGGGAAA TGGAAGGTGATGCAGTCGAAGCCATTGTGGAGGAGTCCGA
CTCFL_sp1	ENST00000433949, ENST00000243914, ENST00000539382, ENST00000371196, ENST00000426658, ENST00000423479, ENST00000422869, ENST00000502686	TGCCAGCAGAGATACCTACAAGCTGAAACGCCACATGAGAACGCACTCAGGTGAGAAGC CTTACGAATGCCACATCTGCCACACCCGCTTCACCCAGAGC
CTCFL_sp2	ENST00000429804	TGCCAGCAGAGATACCTACAAGCTGAAACGCCACATGAGAACGCACTCAGGTGTGCATA TGCGCAACTTGCATGCTTACAGCGCTGCAGAGCTGAAATGC
CTDSL_sp2_T	ENST00000273179, ENST00000443503, ENST00000486978	GCCCCAGTGTGCTTCCGCCACTGGTGGAGGAGAATGGTGGGCTTCAGAAGGGTGACCA GAGGCAGGTCATTCCCATACCAAGTCCACCAGCTAAGTACCT
CTDSP1_common_T	ENST00000273062, ENST00000428361, ENST00000443891, ENST00000452977, ENST00000464255, ENST00000473420, ENST00000482272, ENST00000488627, ENST00000491064, ENST00000497677, ENST00000498160	AGCTGACCTGCTGGACAAATGGGGGGCCTTCCGGGCCCGGCTGTTTCGAGAGTCCTGCG TCTTCCACCGGGGGAACTACGTGAAGGACCTGAGCCGGTT
CTDSP2	ENST00000398073	CCTGTCCTGTACCGAGCTCTGTCTGTTCCAGCCTTCATCCTTCCTGGCTGTTGCTTTTCCTC TTAAGGGCCTCAGAACTCTTGCTCTTCCTGGGCTGAGG
CTDSPL_sp1	ENST00000443503	GCCCCAGTGTGCTTCCGCCACTGGTGGAGGAGAATGGTGGGCTTCAGAAGCCACCAGCT AAGTACCTTCTTCCAGAGGTGACGGTGCTTGACTATGGAAA
DES	ENST00000492726, ENST00000477226, ENST00000373960	GAGAACAATTTGGCTGCCTTCCGAGCGGACGTGGATGCAGCTACTCTAGCTCGCATTGAC CTGGAGCGCAGAATTGAATCTCTCAACGAGGAGATCGCGT
DNMT1_common_T	ENST00000340748, ENST00000359526, ENST00000540357, ENST00000586588, ENST00000587197, ENST00000588913, ENST00000589294, ENST00000592705	CCTTTCCGGACCATCACGGTGCGAGACACGATGTCCGACCTGCCGGAGGTGCGGAATGG AGCCTCGGCACTGGAGATCTCCTACAACGGGGAGCCTCAGT
E2F4	ENST00000379378, ENST00000567007	GTCAGAAATCTTTGATCCCACACGAGAGTGCATGAGCTCGGAGCTGCTGGAGGAGTTGA TGTCCTCAGAAGTGTTTGCCCCTCTGCTTCGTCTTTCTCCA
E2F6_common	ENST00000307236, ENST00000362009, ENST00000381525, ENST00000421117, ENST00000428221, ENST00000437573, ENST00000444832, ENST00000455198, ENST00000468775, ENST00000542100, ENST00000546212	AGGTTGCAACGAAACTGGGAGTCCGAAAGCGGAGAGTGTATGACATCACCAATGTCTTA GATGGAATCGACCTCGTTGAAAAGAAATCCAAGAACCATAT
ebf1_sp1	ENST00000519890, ENST00000380654, ENST00000518836, ENST00000519739, ENST00000313708, ENST00000522192	TGGACAACTGGCCGTGAATGTCTCCGAGGCATCACAAGCCACCAATCAGGGTTTCACCCG CAACTCAAGCAGCGTATCACCACACGGGTACGTGCCGAGC
EGR1	ENST00000239938	CTTCAATGCTAGAAAATCGAGTTGGCAAAATGGGGTTTGGGCCCCTCAGAGCCCTGCCCT GCACCCTTGTACAGTGTCTGTGCCATGGATTTCGTTTTTC
EOMES_common	ENST00000295743, ENST00000449599, ENST00000537516	CAACAAACTAGACATCAGTTCCTATGAATCTGAATATACTTCTAGCACATTGCTCCCATAT GGCATTAAATCCTTGCCCCTTCAGACATCCCATGCCCTG
EP300	ENST00000263253	ACAAATATCCCTTTGGCTCCGTCCAGCGGTCAAGCTCCAGTGTCTCAAGCACAAATGTCTA GTTCTTCCTGCCCGGTGAACTCTCCTATAATGCCTCCAG
esr1_common	ENST00000206249, ENST00000338799, ENST00000406599, ENST00000427531, ENST00000440973, ENST00000443427, ENST00000456483, ENST00000544394	GTAGAGGGCATGGTGGAGATCTTCGACATGCTGCTGGCTACATCATCTCGGTTCCGCATG ATGAATCTGCAGGGAGAGGAGTTTGTGTGCCTCAAATCTA
esr2_sp1	ENST00000353772, ENST00000554572, ENST00000344288, ENST00000358599	ATGCGCCTGGCTAACCTCCTGATGCTCCTGTCCCACGTCAGGCATGCGAGGGCAGAAAA GGCCTCTCAAACACTCACCTCATTTGGAATGAAGATGGAGA
esr2_sp2	ENST00000554520, ENST00000341099, ENST00000267525, ENST00000555483	ATGCGCCTGGCTAACCTCCTGATGCTCCTGTCCCACGTCAGGCATGCGAGTAACAAGGGC ATGGAACATCTGCTCAACATGAAGTGCAAAAATGTGGTCC
ets1_common	ENST00000319397, ENST00000345075, ENST00000526145, ENST00000530924, ENST00000531611, ENST00000535549, ENST00000392668	CAGGAGATGGGGAAAGAGGAAAAACAAACCTAAGATGAATTATGAGAAACTGAGCCGT GGCCTACGCTACTATTACGACAAAAACATCATCCACAAGACA
FBXO15	ENST00000269500, ENST00000581214, ENST00000419743, ENST00000585174, ENST00000583443	CTAGCTGACATTCTCAAACCTGTCAACCCTTACACAGGCCTTCCAGTTAAGACCAAAGAG GCCCTCAGAATATTTGGTTTAGGTTGGGCAATTATACTGA
FOS	ENST00000303562	TCAAGTCCTTACCTCTTCCGGAGATGTAGCAAAACGCATGGAGTGTGTATTGTTCCCAGT GACACTTCAGAGAGCTGGTAGTTAGTAGCATGTTGAGCCA
foxa2_body	ENST00000319993, ENST00000377115, ENST00000419308	CGTTCCGTCCCAAACAGAGGGCCACACAGATACCCCACGTTCTATATAAGGAGGAAAAC GGGAAAGAATATAAAGTTAAAAAAAAGCCTCCGGTTTCCAC

