Published in final edited form as:

Nat Methods. 2023 August ; 20(8): 1187–1195. doi:10.1038/s41592-023-01908-w.

Context-Aware Transcript Quantification from Long Read RNA-Seq data with Bambu

Ying Chen[1,*], **Andre Sim**[1,*], **Yuk Kei Wan**[1,2], **Keith Yeo**[1], **Joseph Jing Xian Lee**[1], **Min Hao Ling**[1], **Michael I. Love**[3,4], **Jonathan Göke**[1,5,+]

[1] Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), 60 Biopolis Street, Genome, Singapore 138672, Republic of Singapore

[2] Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[3] Department of Biostatistics, University of North Carolina-Chapel Hill, Chapel Hill, USA

[4] Department of Genetics, University of North Carolina-Chapel Hill, Chapel Hill, USA

[5] Department of Statistics and Data Science, National University of Singapore, Singapore, Singapore

Abstract

Most approaches to transcript quantification rely on fixed reference annotations. However, the transcriptome is dynamic, and depending on the context, such static annotations contain inactive isoforms for some genes while they are incomplete for others. Here we present Bambu, a method that performs machine-learning based transcript discovery to enable quantification specific to the context of interest using long-read RNA-Seq. To identify novel transcripts, Bambu estimates the novel discovery rate (NDR), which replaces arbitrary per-sample thresholds with a single, interpretable, precision-calibrated parameter. Bambu retains the full-length and unique read counts, enabling accurate quantification in presence of inactive isoforms. Compared to existing methods for transcript discovery, Bambu achieves greater precision without sacrificing sensitivity. We show that context-aware annotations improve quantification for both novel and known transcripts. We apply Bambu to quantify isoforms from repetitive HERVH-LTR7 retrotransposons

[+] corresponding author: gokej@gis.a-star.edu.sg.

[*] contributed equally

Author contribution statement

Y.C. and A.S. designed and implemented the computational method. J.G. conceived the project. Y.C., A.S., and J.G. designed the study and experiments and analysed data. Y.K.W., K.Y., J.J.X.L., and M.H.L. contributed to implementation of the computational method. M.L. contributed to the design of the computation method. Y.C., A.S. and J.G. organised and wrote the paper with contributions from all authors.

Competing Interests Statement

Jonathan Göke received travel and accommodation expenses to speak at the Oxford Nanopore Community Meeting 2018. All other authors declare no competing interest.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Code Availability

Bambu is a R package for transcript discovery and quantification across multiple samples that is maintained on Bioconductor: https://www.bioconductor.org/packages/bambu/. The source code and a detailed documentation is available on GitHub: https://github.com/GoekeLab/bambu/. We used the BambuManuscriptRevision branch version of Bambu for analysis done in this manuscript: https://github.com/GoekeLab/bambu/tree/BambuManuscriptRevision. All analysis code is available on code ocean[56].

in human embryonic stem cells, demonstrating the ability for context-specific transcript expression analysis.

Keywords

Nanopore RNA sequencing; Long read RNA sequencing; Transcriptomics; Transcript discovery; Transcript quantification; Alternative splicing

Introduction

Transcription of DNA is a complex process where multiple alternative RNA transcripts can be expressed from the same gene loci[1–4]. Transcripts which are derived from alternative gene isoforms can be functionally distinct, making it essential to quantify not just the level of transcription for each gene, but for each individual gene isoform[5,6].

The expression levels of these transcripts can be inferred through high throughput sequencing of RNA or cDNA (RNA-Seq)[7–10]. Sequencing reads are then assigned to known gene isoforms that are catalogued in reference genome annotations. Reference annotations aim to be a comprehensive atlas of an organism’s isoforms that capture all possible tissues and cellular stages. However, due to the dynamic nature of the transcriptome only a fraction of the annotated transcripts are expressed in any given sample, while additional, sample-specific isoforms might be missing from the reference[11,12]. This particularly impacts transcripts originating from repetitive sequences such as retrotransposons that are challenging to annotate, or cell types such as early embryonic cells that can have a large number of cell-type specific transcripts[13–15].

The ability of long read RNA-Seq to generate reads corresponding to full-length transcripts provides an opportunity to discover novel transcripts and thereby enable the quantification of isoform expression using context-specific annotations[16]. Tools such as FLAIR[17], TALON[18], StringTie2[19] or IsoQuant[20] have been developed for transcript discovery from long read RNA-Seq and have been shown to identify novel transcripts even in well annotated genomes. However, RNA degradation, sequencing, and alignment artefacts can introduce false positive transcript candidates and impact quantification[21].

