Illuminating the dark side of the human transcriptome with long read transcript sequencing

Richard Kuo (richard.kuo@roslin.ed.ac.uk)
The University of Edinburgh The Roslin Institute https://orcid.org/0000-0002-7867-7594
Yuanyuan Cheng
University of Sydney https://orcid.org/0000-0002-1747-9308
Runxuan Zhang
The James Hutton Institute https://orcid.org/0000-0001-7558-765X
John W.S. Brown
University of Dundee
Jacqueline Smith
The University of Edinburgh The Roslin Institute https://orcid.org/0000-0002-2813-7872
Alan L. Archibald
The University of Edinburgh The Roslin Institute https://orcid.org/0000-0001-9213-1830
Dave W. Burt
University of Queensland https://orcid.org/0000-0002-9991-1028

Research article

Keywords: dark side, human transcriptome, long read transcript sequencing

Posted Date: April 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-23156/v1

Version of Record: A version of this preprint was published on October 30th, 2020. See the published version at https://doi.org/10.1186/s12864-020-07123-7.

Abstract

Background The human transcriptome annotation is regarded as one of the most complete of any eukaryotic species. However, limitations in sequencing technologies have biased the annotation toward multi-exonic protein coding genes. Accurate high-throughput long read transcript sequencing can now provide stronger evidence for genes that were previously either undetectable or impossible to differentiate from sequencing noise such as rare transcripts, mono-exonic, and non-coding genes.

Results We analyzed Sequel II Iso-Seq sequencing data of the Universal Human Reference RNA (UHRR) using the Transcriptome Annotation by Modular Algorithms (TAMA) software. We found that the convention of using mapping identity to measure error correction performance does not reflect actual gain in accuracy of predicted transcript models. In addition, inter-read error correction leads to the thousands of erroneous gene models. Using genome assembly based error correction and gene feature evidence, we identified thousands of potentially functional novel genes.

Conclusions The standard of using inter-read error correction for long read RNA sequencing data could be responsible for genome annotations with thousands of biologically inaccurate gene models. More than half of all real genes in the human genome may still be missing in current public annotations. We require better methods for differentiating sequencing noise from real genes in long read RNA sequencing data.

Introduction

The transcriptome remains a vastly underexplored space despite its significance as a foundation for biology. Major challenges for transcriptome annotation of eukaryotic species stem from biological complexity, RNA preparation, limitations of sequencing technologies, and sequence analysis. The biological complexity of alternative transcription start/stop sites and splice junctions(1) results in a combinatorial array of transcript sequences(2). To complicate matters, RNA samples collected from eukaryotic species contain a mixture of mature functional RNA as well as pre-processed RNA, degraded RNA, and possible genomic contamination(3) (Fig. 1a-b). Meanwhile, low-throughput cDNA sequencing(4) fail to provide coverage for rare/unstable transcripts, while short read RNA sequencing (RNAseq) present computational challenges in accurate transcript reconstruction(5). The ambiguities created by these combined factors forced previous annotation software to adopt conservative algorithms that filtered out many real transcripts/genes such as single exon genes and long non-coding RNA (lncRNA).

High-throughput long read transcript sequencing provides higher confidence in predicting alternative transcripts and distinguishing real transcripts from sequencing noise(6). While there have been many studies using long read transcript sequencing for transcriptome discovery(7), their sensitivity and accuracy were compromised by the use of problematic processing strategies: orthogonal verification/filtering and inter-read error correction. Filtering transcript models based on orthogonal information, such as requiring gene models to have sequence homology to annotated genes from closely related species, reduces gene discovery and is only applicable for a small number of species where such information exists(8). Inter-read error correction relies on correctly grouping reads that originate from the same transcript and using the consensus of the alignment between grouped reads as the corrected sequence. However, lower quality reads have an increased probability of grouping with reads from different transcripts from the same or paralogous genes which could introduce erroneous hybrid sequences via the

correction process. This type of error occurs when the alignment of reads is compromised by regions of high error density (Fig. 1c).

The standard pipeline for processing Iso-Seq data involves intra-read error correction (Circular Consensus Sequences - CCS), removing adapter sequences, and removing poly-A tail sequences to create Full Length Non-Chimeric (FLNC) read sequences (Fig. 1d). This is followed by an inter-read error correction step (Cluster/Polish(9)) to create an amalgamated read sequence with lower error rates. Cluster/Polish differs from the intra-read correction since the sub-reads used in CCS are known to be from the same transcript. While Cluster/Polish has been shown to improve the read mapping rate and identity, the risk of introducing erroneous hybrid sequences and other errors can outweigh the benefits.

To leverage the power of long read transcript sequencing and address the issues with current processing pipelines, we developed the Transcriptome Annotation by Modular Algorithms (TAMA) tool kit. TAMA is designed to both improve transcript model prediction and increase transcriptome discovery. TAMA uses long read transcript data and high-quality reference genome assemblies to produce accurate and informative transcript models. This makes TAMA useful for situations where additional types of data, such as public annotations or short read RNAseq, are not available(10). In addition, by not relying on orthogonal information, TAMA also provides a more agnostic approach to transcriptome annotation which can reveal problems with prior assumptions from previous annotation efforts.

To understand the extent of bias in the human reference annotation and the limitations of currently available long read transcript sequencing analysis software, we analyzed the Universal Human Reference RNA (UHRR) Sequel II Iso-Seq data released by Pacific Biosciences (PacBio) using TAMA.

Results

Standard pipelines filter out thousands of potential novel gene predictions

We processed the UHRR Iso-Seq data using 5 different pipelines to compare the effect of each method on gene discovery and model prediction accuracy (Fig. 1e). The first pipeline, which we will refer to as the “Cupcake pipeline”, uses inter-read error correction (in the form of clustering long reads and using the alignment to polish the sequences) along with Cupcake Collapse (a tool for collapsing redundant transcript models). The Cupcake pipeline represents the current standard pipeline, as is run by default on PacBio machines. The second pipeline, which we will refer to as the “Polish pipeline”, is a modified version of the Cupcake pipeline where TAMA Collapse is used in place of Cupcake Collapse. The third pipeline, which we will refer to as the “Lordec pipeline”, uses inter-read error correction in the form of aligning short read RNAseq data to long reads along with TAMA Collapse. For the Lordec pipeline we used short read RNA-seq data from the UHRR but from another study(11). The fourth pipeline, which we will refer to as the “TAMA Low pipeline”, uses no inter-read error correction. The settings used for the TAMA Collapse run in TAMA Low are also the same settings used in the Polish and Lordec pipelines. The fifth pipeline, which we will refer to as the “TAMA High pipeline”, is identical to the TAMA Low pipeline with the exception of the use of the local density error (LDE) feature which is currently unique to TAMA Collapse. In the TAMA High pipeline, the LDE algorithm is used to remove transcripts with more than 1 error within a 20 bp range of a splice junction. High densities of errors near the splice junctions often cause erroneous splice junction mapping

leading to inflated transcript numbers. We allowed for 1 bp of error due to possible true genomic variation between the UHHR samples and the reference genome assembly. We then only kept transcript models with read support from both SMRT cells (the Iso-Seq data was generated from 2 SMRT cells) to avoid using reads that originated from PCR artefacts.