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
foxa2_sp2_T	ENST00000377115	CGGGTCCCTGGCGGCCGGTGTCTGAGGAGTCGGAGAGCCGAGGCGGCCAGACCGTGCG CCCCGCGCTTCTCCCGAGGCCGTTCCGGGTCTGAACTGTAAC
FOXA3	ENST00000302177	CCCCGTGTTGGCCATGTCGTCACCATTCTCTCTGGCATGGGTTGGGTAGGGGATGGAGG TGAGAATACTCCTTGGTTTTCTCTGAAGCCCACCCTTTCCC
gabpa_sp1_T	ENST00000354828	CCGGACGGGTCTAGGTGAGACAGAAGCCAAACAGGAGGAGGAAGTGGAGGGACTGAT CCTTTGAAATACTCCAGCCATGACTAAAAGAGAAGCAGAGGAG
gabpa_sp2_T	ENST00000400075	CAGCCGGCTCTGGAGTGCGGGCGGGGGCGACAGGGCCGATTCCGGAGTGGGACTGATC CTTTGAAATACTCCAGCCATGACTAAAAGAGAAGCAGAGGAG
GATA1_T	ENST00000376665, ENST00000376670	GTGTCCCACCCGCGAGGACTCTCCTCCCCAGGCCGTGGAAGATCTGGATGGAAAAGGCA GCACCAGCTTCCTGGAGACTTTGAAGACAGAGCGGCTGAGC
HDAC1	ENST00000373548, ENST00000476391	CTGTTTTCGTACCTTCCCACTGGCCTCAAGTGAGCCAAGAAACACTGCCTGCCCTCTGTCT GTCTTCTCCTAATTCTGCAGGTGGAGGTTGCTAGTCTAG
HDAC3_sp1	ENST00000305264, ENST00000495485, ENST00000523353	CATTGACCCATAGCCTGGTCCTGCATTACGGTCTCTATAAGAAGATGATCGTCTTCAAGCC ATACCAGGCCTCCCAACATGACATGTGCCGCTTCCACTC
HDAC4	ENST00000345617	TGTCAGCTCACTCCAGCTTCACAAATGTGCTGAGAGCATTACTGTGTAGCCTTTTCTTTGA AGACACACTCGGCTCTTCTCCACAGCAAGCGTCCAGGGC
HDAC5	ENST00000225983, ENST00000336057, ENST00000393622	CAGGGGAGGATCTGGAGGATCCACTACTGTCTTTAAGATGCAGAGTGGAGGGGAGGTG GGCACCCACCCTGCGATTCTCCACCCTTTCCCCTTCTTTCGT
Hif1a_common	ENST00000323441, ENST00000337138, ENST00000394997, ENST00000555014, ENST00000539097, ENST00000557538	TCCAGCAGACTCAAATACAAGAACCTACTGCTAATGCCACCACTACCACTGCCACCACTG ATGAATTAAAAACAGTGACAAAAGACCGTATGGAAGACAT
HNF1A	ENST00000257555, ENST00000400024, ENST00000402929, ENST00000538646, ENST00000540108, ENST00000541395, ENST00000541924, ENST00000543427, ENST00000544413	GTGCGCTATGGACAGCCTGCGACCAGTGAGACTGCAGAAGTACCCTCAAGCAGCGGCG GTCCCTTAGTGACAGTGTCTACACCCCTCCACCAAGTGTCCC
HNF1B	ENST00000561193, ENST00000225893	GTTTCCATCTGCAATGGTGGTCACAGATACCAGCAGCATCAGTACACTCACCAACATGTC TTCAAGTAAACAGTGTCCTCTACAAGCCTGGTGATGCCCA
Hnf4g_common_T	ENST00000396423	AGAAAATAGTTATCCATTGACTAGAAATTAGTACATGCCCACAGCTGGCTCCCACGGTAG CCAGGAGAATTATCTATAGGTGGAAAGTCTGTGTCAGCCA
HSF1	ENST00000400780, ENST00000528838, ENST00000528988, ENST00000533240	TGTTCGACCAGGGCCAGTTTGCCAAGGAGGTGCTGCCCAAGTACTTCAAGCACAACAAC ATGGCCAGCTTCGTGCGGCAGCTCAACATGTATGGCTTCCG
IGF1R_common_T	ENST00000268035, ENST00000558762	GCGATTGCTGGGTGTGGTGTCCCAAGGCCAGCCAACACTGGTCATCATGGAACTGATGA CACGGGGCGATCTCAAAAGTTATCTCCGGTCTCTGAGGCCA
IKZF1_common	ENST00000331340, ENST00000343574, ENST00000346667, ENST00000349824, ENST00000357364, ENST00000359197, ENST00000438033, ENST00000439701, ENST00000440768, ENST00000471793	CCTGCTGCGCGCCGCCTCCGAGAACTCGCAGGACGCGCTCCGCGTGGTCAGCACCAGCG GGGAGCAGATGAAGGTGTACAAGTGCGAACACTGCCGGGTG
IKZF1_sp1_T	ENST00000357364, ENST00000331340, ENST00000343574, ENST00000413698, ENST00000346667, ENST00000349824, ENST00000440768, ENST00000359197	CGAGGATCAGTCTTGGCCCCAAAGCGCGACGCACAAATCCACATAACCTGAGGACCATG GATGCTGATGAGGGTCAAGACATGTCCCAAGTT
IKZF1_sp2	ENST00000438033, ENST00000492782, ENST00000462201, ENST00000439701	GTGTGGAAAAGGCAGCTCTCACTTGGCCTTGGCGAGGCCTCGGTTGGTTGATAACCTGA GGACCATGGATGCTGATGAGGGTCAAGACATGTCCCAAGTT
IKZF3_sp1_T	ENST00000377958, ENST00000377944, ENST00000346872, ENST00000535189, ENST00000467757	CAAGGAGCGCTGCCGTACATTTCTTCAGAGCACTGACCCAGGGGACACTGCAAGTGCGG AGGCAAGACACATCAAAGCAGAGATGGGAAGTGAAAGAGCT
IKZF3_sp2	ENST00000583368, ENST00000293068, ENST00000348427	TCACTGACCACAGCAGGTACCCCAGGCAAGAATCTGAGCAGTTATAACAGCAAGTGCGG AGGCAAGACACATCAAAGCAGAGATGGGAAGTGAAAGAGCT
IL6	ENST00000258743, ENST00000401630, ENST00000404625, ENST00000407492, ENST00000485300	GCTCTTCGGCAAATGTAGCATGGGCACCTCAGATTGTTGTTGTTAATGGGCATTCCTTCTT CTGGTCAGAAACCTGTCCACTGGGCACAGAACTTATGTT
IL6receptor_common	ENST00000344086, ENST00000368485	TCCAATATTCGCTGTGTCAGCATAGAAGTAACTTACTTAGGTGTGGGGGAAGCACCATAA CTTTGTTTAGCCCAAAACCAAGTCAAGTGAAAAAGGAGGA
IL8	ENST00000307407, ENST00000483500, ENST00000401931	GGAAGGAACCATCTCACTGTGTGTAAACATGACTTCCAAGCTGGCCGTGGCTCTCTTGGC AGCCTTCCTGATTTCTGCAGCTCTGTGTGAAGGTGCAGTT
IL8RA	ENST00000295683	GTCCATTGGGCAGGCAGATGTTCCTAATAAAGCTTCTGTTCCGTGCTTGTCCCTGTGGAA GTATCTTGGTTGTGACAGAGTCAAGGGTGTGTGCAGCATT
IL8RB	ENST00000318507	GATAGACAAATCTCCACCTTCAGACTGGTAGGCTCCTCCAGAAGCCATCAGACAGGAAG ATGTGAAAATCCCCAGCACTCATCCCAGAATCACTAAGTGG
IRF8	ENST00000268638, ENST00000566369	CCCTCTGTCTGGGGTGGGATGCCTTACTTTGCACTTAATTTAATAAGGGCATTCTCGGAG GAGTAGACGTTTAATACGAAGTGGCGGCATAGCCCTGCCG
JUND	ENST00000252818	TGCTACGAGTCCACATTCCTGTTTGTAATCCTTGGTTCGCCCGGTTTTCTGTTTTCAGTAAA GTCTCGTTACGCCAGCTCGGCAAAAAAAAAAAAAAAAA
KAISO_sp1_T	ENST00000326624	CCAGCCTTCCGCGCGTCCGGAGGAGGAGAAGCGGCGGCGCCGGGAAGCAGGCATGGA GAGTAGAAAACTGATTTCTGCTACAGACATTCAGTACTCTGGC
KAP1_common_T	ENST00000253024, ENST00000341753	CAGGCCGAGTGCAAACAGGGCAGCAGGCGGGGCTCCCTCTCGGGTCGCAGGCGCTCTCT GCACACGCCGCAGTGCTCCAGCAGCTCCAGCGCCTCGGCGC
KLF4	ENST00000374672, ENST00000493306, ENST00000497048	CCGAGCATTTTCCAGGTCGGACCACCTCGCCTTACACATGAAGAGGCATTTTTAAATCCCA GACAGTGGATATGACCCACACTGCCAGAAGAGAATTCAG
LEF1_sp1	ENST00000509428, ENST00000438313, ENST00000505379, ENST00000510624, ENST00000504775, ENST00000510135, ENST00000379951	TATCCCTTGTCTCCGGGTGGTGTTGGACAGATCACCCCACCTCTTGGCTGGTTTTCCCATC ATATGATTCCCGGTCCTCCTGGTCCCCACACAACTGGCA
LEF1_sp2	ENST00000265165, ENST00000506680, ENST00000510717, ENST00000504950, ENST00000515500	TCCCTTGTCTCCGGGTGGTGTTGGACAGATCACCCCACCTCTTGGCTGGCAAGGTCAGCC TGTATATCCCATCACGGGTGGATTCAGGCAACCCTACC
LIN28_common	ENST00000254231, ENST00000326279	GTCTGGAATCCATCCGTGTCACCGGACCTGGTGGAGTATTCTGTATTGGGAGTGAGAGG CGGCCAAAAGGAAAGAGCATGCAGAAGCGCAGATCAAAAGG
MAX_sp1	ENST00000556443, ENST00000358664, ENST00000358402, ENST00000553928, ENST00000555667, ENST00000394606, ENST00000284165, ENST00000553951, ENST00000556979, ENST00000556892, ENST00000557746, ENST00000557277	ACAGCTTTCACAGTTTGCGGGACTCAGTCCCATCACTCCAAGGAGAGAAGGCATCCCGG GCCCAAATCCTAGACAAAGCCACAGAATATATCCAGTATAT
mef2a_common_T	ENST00000354410, ENST00000449277, ENST00000557785, ENST00000557942, ENST00000558812, ENST00000561125, ENST00000338042, ENST00000453228	CCCGCAGCCCCAGCCCCGACAGGAAATGGGGCGCTCCCCTGTGGACAGTCTGAGCAGCT CTAGTAGCTCCTATGATGGCAGTGATCGGGAGGATCCACGG
mef2a_sp3	ENST00000558812, ENST00000557785, ENST00000449277, ENST00000557942, ENST00000453228, ENST00000338042	GATTCAGCAAACTAAATGAAGATAGTGATTTTATTTTCAAACGAGGCCCTCCTGGTCTGC CACCTCAGAACTTTTCAATGTCTGTCACAGTTCCAGTGAC
mef2a_sp4	ENST00000354410	AAAAAATTAATGAGGAATTTGATAATATGATGCGGAATCATAAAATCGCACCTGGTCTGC CACCTCAGAACTTTTCAATGTCTGTCACAGTTCCAGTGAC
mef2b_sp10_T	ENST00000477565	CCCCACTGCCACTCCCAGCTGCAAGGACCGTCTCTCAGCTGCGCTGGGAACCGCTGCTTC TCGCTTATTAGAAAACTGTCTCTTTCCTTTTGTCCTGGT
mef2b_sp11_T	ENST00000585679, ENST00000514819, ENST00000462498, ENST00000444486	GTCGCTATGGAGGAGCCGGAGATGCAGCTCAAGGGGAAGAAAGCGCCGTGAAGAACCT GGTGGACAGCAGCGTCTACTTCCGCAGCGTGGAGG
mef2b_sp12	ENST00000591398, ENST00000162023, ENST00000494489, ENST00000462790, ENST00000588208, ENST00000488252, ENST00000354191, ENST00000477565	TGGGAGGAGCAGAGCCAGGGAGCCATCTACACTGTGGAGTACGCCTGCAGCGCCGTGA AGAACCTGGTGGACAGCAGCGTCTACTTCCGCAGCGTGGAGG
mef2b_sp7_T	ENST00000410050, ENST00000424583, ENST00000409224	CAGCCGCCGCGGGTCCGTGCGCCCAGCGTCCCAGGGCCCAGGCCGAGCAGACAAAGAT CATTCCACTCAGCCTGGGACGATGGGGAGGAAAAAAATCCAG
mef2c_sp1	ENST00000508569, ENST00000514015, ENST00000510942, ENST00000514028, ENST00000437473, ENST00000504921, ENST00000503554, ENST00000506554	GGAAAATTAACGAAGATATTGATCTAATGATCAGCAGGCAAAGATTGTGTGCTGTTCCAC CTCCCAACTTCGAGATGCCAGTCTCCATCCCAGTGTCCAG
mef2c_sp3	ENST00000424173, ENST00000340208	GAAGTAAAGAACGGAAGGCAAATGATTGTGGCAGTAAAGAAGTGTATGTGCAGGAACG AATGCAGGAATTTGGGAACTGAGCTGTGCAAGTGCTGAAGAA
mef2c_sp4	ENST00000514015, ENST00000510942, ENST00000514028, ENST00000504921, ENST00000513252, ENST00000506554	TTAAGAAAGGAAAATATCCCAAGGACTAATCTGATCGGGTCTTCCTTCATCAGGAACGAA TGCAGGAATTTGGGAACTGAGCTGTGCAAGTGCTGAAGAA
mef2d_sp4	ENST00000454816, ENST00000360595, ENST00000368240	CCTGCGAGTCATCACTTCCCAGGCAGGAAAGGGGTTAATGCATCACTTGACTGAGGACC ATTTAGATCTGAACAATGCCCAGCGCCTTGGGGTCTCCCA
MYC	ENST00000377970, ENST00000524013	AGGAGCAAAAGCTCATTTCTGAAGAGGACTTGTTGCGGAAACGACGAGAACAGTTGAAA CACAAACTTGAACAGCTACGGAACTCTTGTGCGTAAGGAAA
MYF5	ENST00000228644	TGGATTGCTTATCCAACATAGTGGACCGGATCACCTCCTCAGAGCAACCTGGGTTGCCTC TCCAGGATCTGGCTTCTCTCTCTCCAGTTGCCAGCACCGA