To deal with possible false positive novel transcripts, existing methods rely on user defined thresholds such as the minimum read count to filter novel transcript candidates[17–19]. However as these parameters are dependent on sequencing depth, the same threshold can generate vastly different results across multiple samples[18,22,23] This can partially be addressed with additional thresholds such as a minimum relative isoform expression or transcripts-per-million. While these thresholds correct for sequencing depth, they are still influenced by aspects such as expression level or number of isoforms per gene and may erroneously filter out valid novel transcript candidates. Therefore, as none of these thresholds provide an intuitive way of controlling false positive transcript candidates, bounding the error rate of the resulting novel transcript set remains a challenge.

To overcome these limitations we have developed Bambu, a method for multi-sample transcript discovery and quantification. Bambu estimates the likelihood that a novel transcript is valid, allowing the filtering of transcript candidates with a single, interpretable parameter, the novel discovery rate (NDR), that is calibrated to guarantee a reproducible maximum false discovery rate across different samples and analyses. Bambu then employs a statistical model to assign reads to transcripts that distinguishes full-length and non full-length (partial) reads, as well as unique and non-unique reads, thereby providing additional evidence from long read RNA-Seq to inform downstream analysis. We demonstrate that the NDR implemented in Bambu enables transcript discovery with greater accuracy and with a wider dynamic range compared to existing tools. Additionally we show that Bambu's context-specific annotations improve quantification even for annotated transcripts, and that the ability to track unique and full-length reads reduces the impact of inactive transcripts. We apply Bambu to long read RNA-Seq from human embryonic stem cells (hESCs) and illustrate the ability to quantify individual isoforms from highly repetitive HERVH transposons with full-length read support. Together, Bambu addresses the limitation of static reference annotations while providing a quantitative measure of confidence for novel transcripts, enabling the comprehensive, context-aware quantification of individual isoforms from long read RNA-Seq.

Results

Context-aware quantification of long read RNA-Seq data

Bambu consists of 4 steps: Firstly, a probabilistic model is employed to correct junction alignments[24,25] using reference annotations, genome sequence, and features obtained from the data (see Methods). Corrected reads that use the same splice junctions are summarised into *read classes* (Figure 1a). In the second step, read classes from all samples are combined and a cross-sample NDR is calculated. Read classes below the specified NDR threshold are treated as novel transcripts, resulting in context-specific reference annotations (Figure 1b). Thirdly, each read class is assigned to compatible transcripts allowing for inexact matches to account for possible alignment errors (Figure 1c). In the fourth step, transcript expression estimates are obtained with an expectation-maximisation (EM)[7,9,10,26] algorithm for each sample using the same set of context-specific transcript annotations. Bambu estimates expression levels using reads that are uniquely assigned to a single transcript (*unique* reads) as well as reads which are assigned to multiple transcripts. Bambu then provides final expression estimates that include the number of *full-length* reads and *unique* reads that support each transcript. While Bambu models expression jointly using all reads, the individual contributions from each group are tracked to provide an intuitive measure of evidence for each transcript and gene to support their expression in the samples of interest (Figure 1d). Bambu is available through Bioconductor[27–30], it runs with a single command and a single, interpretable transcript discovery parameter, and can efficiently be applied to a large number of samples (Supplementary Text Section 9, Supplementary Text Figure 4, Supplementary Text Table 4–5), thereby greatly simplifying the quantification with context-specific annotations from long reads.

Bambu infers a probability score to rank novel transcripts

In order to obtain a single parameter for transcript discovery, Bambu extracts nine different features that summarise read count, read alignment, and read sequence characteristics and trains a supervised machine learning model[31] that infers the probability for each read class to be a valid transcript (transcript probability score) (Figure 2a, see Methods and Supplementary Text Section 1). The multi-sample transcript score is then defined as the probability that a transcript is valid in at least one sample (see Methods and Supplementary Text Section 3), enabling the ranking of transcript candidates across multiple samples without the need to apply any threshold on individual samples.

Bambu uses existing annotations as training data, and learns a model for each sample assuming that the majority of valid transcripts are annotated. However, for genomes which are poorly annotated or samples with a very high number of expected novel transcripts, a pretrained model can be used. Bambu includes a model that was trained on Nanopore RNA-Seq data from a human cell line (see Methods). However, for cases where the pretrained model is not suitable due to species and technology differences, Bambu provides the option to train a new model on related data with comprehensive annotations that is more appropriate for the sample of interest (see Supplementary Text Section 5 for additional information).