Out of the 5 pipelines, the Polish and Cupcake pipelines had the lowest sensitivity in terms of the number of predicted genes (25,731 and 25,239 genes, respectively) and transcripts (126,288 and 128,389, respectively) (Table 1). As expected, due to the similarity of algorithms between low stringency TAMA Collapse and Cupcake collapse, these pipelines produced similar numbers of genes and transcripts. The primary reason for the reduced sensitivity in these pipelines is that the Cluster/Polish steps remove any reads that do not cluster with at least one other read (also known as singletons). Thus, low expression transcripts can be filtered out due to a lack of read depth.

The TAMA Low and Lordec pipelines resulted in the greatest sensitivity in terms of the number of genes (168,328 and 166,766 genes, respectively) and transcripts (752,996 and 753,756 transcripts, respectively) predicted (Table 1). The high sensitivity of the Lordec pipeline is not surprising as Lordec does not filter any reads. Thus, the same number of reads were used during genome mapping for the Lordec pipeline as compared to the FLNC Low pipeline and the same TAMA Collapse parameters were used in both pipelines. In theory, the Lordec pipeline should have resulted in greater numbers of genes, since short read error correction is supposed to correct low quality Iso-Seq reads which should then increase the number of mapped reads that pass the filtering thresholds of TAMA Collapse. Surprisingly, we do not see a significant increase in sensitivity from using Lordec.

Table 1 Pipeline Comparison

Comparison of gene and transcript number across pipelines broken down into different categories. Gene level matches refer to gene models which overlap on the same strand. Transcript level matches refer to transcript models with similar exon structures.

Match Type	Polish	Cupcake	Lordec	TAMA Low	TAMA High
Total Genes	25,731	25,241	166,766	168,328	40,821
Total Transcripts	126,288	128,389	753,756	752,996	135,218
Gene Level Matches	19,348	19,342	30,835	30,947	21,284
Transcript Level Matches	20,364	20,428	28,143	28,234	17,932
Novel Genes	8,519	8,068	139,769	141,097	23,302
Novel Transcripts	103,052	104,824	716,241	713,546	115,893

The TAMA High pipeline resulted in gene (38,743) and transcript (135,218) totals that were greater than the Polish and Cupcake pipeline but not nearly as high as the TAMA Low and Lordec pipelines.

To test whether the number of genes identified in the 5 pipelines correlated to true sensitivity (as opposed to erroneous gene predictions from sequencing noise), we compared the resulting annotations from the five pipelines with the Ensembl v96 human annotation(12) using TAMA merge (Fig. 2a-b). The Ensembl human annotation is created by a combination of a semi-automated gene build pipeline and manual annotation by the Havana team.

As a result, the Ensembl human annotation is considered among the most complete and accurate annotations for the human genome. We used two definitions for matching: gene-level matches and transcript-level matches. Gene-level matches are defined as the number of genes with overlap between gene loci (between each pipeline and the Ensembl annotation) on the same strand. Gene level matches represent the upper bound estimate of the number of true known gene models identified in each pipeline. Transcript-level matches are defined as the number of transcripts that have the same exon structures between annotations (see Methods). These represent an estimate of the accuracy of the transcript models predicted by each pipeline since an erroneous transcript model would still be considered a match on the gene level.

The pipelines which predicted the highest number of genes (TAMA Low and Lordec) also resulted in the highest number of both gene-level and transcript-level matches. This suggests that the increased stringency of the other pipelines significantly decreases the detection of real genes and transcripts. However, both the TAMA Low and Lordec pipelines had over 140K mono-exonic genes which is more than 5 times the number of mono-exonic genes as the Polish pipeline (Fig. 2c). This vastly greater number of mono-exonic genes could represent either real genes, RNA sample noise or DNA contamination. Therefore, the increased sensitivity of these pipelines could come at the cost of having a significant number of possible false positives.

The TAMA High pipeline had more gene-level matches as compared to the Polish and Cupcake pipelines but fewer transcript-level matches. This may be because the TAMA High filtration of transcript models with only single SMRT cell read support removed reads representing lowly expressed transcripts. However, the number of mono-exonic genes within the TAMA High annotation appeared to be more reasonable. Therefore, it appears that the TAMA High pipeline resulted in a reasonable balance between sensitivity and specificity.

TAMA HIGH identifies more accurate transcript models without reducing gene discovery

To understand the accuracy of each pipeline for predicting transcript models, we looked at both TAMA's predicted error rates as well as splice junction wobble. Wobble refers to mis-mapping of splice junctions causing small differences in the genomic loci of mapped features such as exon boundaries and splice junction donor/acceptor sites (Fig. 3a). While the error rates of mapped reads are often used to assess the improvement of long read data from different pipelines(13), this metric is actually not as useful for understanding the overall improvement in the transcriptome annotation. In genome-based transcriptome annotations, typically the most important features to identify are the transcription start sites (TSS), transcription end sites (TES), splice junctions, and exon chaining. These features allow for predictions of coding and promoter regions which are often crucial for downstream analyses. Thus, for transcript structure identification, errors near the splice junctions have a greater probability of altering the resulting transcript model than errors occurring farther away from the splice junctions. This means that the percentage of errors within a read may not be as impactful as the distribution of errors. Thus a more accurate metric for the performance of error correction methods is to assess the amount of "splice junction wobble" between the predicted transcripts and known transcripts.

To demonstrate this concept we looked at the error profiles for each mapped read for the TAMA Low, TAMA High, Polish, and Lordec pipelines. Note that the mapped FLNC reads are the same for the FLNC High and FLNC Low pipelines. The Cupcake pipeline was omitted in this analysis because Cupcake Collapse does not provide a report

on the errors in the mapped reads whereas TAMA Collapse provides detailed information on the type and amount of errors for each read.

Using the output from TAMA Collapse we looked at length of coverage, mapping identity, clipping, insertions, deletions, and substitution errors. These values represent the comparison of the mapped reads to the genome assembly and thus only serve as an estimate of the true rates of error. We also looked at average error rates between the pipelines by counting the number of base pairs that were not matching between the mapped read and the genome sequence and dividing this number by the length of the mapped read. This includes soft clipping, insertion, deletion, and substitution errors but does not include hard clipping.

The TAMA Low pipeline had the highest average error rate (2.83%) and the highest amount of each type of error while the Polish pipeline had the lowest error rates (0.52%) with the lowest amount of each type of error. The Lordec pipeline coverage and identity were unexpectedly similar to the FLNC pipelines suggesting that Lordec correction did not provide a large gain in error correction. The Lordec pipeline had a similar amount of clipping errors as compared to the FLNC pipelines but lower rates of insertion, deletion, and substitution errors (Fig. 2d). This indicates that the Lordec correction may have some issues correcting the ends of reads.

The Polish pipeline had longer average mapped read lengths compared to other pipelines. This is most likely due to the Cluster/Polish algorithm which merges read sequences with up to 100 bp length differences on the 5’ end. This behavior essentially absorbs the shorter reads into the longer reads effectively removing their length representation.