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
MYF6	ENST00000228641	GGAGGAGCAAGTATTGATTCGTCAGCCTCGAGTAGCCTTCGATGCCTTTCTTCCATCGTG GACAGTATTTCCTCGGAGGAACGCAAACTCCCCTGCGTGG
MYOD1	ENST00000250003	GCATGGTGTGTGGTGCTACAGGGAATTTGTACGTTTATACCGCAGGCGGGCGAGCCGCG GGCGCTCGCTCAGGTGATCAAAATAAAGGCGCTAATTTATA
NANOG	ENST00000229307, ENST00000526286, ENST00000526434, ENST00000541267	CTTCACCTATGCCTGTGATTTGTGGGCCTGAAGAAAACTATCCATCCTTGCAAATGTCTTC TGCTGAGATGCCTCACACGGAGACTGTCTCTCCTCTTCC
ncor2_sp1_T	ENST00000404621, ENST00000429285, ENST00000397355, ENST00000404121, ENST00000448614	GGCGTCGGGCGTGAGCGGAAATGAGGAGGAGATGGTGGAGGAGGCTGAAGCCACTGT CAACAACAGCTCAGACACCGAGAGCATCCCCTCTCCTCACACT
NFKB1_common	ENST00000226574, ENST00000394820, ENST00000504044, ENST00000505458, ENST00000600343	ACTGAATCTAAAAAGGACCCTGAAGGTTGTGACAAAAGTGATGACAAAAACACTGTAAA CCTCTTTGGGAAAGTTATTGAAACCACAGAGCAAGATCAGG
NOTCH1	ENST00000277541	GAGGGCTTCAGCGGTCCCAACTGCCAGACCAACATCAACGAGTGTGCGTCCAACCCATG TCTGAACCAGGGCACGTGTATTGACGACGTTGCCGGGTACA
NR2F2_sp2_T	ENST00000394166	CCAGTACTGCCGCCTCAAAAAGTGCCTCAAAGTGGGCATGAGACGGGAAGCGGTGCAG AGGGGCAGGATGCCGCCGACCCAGCCGACCCACGGGCAGTTC
OCT4_common	ENST00000259915, ENST00000441888, ENST00000471529, ENST00000512818, ENST00000513407	ATTCAGCCAAACGACCATCTGCCGCTTTGAGGCTCTGCAGCTTAGCTTCAAGAACATGTG TAAGCTGCGGCCCTTGCTGCAGAAGTGGGTGGAGGAAGCT
ONECUT1	ENST00000305901	CTTGGCAAGACAAATGATGAGCAGGAAAACACCACTGGATCTCACACCTTCAATCCATGA CCATCCTCGCTGTGCTTGGCTGTTTAGTGGTTTGGAGCAT
ONECUT2	ENST00000262095, ENST00000491143	CCCAAACCAAAATGCTTGACATAAAGCCAAATCAACTGCCAAGCACACTTTATTTTGCATA GGAGTATGCAGCCTAGGGAACCTTGGTTGAAAAGCAGCA
pbx1_common	ENST00000367897, ENST00000420696, ENST00000465089, ENST00000468104, ENST00000496120, ENST00000560469, ENST00000560641	TTTCTCTCCCAACGCTGAAGCGGTCAGACTGGAGGTCGAAGCAATCAGCAAACACAATAA GAGTCTCCTTCTCTTCTCTTCTTTGGGATGCTATTTCAGC
pbx3_sp1	ENST00000373483, ENST00000373482, ENST00000373492	AATGAAACCAGCGCTCTTCAGCGTCCTGTGTGAGATCAAAGAGAAAACAGGCATGTAAT GAATTTACTACACATGTGATGAACCTTCTCCGAGAACAGAG
pbx3_sp2_T	ENST00000491787, ENST00000447726, ENST00000342287, ENST00000373489, ENST00000373487	AATGAAACCAGCGCTCTTCAGCGTCCTGTGTGAGATCAAAGAGAAAACAGGTCTCAGCA TCAGAGGAGCCCAGGAGGAGGAACCCTCCCGATCCCCAGCTA
PER1	ENST00000581395, ENST00000579065, ENST00000354903, ENST00000582719, ENST00000317276	AGCACATCACGTCTGAGTACACACTTCAGAACCAGGATACCTTCTCAGTGGCTGTCTCCTT CCTGACGGGCCGAATCGTCTACATTTCGGAGCAGGCAGC
POLR2A	ENST00000322644	TTCTACTCCAACATTCAGACTGTCATTAACAACTGGCTCCTCATCGAGGGTCATACTATTG GCATTGGGGACTCCATTGCTGATTCTAAGACTTACCAGG
POU5F1_T	ENST00000259915, ENST00000441888, ENST00000471529, ENST00000512818, ENST00000513407	GAGGCTGCTGGGTCTCCTTTCTCAGGGGGACCAGTGTCCTTTCCTCTGGCCCCAGGGCCC CATTTTGGTACCCCAGGCTATGGGAGCCCTCACTTCACTG
PTEN	ENST00000371953	TTGGATGTGCAGCAGCTTACATGTCTGAAGTTACTTGAAGGCATCACTTTTAAGAAAGCT TACAGTTGGGCCCTGTACCATCCCAAGTCCTTTGTAGCTC
rbpj_sp2	ENST00000506956, ENST00000514807, ENST00000355476, ENST00000504907, ENST00000511546, ENST00000342320, ENST00000509158, ENST00000511451, ENST00000505958, ENST00000511401, ENST00000514730, ENST00000512351, ENST00000514675	CTGTGACTTACCTTAACATGTTCTTGAAGTACCATGGCGTGGATTAAAAGGAAATTTGGT GAGCGGCCTCCACCTAAACGACTTACTAGGGAAGCTATGC
RCOR1	ENST00000262241, ENST00000570597	GAAGGAACCACACCCCAGTTGTGCCGATTACATTAGTGTTGGCACACAGTCGGGTGCTA GTGTAACACAAATGCCGCGTTGTCTGGGTGTACAGTGTTTG
RELL2_common_T	ENST00000297164, ENST00000521367, ENST00000518856, ENST00000517794, ENST00000444782	GAATGAGGACACAGTAGAGAGGATTGTTCGCTGCATCATCCAGAATGAAGCCAATGCTG AGGCCTTGAAGGAGATGCTGGGGGACAGTGAAGGAGAAGGG
rrad_common	ENST00000299759, ENST00000420652, ENST00000566577, ENST00000568915	ACTCAGACGAGAGCGTTTACAAGGTGCTGCTGCTGGGGGCGCCCGGCGTGGGCAAGAG CGCCCTGGCGCGCATCTTCGGCGGTGTGGAGGACGGGCCTGA
Runx1_sp1	ENST00000300305, ENST00000437180, ENST00000475045, ENST00000416754, ENST00000486278	AGACAGCATATTTGAGTCATTTCCTTCGTACCCACAGTGCTTCATGAGAGAATGCATACTT GGAATGAATCCTTCTAGAGACGTCCACGATGCCAGCACG
Runx1_sp2	ENST00000344691, ENST00000358356, ENST00000399240	CCCTGTCGCCGTCTGGTAGGAGCTGTTTGCAGGGTCCTAACTCAATCGGCTTGTTGTGAT GCGTATCCCCGTAGATGCCAGCACGAGCCGCC
SIN3A	ENST00000394949, ENST00000360439, ENST00000394947	CTTCTATGGCAGATGCCAGCAAACATGGTGGTGGAACAGAATCGTTATTTTTTGATAAGG TCCGAAAGGCTCTTCGGAGTGCAGAAGCCTACGAAAATTT
SOX2	ENST00000325404, ENST00000431565	GCCTTTCCAAAAAATAATAATAACAATCATCGGCGGCGGCAGGATCGGCCAGAGGAGGA GGGAAGCGCTTTTTTTGATCCTGATTCCAGTTTGCCTCTCT
SOX4	ENST00000244745, ENST00000543472	GCATGCAGGCTTTTTGGCTTCCTACCTTGCAACAAAATAATTGCACCAACTCCTTAGTGCC GATTCCGCCCACAGAGAGTCCTGGAGCCACAGTCTTTTT
SP1	ENST00000426431	AGCCCTGGTGCTACTTGCTTGAAGTTTTCAGTGTAAGTACCCTGATGCCTTTTGGACCTTG GGATCAGATCAAGAGTTTTGGAGATCAGGTACCAAGGAA
SREBF2	ENST00000361204, ENST00000424354, ENST00000491541	CTGAGTTGCTGTAGCGTCTTGATTCTCTCCCTGGGTCTGCGTTCCCTCCCCTGGGCCTGAC TGAGCCTGCTCATTGTTTTTCCCTTTATTACACAGGACA
SRF	ENST00000265354	AGAGCCTACCTTCACCACCTATATCCAGAAGGGGAGCTTTTTCAGAAACAGGGCAGCAGT GGGGTGAAATTTTCTTAACCCCTAAGACTGCCTTCAGTAG
stat1_common	ENST00000361099, ENST00000392322, ENST00000409465, ENST00000452281, ENST00000540176, ENST00000392323	TGCTGAATGTCACTGAACTTACCCAGAATGCCCTGATTAATGATGAACTAGTGGAGTGGA AGCGGAGACAGCAGAGCGCCTGTATTGGGGGGCCGCCCAA
STAT2	ENST00000314128, ENST00000555665, ENST00000556539, ENST00000557235	CAACATTTTAATAGTTGGTTAGGCTAAACTGGTGCATACTGGCATTGGCCCTTGGTGGGG AGCACAGACACAGGATAGGACTCCATTTCTTTCTTCCATT
stat3_common	ENST00000264657, ENST00000389272, ENST00000404395, ENST00000585517, ENST00000588969	GCTGAAATCATCATGGGCTATAAGATCATGGATGCTACCAATATCCTGGTGTCTCCACTG GTCTATCTCTATCCTGACATTCCCAAGGAGGAGGCATTCG
STAT5A	ENST00000588868, ENST00000345506, ENST00000452307	GCCGCCGGTTTGAGTGAGGGTTTCTGAGCTGCTCTGAATTAGTCCTTGCTTGGCTGCTTG GCCTTGGGCTTCATTCAAGTCTATGATGCTGTTGCCCACG
STAT5B	ENST00000293328	GCCTAGAGAGTGGAGATTTTTGATGAAAGGTGTGCTCGCTCTCTGCGTTCTATCTTCTCTC TCCTCCTTGTTCCTGCAAACCACAAGATAAAGGTAGTGG
TAF1_sp1	ENST00000449580, ENST00000373790	CCAATGAAGAAGGATAAGGACCAGGATTCTATTACTGGTGAGAAAGTGGACTTCAGTAG TTCCTCTGACTCAGAATCTGAGATGGGACCT
TAF1_sp2	ENST00000276072, ENST00000423759	CCAATGAAGAAGGATAAGGACCAGGATTCTATTACTGGTGTGTCTGAAAATGGAGAAGG CATCATCTTGCCCTCCATCATTGCCCCTTCC
TCF12_common_T	ENST00000267811, ENST00000333725, ENST00000438423, ENST00000452095, ENST00000557843, ENST00000559609, ENST00000560190, ENST00000561449, ENST00000343827, ENST00000537840, ENST00000543579, ENST00000559703, ENST00000559710, ENST00000561420	CCTGAACAGAAGATAGAAAGGGAGAAGGAGAGGCGGATGGCTAACAATGCCAGAGAA CGCTTACGCGTGCGGGATATTAATGAAGCATTCAAAGAGCTTG
TCF3_sp2_T	ENST00000593064, ENST00000588136	CCCACCCAGGCCTGAGCGAAGCCCACAACCCCGCCGGGCACATGTGAAAGTAAACAAAA CCTGAAAGCAAGCAACAAAACATACACTTTGTCAGAGAAGA
TCF3_sp3_T	ENST00000262965, ENST00000344749, ENST00000395423, ENST00000585731, ENST00000590684	CCCACCCAGGCCTGAGCGAAGCCCACAACCCCGCCGGGCACATGTGAAAGGTATGCCTC CGTGGGACGAGCCACCCGCTTTCAGCCCTGTGCTCTGGCCC
TCF3_sp4	ENST00000586164, ENST00000593064, ENST00000587425, ENST00000395423, ENST00000592628, ENST00000262965	CACCAGGCTGTCTCGGTCATCCTGAACTTGGAGCAGCAAGTGCGAGAGCGGAACCTGAA TCCCAAAGCAGCCTGTTTGAAACGGCGAGAAGAGGAA
TCF3_sp5_T	ENST00000588136, ENST00000453954, ENST00000344749, ENST00000585731, ENST00000585855, ENST00000590684, ENST00000592395	CCTGCAGCAGGCCGTGCAGGTCATCCTGGGGCTGGAGCAGCAGGTGCGAGAGCGGAAC CTGAATCCCAAAGCAGCCTGTTTGAAACGGCGAGAAAGAGGAA
TCF3_sp6_T	ENST00000395423	CGGGAAGGGCCGGCCCGCCTCCCTGGCCGGGGCGCAGTTCGGAGGTTCAGGCAAGAGC GGTGAGCGGGGCGCCTATGCCTCCTTCGGGAGAGACGCAGGC
TCF3_sp7_T	ENST00000588136, ENST00000344749, ENST00000262965	GTCGCACAGCAGCCTCTCTTCATCCACATTCCTGGGACCGGGACTCGGAGGCAAGAGCG GTGAGCGGGGCGCCTATGCCTCCTTCGGGAGAGACGCAGGC
TNFRSF13B	ENST00000261652, ENST00000437538, ENST00000579315, ENST00000581616, ENST00000582931, ENST00000583789, ENST00000584950	AGGAGCAAGGCAAGTTCTATGACCATCTCCTGAGGGACTGCATCAGCTGTGCCTCCATCT GTGGACAGCACCCTAAGCAATGTGCATACTTCTGTGAGAA
TNFRSF13C	ENST00000291232	GTTTGGTGTGCTTGCCTTTGGCTTCAGACCTCACCATCTTTGACAGCCCTTGAAGGTGGTA GCCCAGCTCCTGTTCCTGTGCCTTCAAAAGGCTGGGGCA
TNFRSF17_sp1	ENST00000053243	GCTAAGGAAGATAAACTCTGAACCATTAAAGGACGAGTTTAAAAACACAGGATCAGGTC TCCTGGGCATGGCTAACATTGACCTGGAAAAGAGCAGGACT
TNFSF13B	ENST00000375887, ENST00000430559	AACAGGAAATGATCCATTCCCTGTGGTCACTTATTCTAAAGGCCCCAACCTTCAAAGTTCA

Probe ID	Targeted trancripts (Ensembl IDs)	Probe sequence
		AGTAGTGATATGGATGACTCCACAGAAAGGGAGCAGTCA
USF1_common_T	ENST00000368019, ENST00000368021, ENST00000435396, ENST00000472217, ENST00000473969, ENST00000528768, ENST00000531842, ENST00000368020	AAGCTTGTGATTATATCCAGGAGCTTCGGCAGAGTAACCACCGCTTGTCTGAAGAACTGC AGGGACTTGACCAACTGCAGCTGGACAATGACGTGCTTCG
YY1	ENST00000262238, ENST00000554579, ENST00000554804, ENST00000555735	ACATGCTAAGGCCAAAAACAACCAGTGAAAAGAAGAGAGAAGACCCTTCTCGACCACGG GAAGCATCTTCCAGAAGTGTGATTGGGAATAAATATGCCTC
ZIC3	ENST00000287538	TCAGTTAGTGGCCATGACATCTCAATCTTGTACTTCAAAGACTGAGAAGCTGGATTTAATC ATCCCTGCCCTACATATATAAACATAAGGTAACCTACTG
ZNF217_sp3_T	ENST00000371471	AGTCCCGGCCGCCGCCGCGCGAGGAAATGGCCGAGGAGCCGGAGCCGCAGGGTTTGGA AATCCCTTGTCTCCAGGTTGCTGGGATTGACTTCTTGCTCAA

Supplementary Table 8. NanoString counts and RPKMs for predominant compatible isoforms

	Nanocounts	mGene graph	Tromer	SLIDE all	mGene	mTim	AUGUSTUS all	Cufflinks	iReckon full	iReckon ends	Trembly all	Transomics all
adar1_sp1	2101	8.019	11.800	0.000	0.000	4.737	0.000	12.261	27.993	23.975	15.097	0.000
adar1_sp2	2321	6.343	10.100	3.529	15.642	0.000	13.182	11.639	0.422	1.950	0.732	18.420
atf2_common	4911	6.016	20.100	6.902	13.868	12.906	9.553	12.499	0.000	15.860	5.974	0.000
atf2_sp1	2299	6.016	19.100	0.000	0.000	12.906	0.000	2.711	0.000	4.636	1.503	0.000
atf2_sp2	2156	4.211	54.300	6.902	13.868	0.000	9.553	12.499	0.000	15.860	5.974	21.380
ATP5J_common	16504	6.764	4.800	10.377	18.577	32.224	33.923	174.528	61.534	50.080	30.978	22.280
Bcl11a_sp1	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Bcl11a_sp2	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
BCL3	1737	0.000	1.700	0.884	0.557	3.767	2.791	6.189	5.926	2.468	0.000	3.780
BHLHB2	17578	39.789	36.600	29.947	62.179	84.314	51.583	0.000	107.521	87.483	53.854	0.000
Blnk_sp1_T	33	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.000	0.000	10.930
Blnk_sp2	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CARM1_sp1_T	1168	4.848	3.700	0.000	0.000	9.981	8.519	19.494	0.000	18.037	9.809	0.000
CD19	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.067	0.071	0.000	0.000
Cd79b_sp1_T	24	0.000	0.000	0.000	0.043	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Cd79b_sp2_T	0	0.000	0.000	0.000	0.000	0.000	0.018	0.000	0.000	0.000	0.000	0.000
cdkn1a_sp1_T	6057	21.884	0.000	0.000	0.000	0.000	21.888	43.565	53.015	42.893	29.102	0.000
cdkn1a_sp2_T	5429	5.581	0.000	0.000	0.000	11.081	0.000	0.000	0.702	0.793	1.044	0.000
CEBPA	7109	0.000	0.000	7.464	0.000	0.000	8.776	0.000	21.206	8.371	0.000	0.000
CTCF_common	255	9.574	10.600	1.337	11.822	10.033	8.977	16.957	20.358	16.725	11.564	0.000
CTCF_sp1	563	9.574	8.400	2.043	11.822	10.033	8.977	16.957	20.358	16.725	11.564	0.000
CTCFL_sp1	19	0.000	0.000	0.000	0.000	0.000	0.261	0.000	0.000	0.000	0.000	0.060
CTCFL_sp2	11	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CTDSL_sp2_T	940	18.993	37.300	0.000	20.476	3.870	9.696	16.728	0.000	22.162	7.568	17.710
CTDSP1_common_T	4117	10.701	22.100	8.134	20.135	14.558	11.711	14.311	25.545	20.837	9.889	42.120
CTDSP2	3330	0.000	0.000	14.322	0.000	0.000	25.306	23.741	44.528	38.755	20.104	0.000
CTDSPL_sp1	4193	18.993	36.700	0.000	20.476	3.870	5.744	11.590	0.000	6.027	5.331	0.000
DES	0	0.000	0.000	0.000	0.000	0.000	0.018	0.000	0.000	0.000	0.000	0.000
DNMT1_common_T	1011	2.208	0.000	2.193	6.619	2.270	3.566	7.063	11.401	6.460	4.820	6.190
E2F4	4533	16.233	0.000	0.000	19.528	14.254	13.402	23.889	29.670	25.409	16.069	17.580
E2F6_common	5693	3.922	1.500	1.555	5.451	0.000	4.099	0.000	0.000	0.000	5.260	2.960
ebf1_sp1	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
EGR1	956	7.684	8.700	2.694	7.895	8.305	4.390	8.910	10.040	8.180	6.463	0.000
EOMES_common	14	0.000	0.000	0.007	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.170
EP300	1914	11.002	16.700	2.775	11.107	10.824	8.552	14.063	16.915	13.800	8.142	12.510
esr1_common	12	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
esr2_sp1	33	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
esr2_sp2	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.020
ets1_common	561	0.000	0.600	0.566	0.307	0.000	0.210	0.376	0.303	0.153	0.368	0.320
FBXO15	14	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010
FOS	755	1.783	2.400	0.542	1.783	0.000	1.255	2.589	2.962	2.395	1.888	0.000
foxa2_body	8700	14.687	16.500	3.488	17.953	0.000	12.329	21.703	14.811	12.235	15.405	0.000
foxa2_sp2_T	250	0.000	0.000	3.488	0.000	17.363	0.000	4.916	11.101	8.888	15.405	0.000
FOXA3	1223	9.923	17.100	3.160	12.663	11.999	8.652	17.657	19.049	15.646	10.745	0.000
gabpa_sp1_T	251	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.681	0.000	1.288	0.000
gabpa_sp2_T	237	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.386	0.256	0.000	0.000
GATA1_T	12	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	11.040
HDAC1	7349	9.321	15.100	7.233	37.448	24.484	0.000	45.483	48.939	38.679	30.272	0.000
HDAC3_sp1	3703	7.760	10.500	0.000	7.760	0.000	13.394	26.599	0.000	23.591	15.355	12.670
HDAC4	902	0.000	3.100	0.786	0.000	1.532	0.000	0.825	3.016	0.093	0.813	0.000
HDAC5	65	0.000	0.000	0.445	0.000	0.000	0.000	0.000	0.000	1.188	13.694	0.000
Hif1a_common	13561	18.760	21.500	7.487	27.504	18.168	13.669	33.054	21.754	18.023	17.016	26.690
HNF1A	1485	8.247	3.600	0.754	8.251	9.176	4.968	2.700	10.798	7.026	3.155	10.810
HNF1B	836	2.287	4.400	1.487	2.903	2.837	1.275	2.439	2.896	2.302	1.823	2.570
Hnf4g_common_T	1906	2.897	3.600	0.793	3.550	0.000	0.000	3.802	4.060	1.977	2.896	0.000
HSF1	4997	13.666	10.900	0.348	19.782	9.781	18.878	22.358	39.006	28.277	8.647	15.140
IGF1R_common_T	2654	5.520	25.300	2.545	12.103	7.032	4.483	12.120	10.999	12.999	5.215	10.780
IKZF1_common	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IKZF1_sp1_T	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IKZF1_sp2	16	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IKZF3_sp1_T	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.060
IKZF3_sp2	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IL6	0	0.000	0.000	0.000	0.000	0.000	0.011	0.000	0.000	0.000	0.000	0.000
IL6receptor_common	2319	6.038	10.300	2.339	6.584	6.117	0.000	8.476	9.792	8.532	5.920	0.000
IL8	84	0.298	0.000	0.817	0.298	0.000	0.031	0.000	0.516	0.231	0.000	0.000
IL8RA	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IL8RB	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IRF8	18	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
JUND	297	38.720	0.000	115.202	38.720	0.000	0.000	0.000	56.163	45.642	0.000	0.000
KAISO_sp1_T	1368	4.853	0.000	2.190	4.853	0.000	2.848	4.740	6.764	5.589	4.147	0.000
KAP1_common_T	0	48.402	22.000	1.195	53.394	48.369	49.948	91.944	2.299	65.311	68.574	62.880
KLF4	226	0.536	1.200	0.504	0.536	0.000	0.000	0.598	0.692	0.466	0.944	0.000
LEF1_sp1	0	0.000	0.000	0.000	0.102	0.000	0.000	0.000	0.046	0.000	0.000	0.000
LEF1_sp2	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
LIN28_common	0	0.000	0.000	0.000	0.043	0.000	0.011	0.000	0.000	0.000	0.000	0.010
MAX_sp1	2984	6.122	7.900	0.000	5.156	6.801	4.591	10.640	0.000	3.786	4.277	8.040
mef2a_common_T	515	1.998	0.000	1.386	4.385	5.082	2.034	4.035	5.451	4.205	2.928	4.700
mef2a_sp3	1470	1.998	8.000	0.309	0.000	5.082	2.034	4.035	5.451	4.205	2.928	0.000
mef2a_sp4	1090	1.266	7.600	1.386	4.385	0.000	0.851	1.627	1.744	1.473	0.892	4.700
mef2b_sp10_T	81	0.000	8.000	0.000	0.000	0.000	4.953	0.000	0.000	1.182	0.000	2.280
mef2b_sp11_T	248	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.074	0.193	0.000