We then looked at transcript model accuracy by measuring the wobble at splice junctions with respect to transcript models annotated in the Ensembl human annotation for the 5 different pipelines (Fig. 3b and 3d). Wobble typically occurs due to high error density near the splice junctions leading to small shifts in mapping the ends of each exon(14). The amount of wobble between the transcript models of each pipeline compared to the reference annotation provides a metric for the accuracy of the transcript models produced by each pipeline. We ignored wobble at the transcript start and end sites due to the high variance of these features in natural RNA(15),(16). We also only assessed Ensembl transcript models that had coverage from all assessed pipelines to account for the differences in sensitivity between the pipelines.

The TAMA Low pipeline produced the highest average wobble per splice junction (0.72 start wobble and 0.69 end wobble) while the TAMA High pipeline had the lowest average wobble values (0.26 start wobble and 0.24 end wobble).

We also looked at the number of transcripts with perfect splice junction matches to the reference annotation. The TAMA Low pipeline had the lowest number of perfect transcript matches, while the TAMA High pipeline had the highest number of perfect transcript matches (Fig. 3c). The Polish pipeline had the second highest number of perfect transcripts. Thus, despite the lower overall error rates in the mapped reads from the Polish pipeline, the TAMA High pipeline produced more accurate transcript models. This suggests that the LDE filtration in the TAMA High pipeline resulted in more accurate identification of splice junctions.

Inter-read error correction leads to read jumbling affecting thousands of genes and transcripts

One of the major concerns when using inter-read error correction methods such as Cluster/Polish and Lordec is the possibility of creating sequencing errors as a result of combining read sequences from different transcripts. The different transcripts can either be from different genes (gene-level jumble) or a combination of alternative transcripts within the same gene (transcript-level jumble). Gene-level jumble typically occurs due to the sequence similarity of paralogues within gene families(17). In both gene-level and transcript-level jumble, it is more likely that the highest expressed gene or transcript within the read clusters will mask the lower expressed genes. This is because the final cluster sequence is determined by sequence coverage. However, in cases where the read coverage within a jumble cluster is similar across unique transcripts, it is more likely that the resulting cluster read will have a mixture of sequences from each unique transcript within the cluster.

To investigate how often these jumble events occur, we compared the read mappings from the TAMA Low pipeline to the inter-read error correction pipelines to find reads which mapped to different genes and transcripts in each comparison. While it is possible that the TAMA Low read mappings are erroneous, they represent the read sequences without any over-correction. Also reads which have different mapping loci between the TAMA Low pipeline and an inter-read error correction pipeline indicate that there is enough sequence ambiguity to call into question the effect of the inter-read error correction.

Since the Cupcake pipeline uses the same Cluster/Polish step as the Polish pipeline, there should be no differences in the read mappings. Similarly, the TAMA Low and TAMA High pipelines used the same read mappings.

Comparing the TAMA Low pipeline to the Polish pipeline, we found 34,637 reads that switched from one gene locus to another after Cluster/Polish correction (Fig. 4b). This gene loci switching involved 6,774 genes, 3,230 of which were only found with the FLNC Low pipeline while 104 genes were only found with the Polish pipeline. The asymmetry of the number of unique genes between the pipelines suggests that Cluster/Polish is incorrectly clustering reads from different genes.

To gain a more detailed understanding, we focus on the PReferentially expressed Antigen of MElanoma (PRAME) gene family. The PRAME gene family is highly associated with cancer development(18),(19),(20),(21) and is used as a biomarker for identifying various forms of cancer. Within the PRAME gene family there are 24 annotated paralogues(22). In this example, the Polish pipeline caused a false negative for the detection of one of the PRAME paralogues, PRAMEF8. The TAMA Low pipeline finds 9 reads mapping to PRAMEF8 (Fig. 4a) while the Polish pipeline shows no reads mapping to PRAMEF8. Of the 9 PRAMEF8 reads from the TAMA Low pipeline, 5 were mapped to another paralogue, PRAMEF15, in the Polish pipeline. We analyzed the sequence similarity between the two paralogues by aligning the PRAMEF8 and PRAMEF15 transcript sequences with Muscle(23) and found that they had 76% identity. While the sequences of the two genes are similar, the genome mapping identity for the reads were higher than the sequence similarity between the two paralogues. The PRAMEF8 read with the lowest identity score during genome mapping in the TAMA Low pipeline had an identity of 89% and 6 PRAMEF8 reads had mapping identities over 98%. Thus, there is strong evidence that the reads mapped correctly in the TAMA Low pipeline and were altered to the point of mis-mapping in the Polish pipeline. This particular type of error could have major consequences for studies aimed at identifying gene biomarker expression.

To assess the effect of short/hybrid inter-read error correction on gene level read jumbling, we compared the TAMA Low pipeline to the Lordec pipeline. There were 19,064 reads which switched from one gene locus to another (Fig. 4c), involving a total of 3,476 genes, 775 of which were only found with the TAMA Low pipeline while 675

genes were only found with the Lordec pipeline. The number of genes found only in the Lordec pipeline is much higher than the number of genes found only in the Polish pipeline which may indicate that Lordec correction is more inclined to produce false positives.

We also examined how erroneous inter-read error correction can lead to transcript level jumbling. In this case, when reads from different transcripts from the same gene are grouped for error correction, the resulting sequence will, at best, represent only the more highly expressed transcript and, at worst, represent an erroneous jumbled sequence.

Comparing the TAMA Low pipeline to the Polish pipeline, we found 477,351 reads that mapped to different transcript models within the same gene. There were 112,891 transcripts affected by transcript-level jumbling, 44,852 of which were found only in the TAMA Low annotation while 1,372 transcript were found only in the Polish annotation. This represents a large difference between the two annotations with the TAMA Low pipeline predicting far more transcript models than the Polish pipeline.

Comparing the TAMA Low pipeline to the Lordec pipeline, we found 187,829 reads that mapped to different transcript models. This involved 142,704 transcripts with 7,117 transcripts found only in the FLNC Low annotation and 11,732 transcript found only in the Lordec annotation. Again, it appears that the Lordec pipeline is more prone to producing false positives.

To summarize, in both the long and short inter-read error correction pipelines we saw a significant number of gene-level and transcript-level read jumbling which may result in the prediction of gene and transcript models that are not biologically accurate. Hence, we argue that the best approach would be to forego inter-read error correction and instead focus on methods, such as the TAMA Collapse LDE algorithm, for removing reads with error profiles that could lead to erroneous transcript model predictions.

Novel genes are dominated by mono-exonic non-coding and coding genes

To gain insight into the 141,097 novel gene models found in the TAMA Low pipeline, we looked at several features: coding potential, number of exons, intronic overlap with other genes, overlap with regulatory features, and the presence of immediately downstream genomic poly-A stretches. Coding potential and splice junctions are often used as strong evidence of a functional gene. Conversely, overlap with introns (from other genes), genomic poly-A stretches immediately downstream of a gene model, and the absence of splice junctions (single exon transcripts) provide evidence that the source of the model could be from either non-functional transcribed products or genomic contamination.

Coding potential was assessed using three complementary methods. First, we used an open reading frame sequence analysis tool, CPAT(24), to detect coding potential. This method only works when the transcripts models do not contain frame shifts caused by erroneous splice junction calling. Second, we used TAMA merge to identify gene models that overlapped the genomic loci (on the same strand) of protein coding genes within the Ensembl annotation. Third, we used the TAMA ORF/NMD pipeline which is a frame shift-tolerant method of matching transcript sequences to peptide sequences from the Uniprot(25) database. We combined these three methods to account for the various errors which can cause false negatives in protein coding gene prediction.

Only a small number of the novel genes (122 out of 141,097) were supported by all features which are considered evidence for functionality (multi-exonic, coding, intergenic, and processed poly-A) (Fig. 5a). This is expected given that these features are used by short read RNA-seq annotation pipelines for validation. Therefore, many of the gene models with these features are likely to have already been annotated by Ensembl.

The two most common sets of features for the TAMA Low predicted novel genes are “single exonic, non-coding, intronic gene overlap, and genomic poly-A” at 32% (45,820) and “single exonic, coding, intronic gene overlap, and genomic poly-A” at 20% (27,524). The first set of features are indicators for lncRNA while the second set of features are indicators for processed pseudogenes or real coding genes. Together, these account for over 52% of the novel genes. While these features are also associated with non-functional sequences such as pre-processed RNA or degraded RNA fragments, the fact that they do not overlap with the exons of known genes suggests that many of these putative novel gene models may represent real genes which were overlooked or undetected in previous annotations efforts.

The third most common set of features (“single exonic, non-coding, intronic gene overlap, and processed poly-A”) are indicators for real mono-exonic lncRNA and make up 10% (14,036) of the novel gene models (Fig. 5a). Since there are no genomic poly-A stretches downstream of these models, the source RNA must have had poly-A tails added during RNA processing which would suggest the RNAs truly exist in that form and may have functional roles. Given that these models did not overlap any gene models in the Ensembl annotation, this would represent a large increase in the number of predicted lncRNA for the human genome.

There were an additional 11,255 (8%) genes with features (single exonic, non-coding, intergenic, and genomic poly-A) that are indicative of either mono-exonic intergenic lncRNA or genomic contamination. With the UHRR being one of the most carefully prepared RNA samples, this would indicate that researchers would require more advanced methods of either RNA preparation or sequencing analysis to confidently identify novel genes.

The Universal Human Reference RNA contains a significant amount of degraded RNA

We developed a metric called the “Degradation Signature” (DegSig) to measure the 5’ transcript completeness. When both a capped selected cDNA library and non-capped selected cDNA library are available on the same sample, we can measure the amount of multi-read-supported, multi-exonic transcripts at a given loci that show a cascading pattern of TSSs along the length of the longest transcript model (Fig. 5b). A higher percent difference indicates that the sample RNA had a greater proportion of degraded RNA.

The DegSig of the Ensembl human reference annotation is 1.5% which represents an estimate of the lower limit for DegSig values. However, since 5’ degraded transcripts are difficult to distinguish from real TSS cascade pattern gene models, it may be that these types of gene models are under-represented in all annotations since most annotation pipelines remove 5’ shorter transcripts models. This analysis supports the idea that a DegSig of 0% is likely impossible due to true TSS cascade models.

We applied the DegSig metric to chicken brain Iso-Seq data, where libraries were prepared using both a non-cap selecting method and the TeloPrime(26) 5’ cap selection. The TeloPrime library should contain a lower percentage of degraded transcript sequences since it selects for complete capped RNAs. The non-cap selected data had a DegSig of 56.3% while the DegSig for the TeloPrime library data was 23.6%, suggesting a large difference in the proportion of degraded RNA sequences captured as cDNA by the two different methods.

We ran DegSig on the UHRR Iso-Seq dataset individually by SMRT cell and chromosome. Almost all chromosomes had a DegSig between 32% and 41% (Fig. 5c). However, the Y chromosome had a DegSig of 26.7% and 27.2% for SMRT Cell 1 and 2, respectively. One explanation for the much lower DegSig on the Y chromosome may be due to the lack of read depth for the Y chromosome (only 629 and 588 reads from SMRT cells 1 and 2, respectively). Lower read depths can decrease the DegSig values due to the lack of coverage for each gene.

The range of DegSig for the human data is higher than that for the chicken 5' cap selected RNA data, suggesting that there may be a significant number of truncated models in the UHRR Iso-Seq transcript annotation. This could also be a source of the unusually high number of novel alternative transcripts predicted in the FLNC Low pipeline.

Discussion

The UHRR PacBio Sequel II Iso-Seq dataset is the result of the most accurate high-throughput long read transcript sequencing technology applied to an RNA library used as a reference for gene profiling experiments. Thus, this dataset represents the technological limits and challenges that are pertinent to all RNA sequencing studies as well as the potential of long read transcript sequencing for discovering novel genes and isoforms. The resulting transcriptome annotation portrays a very different composition of gene models compared to public transcriptome annotations. These differences include over 140 thousand potential novel genes many of which are classified within under-represented biotypes. This raises questions regarding what exactly is present in our sequencing data and what is the best way to further dissect this information to produce biologically meaningful results.

There has been a heavy emphasis on the use of multi-omics or orthogonal data to identify what is real and functional within the transcriptome. While this is certainly a powerful means of investigating novel genes, the pipelines developed for this purpose often overlook the need to properly process individual sources of data before integrating across data types. Using the TAMA tool kit, we have demonstrated some key issues with current long read RNA data pipelines that could have major effects on current transcriptomic studies.

As can be seen in the differences between read errors and splice junction accuracy, one metric, although related, does not have a direct correlation with the other. While sequence error correction is currently the main focus of many long read bioinformatic tools, it cannot be applied at the cost of biological inaccuracies as is the case for the gene and transcript read jumbling events occurring as a result of long read and short read error correction.

The underlying issues in all methodologies is the balance between retaining useful information (sensitivity) and discarding misleading information (specificity). The TAMA pipeline is designed for analyzing long read transcript data to produce high accuracy transcript models, high confidence sets of genes, and increasing the sensitivity for identifying novel genes.

From our analyses of the UHRR PacBio Sequel II Iso-Seq data, we have identified that there may be issues with the RNA preparation methods and/or there are still thousands of novel genes that have not been annotated in the human genome.

Methods

Universal Human Reference RNA and PacBio sequencing

RNA and cDNA library preparation and sequencing were undertaken by Pacific Biosciences. Pacific Biosciences made the data available for public use via a Github repository (https://github.com/PacificBiosciences/DevNet/wiki/Sequel-II-System-Data-Release:-Universal-Human-Reference-(UHR)-Iso-Seq). The RNA library was first created by pooling the Universal Human Reference RNA (Agilent) with SIRV Isoform Mix E0 (Lexogen). cDNA was prepared from the RNA using the Clontech SMARTer kit. The sequencing library was prepared using the Iso-Seq Template Preparation for Sequel Systems (PN 101-070-200) and Sequencing Sequel System II with "Early Access" binding kit (101-490-800) and chemistry (101-490-900). The sequencing library was sequenced on two Sequel II SMRT cells.

Chicken Brain RNA and PacBio sequencing

The non-cap selected chicken brain Iso-Seq data is from the European Nucleotide Archive submission PRJEB13246 which was previously analyzed and published(5).

The cap selected chicken brain Iso-Seq data was from an adult Advanced Intercross Line chicken whole brain sample. The RNA was extracted from the tissue sample using the Qiagen RNeasy Mini Kit. The RNA was converted to cDNA using the Lexogen TeloPrime kit. The resulting cDNA library was sent to Edinburgh Genomics for sequencing on the Sequel system using 2.0 chemistry.