Nature Methods: doi:10.1038/nmeth.2714

	Nanocounts	mGene graph	Tromer	SLIDE all	mGene	mTim	AUGUSTUS all	Cufflinks	iReckon full	iReckon ends	Trembly all	Transomics all
mef2b_sp12	311	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.182	0.882	0.000
mef2b_sp7_T	215	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.424	0.000
mef2c_sp1	9	3.852	2.200	1.395	3.852	0.000	2.842	3.425	0.000	0.000	0.000	0.000
mef2c_sp3	18	60.369	54.200	10.330	72.155	47.513	45.253	60.126	0.000	0.000	0.000	0.000
mef2c_sp4	16	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
mef2d_sp4	80	0.000	0.000	0.000	0.000	0.000	0.000	0.000	3.045	2.576	3.004	0.000
MYC	28966	0.000	0.000	0.000	0.000	0.000	0.000	0.000	78.566	69.472	40.507	0.000
MYF5	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.070
MYF6	51	3.416	9.700	1.049	0.000	2.773	1.998	6.599	0.000	0.000	0.000	7.360
MYOD1	32	5.066	0.000	1.638	3.331	3.444	2.847	9.295	0.000	0.000	0.000	4.760
NANOG	64	0.226	0.000	0.000	0.226	0.000	2.475	0.000	0.000	0.011	0.000	0.190
ncor2_sp1_T	1143	8.053	0.000	0.000	8.053	0.000	5.163	9.870	0.000	1.034	2.510	11.050
NFKB1_common	355	1.011	0.000	0.000	1.011	3.434	1.299	0.504	4.062	4.897	3.380	0.980
NOTCH1	27	6.427	8.000	1.518	6.427	0.000	3.314	6.747	0.284	0.032	0.000	0.000
NR2F2_sp2_T	800	4.323	4.400	0.963	4.323	0.000	2.183	2.163	4.223	3.783	7.895	0.000
OCT4_common	284	0.000	0.000	0.023	0.000	0.000	0.000	0.000	0.084	1.796	0.000	0.000
ONECUT1	2013	0.000	0.000	0.000	0.000	0.000	0.000	0.823	7.115	5.992	5.487	0.000
ONECUT2	805	0.752	0.000	0.000	1.089	1.154	0.774	1.008	4.046	3.806	3.005	0.960
pbx1_common	0	2.905	0.000	1.231	2.905	2.990	2.404	4.568	0.001	0.000	0.000	3.320
pbx3_sp1	97	15.249	0.000	7.541	17.974	15.275	13.497	22.026	0.702	0.592	0.000	20.140
pbx3_sp2_T	89	0.000	0.000	0.828	0.000	3.434	1.299	0.000	1.215	1.032	0.860	0.980
PER1	452	3.611	2.700	0.000	4.498	0.000	0.000	0.000	5.530	4.592	3.320	0.000
POLR2A	4377	0.000	3.900	0.000	0.000	0.000	2.728	5.559	31.669	26.517	18.840	0.000
POU5F1_T	94	0.000	7.800	0.888	0.000	5.158	3.436	0.000	0.084	1.796	0.000	0.000
PTEN	1128	0.629	0.000	0.258	2.280	0.000	1.567	1.227	4.292	1.040	0.000	3.240
rbpj_sp2	82	0.577	1.300	1.881	0.577	0.602	0.407	1.140	0.000	5.918	2.401	0.700
RCOR1	2199	0.000	0.000	0.000	0.000	0.000	0.000	0.000	8.782	6.885	5.339	0.000
RELL2_common_T	226	0.000	0.000	0.600	0.000	0.000	1.474	3.147	0.000	0.000	1.725	0.000
rrad_common	82	5.612	10.400	1.445	6.486	5.683	4.965	8.384	0.409	0.740	0.000	6.620
Runx1_sp1	24	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Runx1_sp2	338	0.000	4.600	1.114	0.000	0.000	0.000	0.000	1.734	1.439	1.291	0.000
SIN3A	1241	0.000	7.300	1.759	0.000	32.098	0.000	6.523	5.741	8.155	0.000	0.000
SOX2	0	28.334	25.500	14.937	50.738	27.716	37.410	80.463	0.000	0.034	0.000	0.000
SOX4	528	0.000	12.200	4.700	0.000	14.379	7.759	13.846	0.063	2.135	0.000	0.000
SP1	513	8.426	0.000	3.291	9.046	7.994	7.303	12.281	7.665	6.247	5.301	3.290
SREBF2	7451	3.716	6.000	1.516	6.431	0.000	2.611	8.957	84.859	3.228	46.195	0.000
SRF	790	8.985	3.300	7.878	26.855	10.632	10.560	30.035	17.672	14.304	12.139	21.140
stat1_common	1154	1.007	1.700	0.416	1.491	0.000	0.000	1.743	13.717	10.893	8.726	0.000
STAT2	1235	7.422	0.000	4.175	9.714	0.000	6.814	12.711	2.849	6.198	1.597	0.000
stat3_common	7640	0.000	2.600	0.000	0.000	0.000	0.000	0.000	0.000	0.000	5.170	0.000
STAT5A	396	0.960	3.000	1.889	2.421	0.000	1.776	2.442	1.113	1.393	1.264	3.290
STAT5B	1717	2.932	6.700	1.193	7.708	5.653	5.623	8.754	15.297	12.872	8.823	5.400
TAF1_sp1	270	0.000	0.000	0.000	0.000	0.000	3.949	5.753	0.136	0.000	0.000	0.000
TAF1_sp2	326	0.000	16.900	3.247	0.000	0.000	0.000	7.678	2.182	0.000	2.013	0.000
TCF12_common_T	2427	1.818	0.600	0.000	0.000	0.000	0.000	7.183	8.493	6.724	4.822	0.000
TCF3_sp2_T	1214	1.903	0.000	3.247	8.286	0.000	3.949	7.678	0.000	10.253	4.776	10.600
TCF3_sp3_T	1102	1.818	5.700	0.000	0.000	1.393	0.000	0.000	0.000	5.126	0.000	0.000
TCF3_sp4	1153	1.650	3.900	0.000	0.000	4.494	3.949	7.183	0.000	8.657	4.776	10.600
TCF3_sp5_T	1031	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	10.253	2.450	0.090
TCF3_sp6_T	1111	0.000	0.400	0.161	0.158	0.000	0.000	0.000	0.000	2.463	4.776	0.000
TCF3_sp7_T	1646	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	10.253	2.450	0.000
TNFRSF13B	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	2.110
TNFRSF13C	22	2.951	5.400	2.270	4.313	2.831	3.584	5.177	0.203	0.125	0.000	3.250
TNFRSF17_sp1	0	17.687	10.100	5.461	17.687	17.874	11.013	13.888	0.000	0.000	0.000	0.000
TNFSF13B	84	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.018	0.000	0.000
USF1_common_T	623	0.000	0.800	5.349	0.000	0.000	0.000	13.414	5.824	4.726	3.154	0.000
YY1	11801	8.019	11.800	0.000	0.000	4.737	0.000	12.261	22.801	18.355	11.602	0.000
ZIC3	46	6.343	10.100	3.529	15.642	0.000	13.182	11.639	0.000	0.000	0.000	18.420
ZNF217_sp3_T	1401	6.016	20.100	6.902	13.868	12.906	9.553	12.499	20.481	15.954	8.342	0.000