Iso-Seq Processing

The UHRR Sequel II Iso-Seq data was processed into CCS reads using the *ccs* tool with the parameters *"--noPolish --minPasses = 1"*. CCS reads with cDNA primers and polyA tails were identified as full-length, non-concatemer (FLNC) reads using *lima (--isoseq --dump-clips*) and *isoseq3 refine (--require-polya*).

TAMA Low Pipeline

Full descriptions of the TAMA algorithms can be found in the wiki pages of the Github repositiory (https://github.com/GenomeRIK/tama/wiki).FLNC reads were mapped to GRCh38 using minimap2 *(--secondary = no -ax splice -uf -C5 -t 8*). The resulting bam files were then split into 12 smaller bam files using *tama_mapped_sam_splitter.py* which splits bam files by chromosome thus preventing splitting between reads from the same gene. Split bam files were annotated using *TAMA collapse (-d merge_dup -x no_cap -a 100 -z 100 -sj sj_priority -lde 5 -sjt 20 -log log_off)* then merged into a single bed file using *TAMA merge (-a 100 -z 100*).

TAMA High Pipeline

TAMA collapse was run on the split bam files using more stringent parameters that filter out any mapped read with more than 1 error within 20 bp of a splice junction (*-d merge_dup -x no_cap -a 100 -z 100 -sj sj_priority -lde 1 -sjt 20 -log log_off*). Merging was done in the same manner as the FLNC Low Pipeline. Transcript models supported only by a single read were filtered out using *tama_remove_single_read_models_levels.py (-l transcript -k remove_multi -s 2*).

Polish Pipeline

FLNC reads from the *isoseq3 refine* step were clustered using *isoseq3 cluster* and *isoseq3 polish* with default parameters. The output high-quality transcripts were mapped to the genome using minimap2 *(--secondary = no -ax*

splice -uf -C5 -t 8) and processed using *TAMA collapse (-d merge_dup -x no_cap -a 100 -z 100 -sj sj_priority -lde 5 -sjt 20 -log log_off*).

Cupcake Pipeline

The analysis followed the same steps as the Polish pipeline up to the mapping step. Mapped bam files were processed using Cupcake *collapse_isoforms_by_sam.py (–dun-merge-5-shorter*).

Lordec Pipeline

FLNC reads from the *isoseq3 refine* step were error corrected using Lordec *(-k 31 -s 3*) with short read RNA-seq data from the Universal Human Reference RNA (Agilent) (https://www.ncbi.nlm.nih.gov/sra/SRX1426160) (https://rnajournal.cshlp.org/content/22/4/597.full.pdf). The resulting error-corrected reads were processed in the same way as the FLNC Low starting from the mapping step.

Wobble Analysis and Novel Genes

To assess the wobble between each pipeline and the Ensembl annotation, we used *TAMA merge* with parameter settings *(-a 300 -z 300 -m 30 -d merge_dup*) which considers any transcripts which have up to 300 bp difference in their transcription start and end and up to 30 bp difference in their splice junctions starts and ends to have “nearly identical structures”. This is the definition for matching at transcript level.

Novel genes were identified by using the *TAMA merge* output from the wobble analysis. All genes that had no exonic overlap on the same strand as a gene annotated in the Ensembl human annotation v96 were classified as novel.

Coding Potential Analysis

For the Ensembl match evidence of coding potential, we labelled the Iso-Seq annotation genes as coding if they had any overlap on the same strand as an Ensembl-annotated protein coding gene.

CPAT was used with default parameters and the built-in Human Hex models. A cutoff score of 0.364 (suggested by the CPAT creators(24)) was used to segregate between coding and non-coding transcripts.

We used the TAMA ORF/NMD pipeline for the third source of coding evidence. ORFs were predicted for each transcript then translated into amino acid sequences. Blastp (*-evalue 1e-10 -ungapped -comp_based_stats F*) was used to match the amino acid sequences to the UniRef90 database, where the top hits were selected as the best ORF prediction. Transcripts with no hits were considered to be non-coding.

Code Availability

TAMA is available from https://github.com/GenomeRIK/tama.

Data Availability

The PacBio Universal Human Reference RNA Sequel II Iso-Seq dataset is available from https://github.com/PacificBiosciences/DevNet/wiki/Sequel-II-System-Data-Release:-Universal-Human-Reference-(UHR)-Iso-Seq. The short read Illumina RNA-seq data used for Lordec error correction are available in the National

Center for Biotechnology Information Sequence Read Archive under accession number SRP066009 (https://www.ncbi.nlm.nih.gov/sra/SRX1426160). The non-cap selected chicken brain Iso-Seq data is available from the European Nucleotide Archive under accession number PRJEB13246. The TeloPrime cap selected chicken brain Iso-Seq data is available from the European Nucleotide Archive under accession number PRJEB25416.

Declarations

Author's contributions

RIK developed TAMA and implemented the different Iso-Seq pipelines. RIK, DWB, and YC conceived the idea of this study. DWB provided guidance on the focus of the study. YC ran the ORF/NMD pipeline and identified issues with gene swapping. JWSB and RZ tested the TAMA Collapse LDE feature. JS, ALA, JWSB, and RZ reviewed and edited the manuscript.

Acknowledgements

We would like to thank Dr. Elizabeth Tseng and Pacific Biosciences for releasing the Universal Human Reference RNA Sequel II Iso-Seq dataset and providing guidance on the analyses.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

We acknowledge funding support from the UK's Biotechnology and Biological Sciences Research Council (Institute Strategic Programme grant BBS/E/D/10002070; and BB/N019202/1, BB/M011461/1, BB/M01844X/1). The funding bodies did not contribute to the design of the study, sample collection, analysis, interpretation of data, or in writing the manuscript.

References

1. Salzberg SL. Next-generation genome annotation: We still struggle to get it right. Genome Biol. 2019;20(1):19–21.
2. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. 2012;22:1760–74.
3. Adams MD, Kelley JM, Gocayne JD, Dubnick MAK, Polymeropoulos MH, Xiao H, et al. Complementary DNA sequencing: Expressed sequence tags and human genome project. Science. 1991;252(5013):1651–6.
4. Sanger F, Coulson AR, Barrell BG, Smith AJH, Roe BA. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. J Mol Biol. 1980;143(2):161–78.
5. Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genom. 2017;18(1):1–19.

6. 10.1038/ncomms11708
Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun [Internet]. 2016;7:11708. Available from: http://www.nature.com/doifinder/10.1038/ncomms11708.
7. 10.1038/ncomms11706
Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, et al. A survey of the sorghum transcriptome using single-molecule long reads. Nat Commun [Internet]. 2016;7:11706. Available from: http://www.nature.com/doifinder/10.1038/ncomms11706.
8. 10.1186/s12864-017-3757-8
Hoang NV, Furtado A, Mason PJ, Marquardt A, Kasirajan L, Thirugnanasambandam PP, et al. A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. BMC Genomics [Internet]. 2017;18(1):395. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3757-8.
9. 10.1371/journal.pone.0132628
Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, et al. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. PLoS One [Internet]. 2015;10(7):1–15. Available from: http://dx.doi.org/10.1371/journal.pone.0132628.
10. Koepfli K-P, Paten B, O'Brien SJ. The Genome 10K Project: A Way Forward. Annu Rev Anim Biosci. 2015;3(1):57–111.
11. Yao J, Qin Y, Wu DC, Nottingham RM, Lambowitz AM, Hunicke-Smith S. RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. Rna. 2016;22(4):597–613.
12. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018;46(D1):D754–61.
13. 10.1186/s12859-016-1316-y
Hu R, Sun G, Sun X. LSCplus:. A fast solution for improving long read accuracy by short read alignment. BMC Bioinformatics [Internet]. 2016;17(1):1–9. Available from: http://dx.doi.org/10.1186/s12859-016-1316-y.
14. Holmes I, Durbin R. Dynamic programming alignment accuracy. J Comput Biol. 1998;5(3):493–504.
15. 10.1038/ng.3791
Schor IE, Degner JF, Harnett D, Cannavò E, Casale FP, Shim H, et al. Promoter shape varies across populations and affects promoter evolution and expression noise. Nat Genet [Internet]. 2017;(February). Available from: http://www.nature.com/doifinder/10.1038/ng.3791.
16. Djebali S, Davis C, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature [Internet]. 2012 Sep 6 [cited 2014 Jul 9];489(7414):101–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3684276&tool=pmcentrez&rendertype=abstract.
17. 10.1038/s41467-018-06910-x
Sahlin K, Tomaszkiewicz M, Makova KD, Medvedev P. Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. Nat Commun [Internet]. (2018):1–12. Available from: http://dx.doi.org/10.1038/s41467-018-06910-x.
18. Epping MT, Hart AAM, Glas AM, Krijgsman O, Bernards R. PRAME expression and clinical outcome of breast cancer. Br J Cancer. 2008;99(3):398–403.

19. Field MG, Decatur CL, Kurtenbach S, Gezgin G, Van Der Velden PA, Jager MJ, et al. PRAME as an independent biomarker for metastasis in uveal melanoma. Clin Cancer Res. 2016;22(5):1234–42.
20. Roszik J, Wang W-L, Livingston JA, Roland CL, Ravi V, Yee C, et al. Overexpressed PRAME is a potential immunotherapy target in sarcoma subtypes. Clin Sarcoma Res. 2017;7(1):1–7.
21. Zhang W, Barger CJ, Eng KH, Klinkebiel D, Link PA, Omilian A, et al. PRAME expression and promoter hypomethylation in epithelial ovarian cancer. Oncotarget. 2016;7(29).
22. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res [Internet]. 2014 Oct 28 [cited 2014 Nov 25];43(October 2014):662–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25352552.
23. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
24. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li WCPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res [Internet]. 2013 Apr 1 [cited 2015 Feb 19];41(6):e74. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3616698&tool=pmcentrez&rendertype=abstract.
25. The UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res [Internet]. 2014;43(Database issue):D204-12. Available from: http://nar.oxfordjournals.org/content/43/D1/D204%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4384041&tool=pmcentrez&rendertype=abstract.
26. 10.1371/journal.pone.0157779
Cartolano M, Huettel B, Hartwig B, Reinhardt R, Schneeberger K. cDNA library enrichment of full length transcripts for SMRT long read sequencing. PLoS One [Internet]. 2016;11(6):1–10. Available from: http://dx.doi.org/10.1371/journal.pone.0157779.

Figures

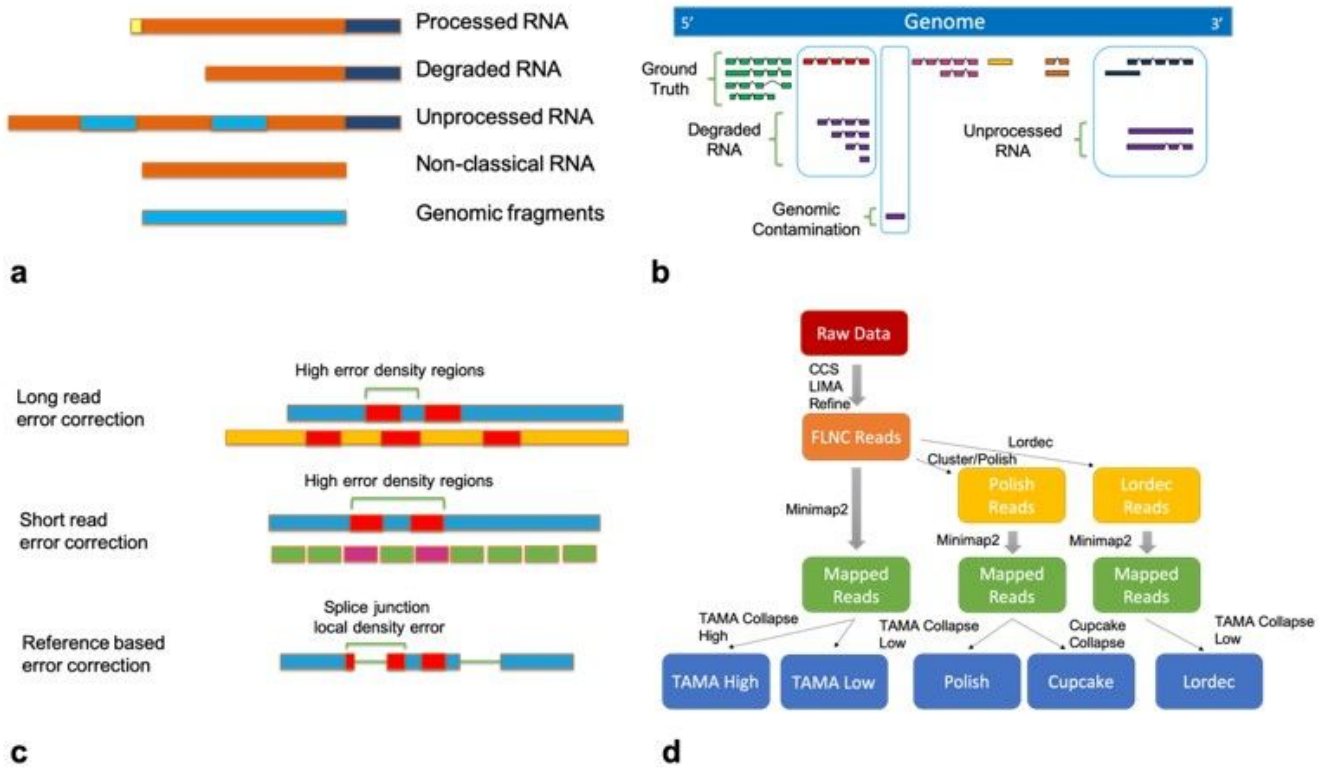

Figure 1

Long read RNA diversity and artefacts. (a) RNA samples are typically comprised of a mixture of degraded and immature RNA as well as DNA fragments which can be erroneously identified as novel genes and transcripts. (b) Representation of RNA sample sequences relative to the genome provides some clues as to the nature of the source for predicted gene and transcript models. (c) Illustration of problems arising from different error correction methods for long reads. Error dense regions within reads contribute to the erroneous clustering of reads from different genes or transcripts. A reference based approach allows for the identification of reads which have regions of high error density which may impact the model prediction. (d) Standard Iso-Seq pipeline. (e) Work flow for the different pipelines analyzed.