Nature Methods: doi:10.1038/nmeth.2714

Supplementary Table 9. NanoString counts and RPKMs for predominant isoforms

	Nanocounts	AUGUSTUS all	Cufflinks	iReckon full	iReckon ends	mGene	mGene graph	mTim	SLIDE all	Transomics all	Trembly all	Tromer
adar1_sp1	2101	13.182	12.261	27.993	23.975	15.642	8.019	4.737	10.275	18.420	18.420	29.500
adar1_sp2	2321	13.182	12.261	27.993	23.975	15.642	8.019	4.737	10.275	18.420	18.420	29.500
atf2_common	4911	9.553	12.499	0.000	15.860	13.868	6.016	12.906	6.902	21.380	21.380	54.300
atf2_sp1	2299	9.553	12.499	0.000	15.860	13.868	6.016	12.906	6.902	21.380	21.380	54.300
atf2_sp2	2156	9.553	12.499	0.000	15.860	13.868	6.016	12.906	6.902	21.380	21.380	54.300
ATP5J_common	16504	33.923	174.528	61.534	50.080	18.577	6.764	32.224	10.377	22.280	22.280	59.500
Bcl11a_sp1	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.140	0.140	0.000
Bcl11a_sp2	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.140	0.140	0.000
BCL3	1737	2.791	6.189	5.926	3.573	4.486	4.486	3.767	1.557	3.780	3.780	7.200
BHLHB2	17578	51.583	82.714	107.521	87.483	62.179	39.789	84.314	29.947	88.550	88.550	82.600
Blnk_sp1_T	33	0.007	0.000	0.008	0.000	7.716	7.716	0.000	0.011	10.930	10.930	0.000
Blnk_sp2	0	0.007	0.000	0.008	0.000	7.716	7.716	0.000	0.011	10.930	10.930	0.000
CARM1_sp1_T	1168	8.519	19.494	0.000	18.037	20.105	9.535	9.981	3.743	9.540	9.540	43.900
CD19	0	0.612	2.669	0.067	0.071	0.681	0.000	0.000	7.420	0.190	0.190	1.600
Cd79b_sp1_T	24	0.018	0.000	0.000	0.000	0.043	0.000	0.000	0.000	0.000	0.000	0.000
Cd79b_sp2_T	0	0.018	0.000	0.000	0.000	0.043	0.000	0.000	0.000	0.000	0.000	0.000
cdkn1a_sp1_T	6057	21.888	43.565	53.015	42.893	27.358	21.884	11.361	5.722	37.470	37.470	17.200
cdkn1a_sp2_T	5429	21.888	43.565	53.015	42.893	27.358	21.884	11.361	5.722	37.470	37.470	17.200
CEBPA	7109	8.776	0.000	21.206	8.371	5.980	5.274	0.000	7.464	6.380	6.380	0.000
CTCF_common	255	8.977	16.957	20.358	16.725	11.822	9.574	10.033	3.169	13.750	13.750	21.100
CTCF_sp1	563	8.977	16.957	20.358	16.725	11.822	9.574	10.033	3.169	13.750	13.750	21.100
CTCFL_sp1	19	0.261	0.000	0.012	0.041	0.099	0.000	0.000	0.721	0.060	0.060	0.200
CTCFL_sp2	11	0.261	0.000	0.012	0.041	0.099	0.000	0.000	0.721	0.060	0.060	0.200
CTDSL_sp2_T	940	9.696	16.728	0.000	22.162	20.476	18.993	19.435	6.404	17.710	17.710	37.300
CTDSP1_common_T	4117	11.711	14.311	25.545	20.837	20.135	10.701	14.558	8.134	42.120	42.120	38.700
CTDSP2	3330	25.306	23.741	44.528	38.755	30.798	25.746	24.823	14.322	20.100	20.100	22.300
CTDSPL_sp1	4193	9.696	16.728	0.000	22.162	20.476	18.993	19.435	6.404	17.710	17.710	37.300
DES	0	0.018	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
DNMT1_common_T	1011	3.566	12.116	11.401	6.460	6.619	2.208	2.270	2.888	6.190	6.190	16.400
E2F4	4533	13.402	23.889	29.670	25.409	19.528	16.233	14.254	3.340	17.580	17.580	35.100
E2F6_common	5693	4.099	7.950	0.000	0.000	5.451	3.922	4.733	2.139	2.960	2.960	8.000
ebf1_sp1	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
EGR1	956	4.390	8.910	10.040	8.180	7.895	7.684	8.305	2.694	9.340	9.340	8.700
EOMES_common	14	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.007	0.170	0.170	0.000
EP300	1914	8.552	14.063	16.915	13.800	11.107	11.002	10.824	2.775	12.510	12.510	18.200
esr1_common	12	0.000	0.000	0.000	0.001	0.005	0.000	0.000	0.001	0.000	0.000	0.000
esr2_sp1	33	0.000	0.000	0.000	0.081	0.087	0.000	0.000	1.196	0.020	0.020	0.200
esr2_sp2	0	0.000	0.000	0.000	0.081	0.087	0.000	0.000	1.196	0.020	0.020	0.200
ets1_common	561	0.210	0.376	0.303	0.153	0.307	0.000	0.000	0.566	0.320	0.320	0.600
FBXO15	14	0.000	0.000	0.000	0.000	0.016	0.000	0.000	0.938	0.010	0.010	0.300
FOS	755	1.255	2.589	2.962	2.395	1.783	1.783	0.000	1.000	2.290	2.290	2.400
foxa2_body	8700	12.329	21.703	14.811	12.235	17.953	14.687	17.363	3.488	18.410	18.410	33.200
foxa2_sp2_T	250	12.329	21.703	14.811	12.235	17.953	14.687	17.363	3.488	18.410	18.410	33.200
FOXA3	1223	8.652	17.657	19.049	15.646	12.663	9.923	11.999	3.160	11.930	11.930	17.100
gabpa_sp1_T	251	2.389	6.092	2.225	0.256	2.723	1.769	0.000	1.391	3.480	3.480	6.400
gabpa_sp2_T	237	2.389	6.092	2.225	0.256	2.723	1.769	0.000	1.391	3.480	3.480	6.400
GATA1_T	12	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	11.040	11.040	0.000
HDAC1	7349	27.696	86.127	48.939	38.679	37.448	9.321	24.484	7.233	27.570	27.570	44.600
HDAC3_sp1	3703	13.394	26.599	0.000	23.591	20.748	20.748	19.399	4.831	12.670	12.670	21.700
HDAC4	902	0.863	1.522	3.016	2.396	2.217	0.943	1.532	0.786	1.720	1.720	8.300
HDAC5	65	12.839	26.933	0.000	47.104	33.155	7.620	15.081	18.303	26.610	26.610	72.300
Hif1a_common	13561	13.669	33.054	21.754	18.023	27.504	18.760	18.168	7.487	26.690	26.690	32.500
HNF1A	1485	4.968	2.700	10.798	7.026	8.251	8.247	9.176	2.089	10.810	10.810	3.600
HNF1B	836	1.275	2.439	2.896	2.302	2.903	2.287	2.837	1.487	2.570	2.570	6.600
Hnf4g_common_T	1906	2.359	3.802	4.060	1.977	3.550	2.897	2.040	0.818	1.220	1.220	4.200
HSF1	4997	18.878	22.358	39.006	28.277	19.782	13.666	9.781	5.110	15.140	15.140	65.500
IGF1R_common_T	2654	9.782	12.120	10.999	12.999	17.784	17.784	7.032	2.545	14.320	14.320	25.300
IKZF1_common	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IKZF1_sp1_T	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IKZF1_sp2	16	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IKZF3_sp1_T	0	0.000	0.000	0.000	0.015	11.141	3.412	0.000	0.000	0.060	0.060	0.000
IKZF3_sp2	0	0.000	0.000	0.000	0.015	11.141	3.412	0.000	0.000	0.060	0.060	0.000
IL6	0	0.011	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.010	0.000
IL6receptor_common	2319	4.373	8.476	9.792	8.532	6.584	6.038	6.117	3.045	5.370	5.370	14.900
IL8	84	0.031	0.475	0.516	0.231	0.298	0.298	0.000	0.817	0.160	0.160	0.500
IL8RA	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.030	0.000
IL8RB	0	0.013	0.000	0.000	0.000	0.029	0.000	0.000	0.262	0.010	0.010	0.000
IRF8	18	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
JUND	297	25.218	0.000	56.163	45.642	38.720	38.720	0.000	115.202	46.830	46.830	0.000
KAISO_sp1_T	1368	2.848	4.740	6.764	5.589	4.853	4.853	0.000	2.190	4.920	4.920	6.000
KAP1_common_T	0	49.948	91.944	119.818	65.311	53.394	48.402	48.369	23.632	62.880	62.880	145.100
KLF4	226	0.137	0.598	0.692	0.466	0.536	0.536	0.000	0.504	0.370	0.370	1.200
LEF1_sp1	0	0.030	0.000	0.046	0.006	0.102	0.000	0.000	0.000	0.040	0.040	0.300
LEF1_sp2	0	0.030	0.000	0.046	0.006	0.102	0.000	0.000	0.000	0.040	0.040	0.300
LIN28_common	0	0.011	0.000	0.000	0.000	0.043	0.000	0.000	0.000	0.010	0.010	0.000
MAX_sp1	2984	4.591	10.640	0.000	3.786	5.156	6.122	6.801	2.226	8.040	8.040	7.900
mef2a_common_T	515	2.034	4.035	5.451	4.205	4.385	1.998	5.082	1.386	4.700	4.700	9.200
mef2a_sp3	1470	2.034	4.035	5.451	4.205	4.385	1.998	5.082	1.386	4.700	4.700	9.200
mef2a_sp4	1090	2.034	4.035	5.451	4.205	4.385	1.998	5.082	1.386	4.700	4.700	9.200
mef2b_sp10_T	81	4.953	3.464	0.000	1.182	1.384	1.088	1.294	1.760	2.280	2.280	8.000
mef2b_sp11_T	248	0.000	0.000	0.000	3.168	0.000	0.000	0.000	0.000	0.000	2.280	0.000