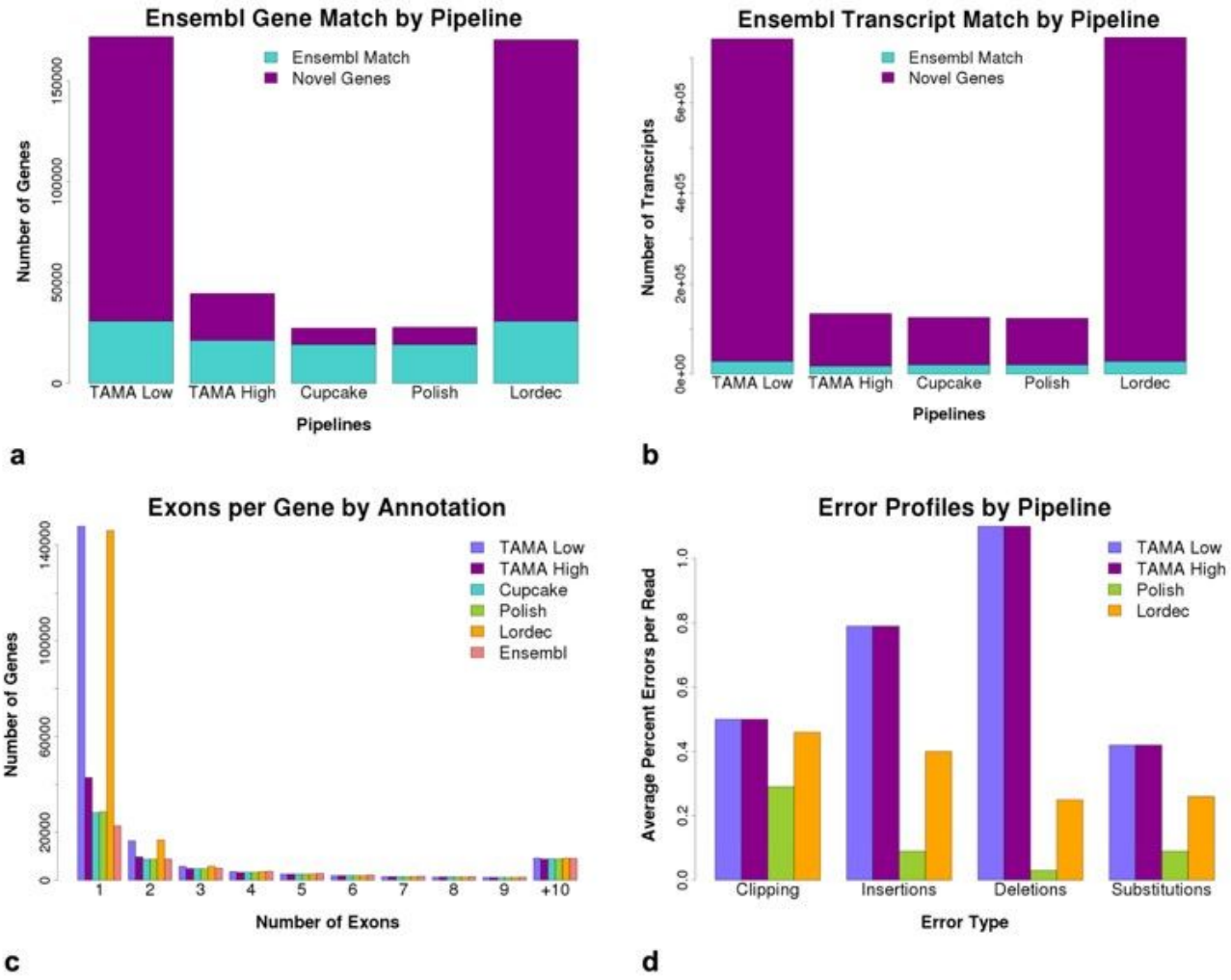

Figure 2

Comparing the results of the different Iso-Seq pipelines. (a) Number of novel and Ensembl matching genes by pipeline. (b) Number of novel and Ensembl matching transcripts by pipeline. (c) Maximum number of exons per gene by pipeline. (d). Type and amount of error per pipeline.