Nature Methods: doi:10.1038/nmeth.2714

	Nanocounts	AUGUSTUS all	Cufflinks	iReckon full	iReckon ends	mGene	mGene graph	mTim	SLIDE all	Transomics all	Trembly all	Tromer
mef2b_sp12	311	0.000	0.000	0.000	3.168	0.000	0.000	0.000	0.000	0.000	2.280	0.000
mef2b_sp7_T	215	0.000	0.000	0.000	3.168	0.000	0.000	0.000	0.000	0.000	2.280	0.000
mef2c_sp1	9	2.842	3.425	0.002	0.000	3.852	3.852	0.000	1.395	3.900	0.000	8.200
mef2c_sp3	18	45.253	60.126	0.002	0.000	72.155	60.369	47.513	16.867	103.100	0.000	54.200
mef2c_sp4	16	0.733	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
mef2d_sp4	80	0.733	0.000	3.045	2.576	0.000	0.000	0.000	0.000	0.000	3.900	0.000
MYC	28966	0.000	0.000	78.566	69.472	0.000	0.000	0.000	0.000	0.000	103.100	0.000
MYF5	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.070	0.000	0.000
MYF6	51	1.998	6.599	0.000	0.000	8.090	3.416	2.773	5.106	7.360	0.000	24.700
MYOD1	32	2.847	9.295	0.000	0.000	3.331	5.066	3.444	1.638	4.760	0.000	9.700
NANOG	64	2.475	0.496	0.000	0.011	0.226	0.226	0.000	0.310	0.190	0.070	0.500
ncor2_sp1_T	1143	5.163	9.870	0.000	7.648	8.053	8.053	9.022	2.367	11.050	7.360	5.800
NFKB1_common	355	1.299	0.504	4.062	4.897	1.011	1.011	3.434	0.828	0.980	4.760	0.000
NOTCH1	27	3.314	6.747	0.284	0.062	6.427	6.427	4.973	1.518	4.420	0.190	8.000
NR2F2_sp2_T	800	2.183	2.163	6.076	4.056	4.323	4.323	0.000	0.963	1.900	11.050	4.400
OCT4_common	284	0.000	0.000	0.084	1.796	0.000	0.000	0.000	0.023	0.010	0.980	0.000
ONECUT1	2013	0.774	1.008	7.115	5.992	1.089	0.752	1.154	0.476	0.960	4.420	2.100
ONECUT2	805	0.774	1.008	4.046	3.806	1.089	0.752	1.154	0.476	0.960	1.900	2.100
pbx1_common	0	2.404	4.568	0.001	0.000	2.905	2.905	10.530	1.598	3.320	0.010	2.300
pbx3_sp1	97	13.497	22.026	1.215	1.032	17.974	15.249	15.275	7.810	20.140	0.960	30.500
pbx3_sp2_T	89	1.299	0.504	1.215	1.032	1.011	1.011	3.434	0.828	0.980	0.960	0.000
PER1	452	5.492	7.897	5.530	4.592	4.498	3.611	0.000	3.308	35.660	3.320	7.000
POLR2A	4377	2.927	11.227	31.669	26.517	6.303	3.903	6.441	2.160	1.790	20.140	9.700
POU5F1_T	94	3.436	6.237	0.084	1.796	4.716	4.716	5.158	0.888	2.200	0.980	7.800
PTEN	1128	1.567	1.227	4.292	4.046	2.280	0.629	0.000	0.696	3.240	35.660	3.100
rbpj_sp2	82	0.407	1.140	0.000	5.918	0.577	0.577	0.602	1.881	0.700	1.790	21.000
RCOR1	2199	1.474	3.147	8.782	6.885	1.992	1.741	2.388	0.991	1.860	2.200	4.600
RELL2_common_T	226	1.474	3.147	0.000	1.064	1.992	1.741	2.388	0.991	1.860	3.240	4.600
rrad_common	82	4.965	8.384	0.409	0.740	6.486	5.612	5.683	1.445	6.620	0.700	13.800
Runx1_sp1	24	0.008	0.000	1.734	1.439	0.022	0.022	0.000	0.009	0.020	1.860	0.000
Runx1_sp2	338	0.608	0.000	1.734	1.439	1.566	1.566	0.000	1.114	1.950	1.860	4.600
SIN3A	1241	61.269	6.523	5.741	8.155	5.816	5.799	32.098	1.759	6.870	6.620	8.800
SOX2	0	37.410	80.463	0.000	0.034	50.738	28.334	27.716	14.937	60.260	0.020	104.400
SOX4	528	7.759	13.846	0.063	2.135	11.998	11.998	14.379	5.375	11.910	1.950	12.200
SP1	513	7.303	16.601	7.665	6.247	9.046	8.426	7.994	3.291	3.290	6.870	18.900
SREBF2	7451	2.611	8.957	84.859	70.533	6.431	3.716	1.655	2.549	5.450	60.260	9.500
SRF	790	10.560	30.035	17.672	14.304	26.855	8.985	10.632	10.479	21.140	11.910	66.800
stat1_common	1154	0.911	1.743	13.717	10.893	1.491	1.007	1.712	0.416	5.160	3.290	2.200
STAT2	1235	6.814	12.711	2.849	6.198	9.714	7.422	8.864	4.175	8.570	5.450	19.700
stat3_common	7640	1.776	5.084	0.000	0.000	2.421	0.960	10.924	1.889	3.290	21.140	5.000
STAT5A	396	1.776	5.084	1.113	1.393	2.421	0.960	10.924	1.889	3.290	5.160	5.000
STAT5B	1717	5.623	9.242	15.297	12.872	7.708	2.932	5.653	3.924	5.400	8.570	13.600
TAF1_sp1	270	3.949	7.678	3.525	6.240	8.286	1.903	4.494	3.247	10.600	3.290	16.900
TAF1_sp2	326	3.949	7.678	3.525	6.240	8.286	1.903	4.494	3.247	10.600	3.290	16.900
TCF12_common_T	2427	3.949	7.678	8.493	6.724	8.286	1.903	4.494	3.247	10.600	5.400	16.900
TCF3_sp2_T	1214	3.949	7.678	0.000	10.253	8.286	1.903	4.494	3.247	10.600	10.600	16.900
TCF3_sp3_T	1102	3.949	7.678	0.000	10.253	8.286	1.903	4.494	3.247	10.600	10.600	16.900
TCF3_sp4	1153	3.949	7.678	0.000	10.253	8.286	1.903	4.494	3.247	10.600	10.600	16.900
TCF3_sp5_T	1031	0.000	0.000	0.000	10.253	0.338	0.000	0.000	0.031	0.090	10.600	0.000
TCF3_sp6_T	1111	0.000	0.000	0.000	10.253	0.158	0.000	0.000	0.161	0.110	10.600	0.400
TCF3_sp7_T	1646	0.000	0.000	0.000	10.253	0.000	0.000	0.000	0.000	0.000	10.600	0.000
TNFRSF13B	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	2.110	0.090	0.000
TNFRSF13C	22	3.584	5.177	0.203	0.125	4.313	2.951	2.831	2.270	3.250	0.110	13.500
TNFRSF17_sp1	0	11.013	13.888	0.000	0.000	17.687	17.687	17.874	5.461	10.590	0.000	10.100
TNFSF13B	84	0.000	0.000	0.008	0.018	0.000	0.000	0.000	0.000	0.170	2.110	0.000
USF1_common_T	623	7.432	13.414	5.824	4.726	12.480	9.747	0.000	5.349	17.100	3.250	11.200
YY1	11801	13.182	12.261	22.801	18.355	15.642	8.019	4.737	10.275	18.420	10.590	29.500
ZIC3	46	13.182	12.261	0.000	0.000	15.642	8.019	4.737	10.275	18.420	0.170	29.500
ZNF217_sp3_T	1401	9.553	12.499	20.481	15.954	13.868	6.016	12.906	6.902	21.380	17.100	54.300

Nature Methods: doi:10.1038/nmeth.2714

Supplementary Table 10. Summary of transcript reconstruction tools

Method	URL	Main application	Additional features
AUGUSTUS[1,2]	bioinf.uni-greifswald.de/augustus	Gene prediction, genome annotation	Can incorporate external expression data (e.g. SAGE or CAGE). Can make use of protein homology information.
Cufflinks[3]	cufflinks.cbcb.umd.edu	Transcript assembly and quantification	Can be run without or without gene annotation, or optionally applied to quantify known transcripts. Can correct for fragment bias and improve transcript quantification if provided with estimated pre-mRNA levels.
Exonerate[4]	www.ebi.ac.uk/~guy/exonerate	Transcript assembly and quantification	
GSTRUCT	—	Transcript assembly and quantification	
iReckon[5]	compbio.cs.toronto.edu/ireckon	Transcript assembly and quantification	Can identify pre-mRNAs and retained introns. Can incorporate external expression data (e.g. SAGE or CAGE).
mGene[6,7]	mgene.org	Gene prediction, genome annotation	
mTim	galaxy.raetschlab.org	Transcript assembly and quantification	Can use external gene expression, protein, RepeatMasker or sequence conservation data. Can identify open reading frames and thus predict coding exons.
NextGeneid[8]	—	Gene prediction, genome annotation	
Oases[9]	www.ebi.ac.uk/~zerbino/oases	De-novo transcriptome assembly	
SLIDE[10]	sites.google.com/site/jingyijli/SLIDE.zip	Transcript assembly and quantification	Can be applied to independent transcript quantification. Can incorporate external expression data (e.g. SAGE or CAGE).
Transomics	www.softberry.com	Gene prediction, genome annotation	
Trembly	—	Transcript assembly and quantification	
Tromer[11]	tromer.sourceforge.net	Transcript assembly and quantification	
Velvet[12]	www.ebi.ac.uk/~zerbino/velvet	De-novo genome assembly	

Nature Methods: doi:10.1038/nmeth.2714

Supplementary Note: Description of transcript reconstruction protocols

This document provides details about specific transcript reconstruction methods. For methods that underlie several protocols, subheadings designate procedural variants. All protocols used the same reference genome sequences: *H. sapiens* assembly GRCh37, *D. melanogaster* release 5 from the Berkeley Drosophila Genome Project, and *C. elegans* assembly WS200. See also Supplementary Table 1, where details of the protocols are tabulated, including the alignment programs used to map RNA-seq reads to the genome sequences.

1. AUGUSTUS

The gene finder AUGUSTUS was initially built to predict gene structures from genomic sequences alone[1]. An extended version of the generalized hidden Markov model used by AUGUSTUS was developed to incorporate evidence from external sources, such as syntenic genomic sequences or expressed sequence tags[2]. Such data is used to inform the detection of start and stop codons, acceptor and donor splice sites, and exonic regions. For each triple of a genomic sequence, a gene structure, and a set of hints, AUGUSTUS assigns a joint probablility and then finds the gene structure that maximizes the posterior probability. The software is available at http://bioinf.uni-greifswald.de/augustus.

1.1. AUGUSTUS all

AUGUSTUS was run using gene expression evidence generated from RNA-seq data. All reconstructed transcripts are reported.

1.2. AUGUSTUS high

AUGUSTUS was run using gene expression evidence generated from RNA-seq data. Only genes with RPKM > 0 are reported.

1.3. AUGUSTUS de-novo

AUGUSTUS was run purely on the genomic sequences. No RNA-seq information is provided.

1.4 Additional features

AUGUSTUS can optionally make use of protein homology information to identify coding genes. Other types of external transcriptomic information can also be incorporated, such as SAGE or CAGE data.

2. Cufflinks

Cufflinks is designed to reconstruct transcripts using RNA-seq reads mapped to the genome with the aligner TopHat[3], but can also process output from other spliced aligners. A overlap graph is generated based on the read alignments, including both spliced and unspliced mappings. Reads that are incompatible, i.e. must have originated from different transcript isoforms, are not connected in the graph, whereas compatbile reads are connected. Paths through the graph correspond to different transcript isoforms. The Cufflinks algorithm aims to find a minimal set of paths that covers all fragments, by searching for the largest set of reads with the property that no two of them could have originated from the same transcript isoform.

Here, Cufflinks version 2.0.2 was used together with TopHat version 2.0.3. Both programs were executed with intron settings tailored to the characteristics of each species: minimum intron lengths were set to 30, 40, and 50 bp for *C. elegans*, *D. melanogaster*, and *H. sapiens*, respectively. The software can be obtained from http://cufflinks.cbcb.umd.edu.

2.1 Additional features

Cufflinks can optionally be provided with genome annotation to improve transcript assembly. Two modes of operation are implemented: the first will use annotation as a guide, but novel transcript isoforms are still predicted; the second mode uses RNA-seq data solely to quantify annotated transcripts. Cufflinks offers additional options to correct for fragment bias and improve transcript assembly by estimating the expected pre-mRNA fraction within the sample.

3. Exonerate SM

This method, which is based on Exonerate[4], was developed by Steve Searle and colleagues at the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk). Briefly, sequencing reads are aligned to the genome and processed to build approximate transcript models. This is followed by a refinement stage, where reads are realigned against the models using a method that takes splicing signals into account. Exonerate SM is currently unreleased.

3.1. Exonerate SM high

This set of transcripts contains the highest scoring model for each locus only.

3.2. Exonerate SM all

This output contains additional alternative isoforms.

4. GSTRUCT

The GSTRUCT pipeline was developed by Thomas Wu and colleagues at Genentech. GSTRUCT uses bounded graph analysis to assemble transcripts based on RNA-seq read mappings produced with the aligner GSNAP[13]. GSTRUCT is currently unreleased, but GSNAP can be obtained from http://research-pub.gene.com/gmap.

5. iReckon

The iReckon algorithm first uses spliced alignments, and if applicable annotated introns, to build a splice graph[5]. All possible transcript isoforms are then identifed by enumerating paths from each of the possible transcription start sites to the end sites. For each putative isoform, the sequence is extracted and reads are realigned using the program BWA[14]. Finally, expressed isoforms and their abundances are predicted by a regularized expectation maximization algorithm, which penalizes low-abundance isoforms. iReckon also reports pre-spliced mRNAs and isoforms with retained introns.

iReckon version 1.0.7 was applied using initial alignments from TopHat version 2.0.3 and BWA version 0.6.2 for realignment. These software components are available from http://compbio.cs.toronto.edu/ireckon, http://tophat.cbcb.umd.edu and http://bio-bwa.sourceforge.net, respectively.

5.1 iReckon full

The program was provided with complete annotation for protein coding genes. Unspliced or retained intron transcripts and were removed from the program output so as not to bias evaluations based on protein-coding annotation.

5.2 iReckon ends

The program was provided with transcript boundary coordinates (start and end coordinates, but not intron information) from the reference annotation used for the evaluation. Unspliced transcripts were removed from the program output so as not to bias evaluations based on protein-coding annotation.

5.3 Additional features

iReckon can predict pre-mRNAs and retained introns from RNA-seq data. The software distribution also includes a plug-in for the Savant Genome Browser to visualize read assignment to transcript isoforms.

6. mGene and mTim

The protocols mGene, mGene graph and mTim are based on combinations of the individual programs PALMapper[15], mGene[6,16], mTim, SplAdder and rQuant[17], all developed by the same group.

PALMapper was used to align the RNA-seq reads by allowing spliced and unspliced alignments. The program considers base call quality scores and computational splice site predictions during alignment. Alignments were filtered with different settings for mGene and mTim (see below).

The basic mGene algorithm uses a two-layered machine learning approach. The first layer employs support vector machine models to scan genomic sequence for transcription start and stop sites, translation start and

stop sites, and splice donors and acceptors. The second uses hidden semi-Markov support vector machines to combine those features into valid coding gene predictions. While the fundamental strategy relies only on genomic sequence, mGene can also include information from features tracks in the scoring function. Various tracks were used in this protocol, mostly from RNA-seq alignments. The balance between signal predictions and feature tracks is optimized during training. This extension of mGene has recently been documented.

In contrast to mGene, mTim uses a simpler hidden Markov support vector machine approach, in which states directly correspond to intergenic, exonic and intronic nucleotides with a certain expression level (five submodels were used, each corresponding to an expression quintile). Based on features derived from the RNA-seq alignments, the most likely state is inferred for each nucleotide, taking context-dependencies into account (in the form of a state-transition model). The model does not distinguish between coding and non-coding regions. The parameters of the model are trained on a small subset of the reference gene annotation.

SplAdder builds a splice graph based on initial predictions and RNA-seq evidence. The splice graph is then used to generate possible transcript isoforms. For each transcript SplAdder determines the maximal open reading frame to predict coding regions.

Expression levels of predicted isoforms were estimated using the program rQuant that can take fragment biases into account. Those transcripts that scored low using an SVM classifier were removed from the prediction set. The SVM was trained on a small fraction of the genome annotation using estimated abundance, length, coding sequence length and number of exons as features.

The programs PALMapper, mGene, mTim, SplAdder, and rQuant can be obtained from http://raetschlab.org/suppl/rgasp2.

Based on these algorithms, prediction sets were created as follows:

6.1. mGene

For each organism mGene was trained with multiple features from PALMapper RNA-seq alignments, as well as other genomic features, output from SVM-based signal predictors that were previously trained on a part of the annotation. The RNA-seq based features included exon coverage, intron coverage (number of spliced reads spanning a given position, indicating that this position may be part of an intron) and intron lists including the count of supporting reads. Alignments were filtered by excluding reads with more than one mismatch, fewer than eight aligned nucleotides in any exon flanking a spliced alignment, or a spliced alignment indicating and intron longer than 20 kb (100 kb for human).

For human, repeat elements identified by RepeatMasker were included, in addition to other sequence-based features. Similar to exon and intron coverage tracks, 15 additional tracks were included: "DNA", "LINE", "Low_complexity", "LTR", "Other", "RC", "tRNA", "Satellite", "Simple_repeat", "SINE", "Unknown", "rRNA", "scRNA", "snRNA" and "RNA".

6.2. mGene graph

As the mGene protocol above, with the addition that alternative transcripts were predicted using SplAdder, subsequently quantified using rQuant and filtered using the SVM classifier.

6.3. mTim

RNA-seq alignments generated by PALMapper were filtered specifically for each organism. Spliced alignments for *C. elegans*, *D. melanogaster* and *H. sapiens* were filtered out if the minimal segment length within an alignment was shorter than 15nt, 20nt, or 15nt, respectively. A general threshold of at most one edit operation was applied to all alignments for all organisms.

mTim was trained with multiple features from RNA-seq alignments as well as splice signals from genomic sequence (based on an SVM classifier previously trained on a subset of the annotation). Features derived from the RNA-seq alignments included exon coverage, intron coverage (number of spliced reads spanning a given position), scores for acceptor and donor splice sites deduced from spliced alignments (number of

spliced alignments with a given junction and alignment confidence scores) as well as mate-pair coverage (number of read pairs with an insert spanning a given position). SplAdder was applied to the raw mTim transcript predictions to generate alternative transcripts, which were quantified by rQuant and filtered using the SVM approach.

7. NextGeneid

NextGeneid is a modified version of Geneid (version 1.3)[8]. Geneid identifies splice sites and start/stop codons from genomic sequence. These features are then combined to predict exons, which are scored based on supporting features and coding potential. From the set of predicted exons, gene structures are assembled. NextGeneid additionally incorporates RNA-seq read alignments to the genome, produced with the GEM mapper[18], including spliced alignments from the GEM component gem-split-mapper. Read alignments were used to modify the scores of potential exons, determine transcript start and end coordinates, and constrain the exon-chaining algorithm in Geneid based on spliced alignments. NextGeneid has not been released.

7.1. NextGeneid

Geneid was previously trained on several species, including the three used in this study. No modifications to the signal or coding potential position weight arrays were performed. Transcripts reported by NextGeneid with RPKM > 1 were retained.

7.2. NextGeneidAS

As NextGeneid, but iterated to increase intron detection sensitivity.

7.3. NextGeneidAS de-novo

As NextGeneid, with the addition of *ab initio* Geneid predictions falling within intergenic space to the final set of transcripts.

8. Oases

Oases assembles transcripts from RNA-seq data without using genomic sequence[9]. It is based on the short-read assembler Velvet[12], adapting several steps to the different characteristics of RNA-seq data, such as uneven coverage across transcripts and the expression of alternative isoforms. Reads of low quality were first removed from the data and low-quality bases were trimmed from from both ends of reads. Velvet was then used to build a de Bruijn graph from the RNA-seq data using a *k*-mer size of 33. Mate pair and coverage information was used to predict and assemble transcripts using Oases v0.1. The resulting transcripts were then aligned to each genome using BLAT[19]. Oases is available at http://www.ebi.ac.uk/~zerbino/oases.

9. SLIDE

SLIDE is a stochastic method based on a linear model with a design matrix that computes the sampling probability of RNA-seq reads from different transcript isoforms[10]. It utilizes exon boundary information from annotations to enumerate all possible isoforms. Discovery of expressed isoforms is implemented as a sparse estimation problem, related to the number of isoforms that are expected to be expressed. Sparse estimation is achieved by a modified lasso method[20]. SLIDE is available at https://sites.google.com/site/jingyijli.

9.1. SLIDE all

All transcript isoforms reported by SLIDE.

9.2. SLIDE high

The subset of isoforms identified as “high confidence” by SLIDE.

10. Transomics

The Transomics pipeline is based on the gene finding pipeline Fgenesh++[21], extended to incoporate splice site information from RNA-seq read mappings to the genome. The relative abundance of alternative transcripts generated from the same gene locus is estimated using a solution of a system of linear equations. Further details are available at http://linux5.softberry.com/cgi-bin/berry/programs/Transomics.

10.1. Transomics all

All predicted genes are reported.

10.2. Transomics high

Only transcripts with RPKM > 0.02 are reported.

11. Trembly

Trembly is an unpublished software package for transcript reconstruction from RNA-seq data developed in Mark Gerstein's group at Yale University (http://www.gersteinlab.org). Trembly was applied to RNA-seq reads aligned with TopHat. A signal track of mapped reads is generated and a set of transcriptionally active regions (TARs) is identified. Splice junctions are then inferred from adjacent TARs. Both the predicted splice junctions and TARs are provided as input for transcript assembly, which generates all possible transcript isoforms compatible with the data. Expression levels of predicted transcripts were estimated using the program IQSeq developed by the same group[22].

11.1. Trembly all

The full output from Trembly.

11.2. Trembly high

The subset of transcripts with RPKM above 0.1.

12. Tromer

The Tromer pipeline first maps reads to the genome using fetchGWI to identify unique exact matches[11]. MegaBLAST is used to recover unmapped reads. In a third step, spliced alignment is carried out with SIBsim4, taking mate pair information into account. The output of these three steps is combined to create graphs representing all possible alternative splice variants of a gene. A greedy algorithm is applied to the graphs, designed to output a set of transcripts such that each edge is covered at least once. The algorithm proceeds in three steps: 1) select a seed edge, 2) extend toward the 5´ end, and 3) extend toward the 3´ end. The seed edge is first selected among unused 5´-most exons, and then among remaining unused edges. The extension process attempts to include unused edges derived from the same read pair as the seed edge. Further details are available at http://tromer.sourceforge.net.

13. Velvet

These protocols are based on the genome assembly program Velvet[12]. Transcripts assembled by Velvet were mapped to the respective genomes using BLAT[19]. Exons were quantified using ERANGE[23]. These programs are available from http://www.ebi.ac.uk/~zerbino/velvet, http://genome.ucsc.edu (BLAT) and http://woldlab.caltech.edu (ERANGE).

13.1. Velvet

This protocol corresponds to the Velvet pipeline outlined above.

13.2. Velvet + Augustus

Transcripts structures assembled by the Velvet pipeline were provided to AUGUSTUS as evidence. The final set of transcripts consisted of AUGUSTUS models that agreed with the original isoforms from Velvet over more than 25% of their length.

References

1. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34,** W435–9 (2006).
2. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7,** 62 (2006).

3. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27,** 2325–2329 (2011).
4. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 31 (2005).
5. Mezlini, A. M. *et al.* iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res.* **23,** 519–529 (2013).
6. Schweikert, G. *et al.* mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Res.* **19,** 2133–2143 (2009).
7. Schweikert, G. *et al.* mGene.web: a web service for accurate computational gene finding. *Nucleic Acids Res.* **37,** W312–6 (2009).
8. Blanco, E., Parra, G. & Guigo, R. Using geneid to identify genes. *Curr Protoc Bioinformatics* **18**, 4.3.1–4.3.28 (2007).
9. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28,** 1086–1092 (2012).
10. Li, J. J., Jiang, C.-R., Brown, J. B., Huang, H. & Bickel, P. J. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 19867–19872 (2011).
11. Sperisen, P. *et al.* trome, trEST and trGEN: databases of predicted protein sequences. *Nucleic Acids Res.* **32,** D509–11 (2004).
12. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
13. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26,** 873–881 (2010).
14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
15. Jean, G., Kahles, A., Sreedharan, V. T., De Bona, F. & Rätsch, G. RNA-Seq read alignments with PALMapper. *Curr Protoc Bioinformatics* **32**, 11.6.1–11.6.37 (2010).
16. Gan, X. *et al.* Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477,** 419–423 (2011).
17. Bohnert, R. & Rätsch, G. rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res.* **38,** W348–51 (2010).
18. Marco-Sola, S., Sammeth, M., Guigo, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9,** 1185–1188 (2012).
19. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
20. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Statist. Soc.* **B 58,** 267-288(1996).
21. Solovyev, V., Kosarev, P., Seledsov, I. & Vorobyev, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7** Suppl 1, S10.1–12 (2006).
22. Du, J. *et al.* IQSeq: integrated isoform quantification analysis based on next-generation sequencing. *PLoS ONE* **7,** e29175 (2012).
23. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).