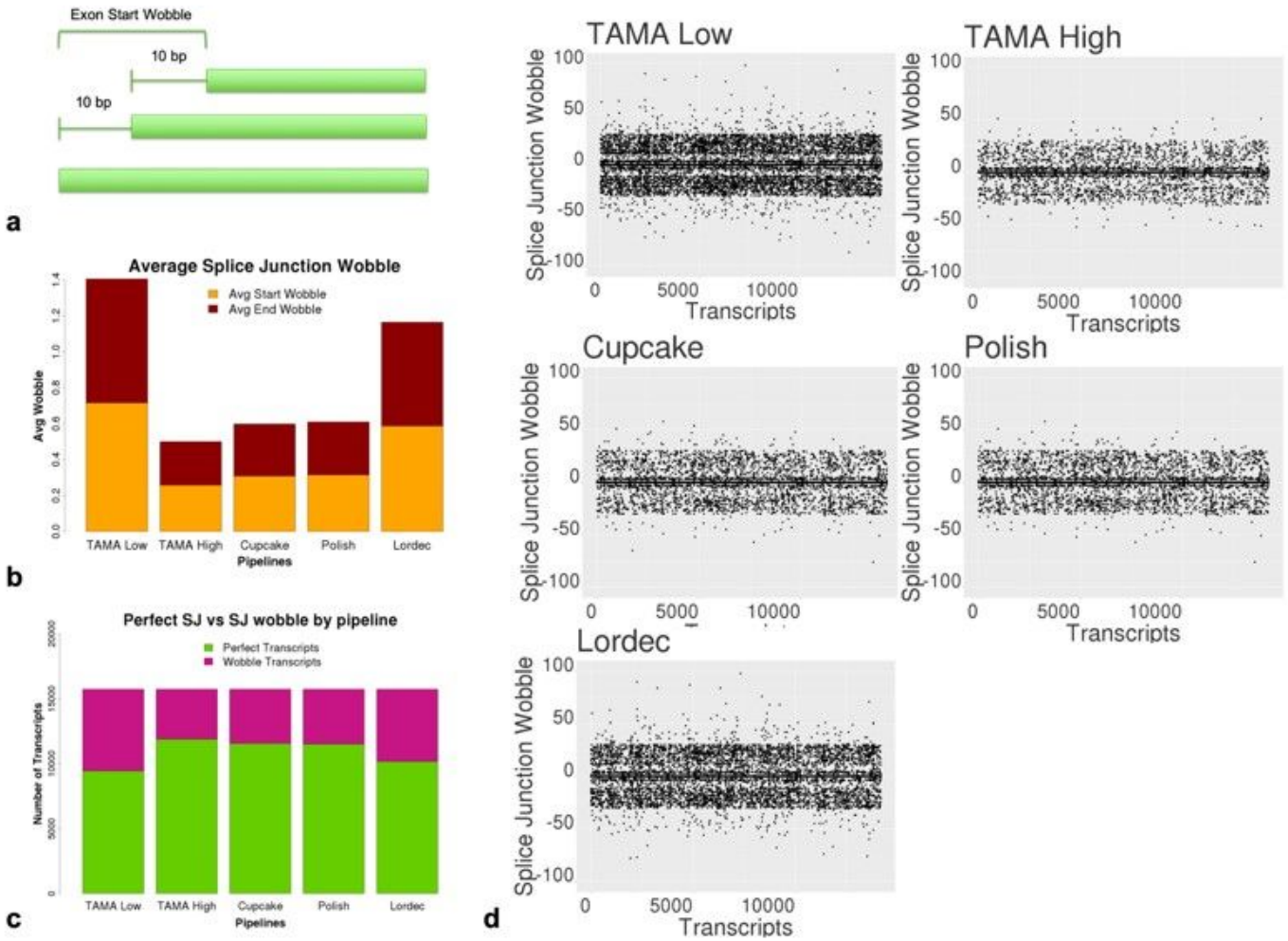

Figure 3

Assessing wobble across pipelines. (a) Illustration of wobble which is defined as the difference in exon start and/or end between transcript models from the pipelines and Ensembl v96.. This occurs due to mapping uncertainty when there are errors in the reads near splice junctions. (b) Average amount of wobble across splice junctions by pipeline. (c) Number of perfect transcript models and transcript models with wobble as compared to the Ensembl annotation by pipeline. (d) Scatter plot of splice junction wobble per transcript by pipeline. Wobble is shown as base pair distance from true splice junction exon start and end. Positive wobble represents exon start wobble and negative wobble represents exon end wobble. Each x-axis unit represents a single transcript model. A 30bp wobble threshold was used for the TAMA Merge run thus the apparent drop off in wobble outside this range.

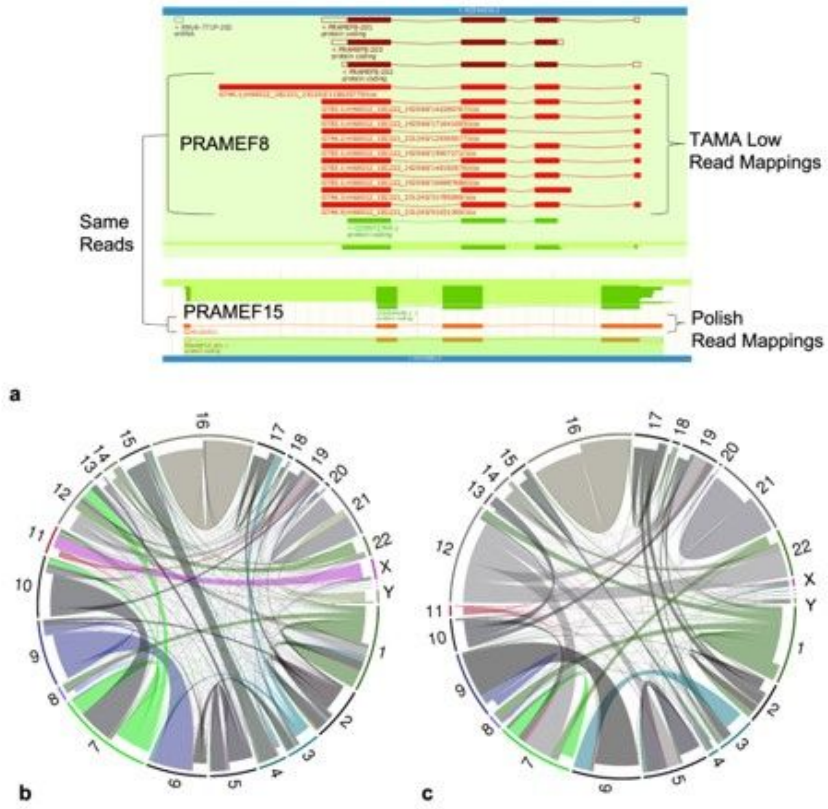

Figure 4

Gene and transcript read swapping from error correction. (a) PRAMEF8 gene with a coverage of 9 TAMA Low reads which are clustered and merged in the Polish pipeline resulting in jumbled cluster read mapping to PRAMEF15 gene. (b) Circos plot showing reads swapping genes after correction with Cluster/Polish. Indented shows true read location and non-indented shows read allocation after error correction. Each line represents a single read moving from one gene to another with 34,637 reads from 4,799 genes moving to 2,793 genes after Cluster/Polish.(c) Circos plot for reads swapping genes after correction with Lordec. Each line represents a single read moving from one gene to another with 19,064 reads from 2,292 genes moving to 2,319 genes after Lordec error correction

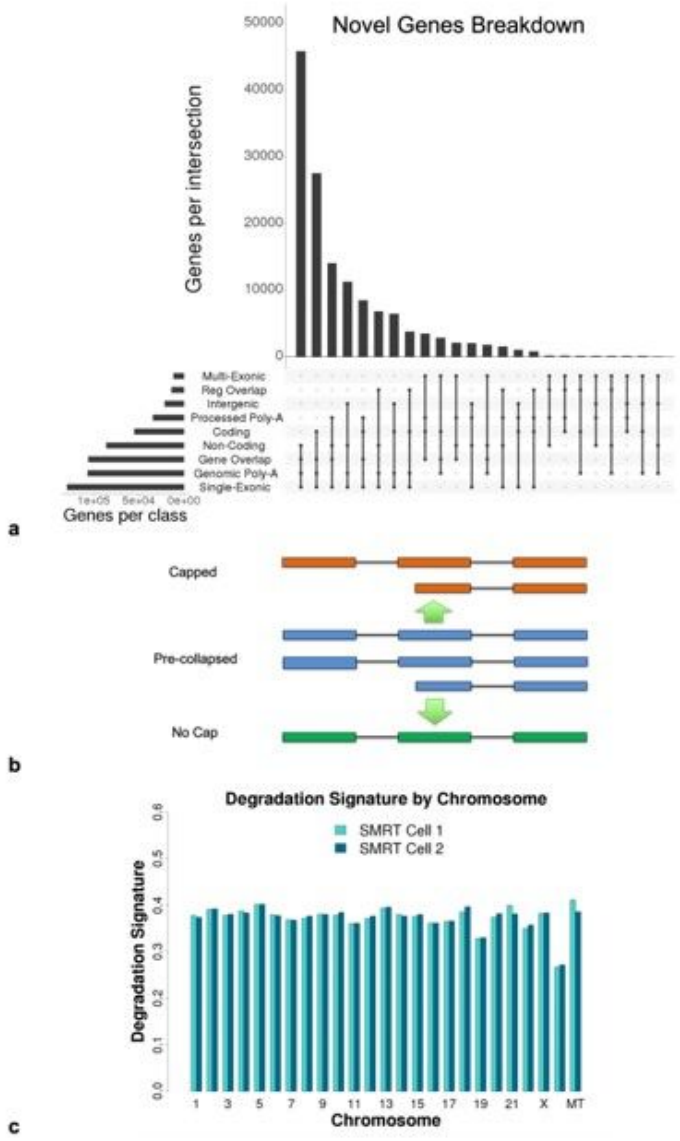

Figure 5

Novel genes breakdown and degradation signature analysis. (a).Novel gene breakdown by features. (b) Collapsing algorithms for 5’ cap selected RNA and non-cap selected RNA. (c). Degradation signature by chromosome per SMRT Cell run.