To evaluate the performance of the transcript score to identify valid transcripts, we first applied Bambu on sequin spike-in RNAs[32] where the ground truth is known, using Nanopore long read RNA-Seq samples from the Singapore Nanopore Expression (SG-NEx) project[33]. On these data, the transcript score shows a high level of precision[34], outperforming the baseline parameters of read counts and relative isoform gene expression (gene proportion) as parameters for transcript discovery (Figure 2b). While the sample-specific model shows the highest performance, the pre-trained model still outperforms baseline statistics (Figure 2b). Next we tested the performance for transcript discovery on the human chromosome 1[35] for which annotations were removed before running Bambu. The results similarly show that both the sample-specific and the pre-trained model have a higher level of precision compared to the baseline parameters, and it is consistent when applied on single samples or multiple samples (Figure 2c, Supplementary Figure 1a). We repeated this analysis using Arabidopsis[36,37] and PacBio RNA-Seq data[38], which further confirmed that the sample-specific model and the pre-trained model show better performance than read count or gene proportion (Supplementary Figure 1b–c). The TPS was even able to rank candidates that were lowly expressed or which consisted only of a single isoform, something not possible when using baseline read count and gene proportion thresholds (Supplementary Figure 1d–e).

To test the robustness of the supervised approach used in Bambu when annotations are incomplete, we trained a model providing 25%, 50%, 75% and 100% of human annotations respectively. Even with 50% of annotations we observe that the model trained in Bambu shows a high level of accuracy (Supplementary Figure 1f, Supplementary Text Table 1). When less than 50% of annotations are known, the model still generates accurate predictions, however a pretrained model using more complete annotations is able to provide improved performance (Supplementary Figure 1f). To catch such a scenario in practice, Bambu automatically infers the completeness of reference annotations and recommends using a pre-trained model when the expected fraction of missing transcripts is higher than 50% (Supplementary Figure 1g–h, see Methods and Supplementary Text Section 5). Even when trained on very poorly annotated genomes (25% of annotations), the sample-specific model shows a higher overall accuracy than read count (Supplementary Figure 1f, Supplementary Text Table 1). Together, these results indicate that the transcript score predictions in Bambu provide a robust and accurate way to rank and identify valid transcripts on both well annotated and poorly annotated genomes (Figure 2c, Supplementary Figure 1f, Supplementary Text Table 1).

The NDR, a single, interpretable, and comparable parameter

In order to make the transcript discovery parameter interpretable and comparable across samples, we define the novel discovery Rate (NDR) as the fraction of unannotated transcripts among all transcripts with an equal or higher transcript score. The NDR can be interpreted as an upper limit on the false discovery rate (or 1-precision) under the assumption that reference annotations are complete: a NDR of 0.1 indicates that at least 90% of transcripts with a similar score or higher are annotated, thereby providing an intuitive estimate of precision. As the expected number of novel transcripts differs depending on the completeness of the annotations, the same NDR can correspond to different levels of precision across different species or when alternative annotations are used. However, in practice most analyses are done within the same species and annotations where the number of expected novel transcripts is similar, in which case the same NDR threshold guarantees a similar FDR and precision across independent analysis.

To evaluate this, we identified novel transcripts with different NDR thresholds on the core SG-Nex data (Supplementary Table 1). Here we provide the human genome annotations without chromosome 1 during transcript discovery, using annotations from chromosome 1 as ground truth to estimate the precision of transcript discovery for each sample. A comparison of the observed precision for different NDR thresholds confirms that it is indeed well calibrated (Figure 2d). We find that the same NDR threshold provides a very similar level of precision across all samples, whereas an equivalent read count or gene proportion threshold results in a wide range of precision (Figure 2d–f). Furthermore, unlike thresholds such as read count, TPM, or relative read count that are used by other methods, the NDR provides a continuous metric that is linearly related to the expected precision (Figure 2d–f). This property enables transcript discovery across the complete dynamic range of precision, facilitating either conservative but accurate extension of annotations in the case of well annotated genomes, or more sensitive transcript discovery for genome annotation of species or samples with many unknown transcripts. To optimise results for each analysis where samples have varying levels of annotation completeness, Bambu infers an analysis-specific default NDR threshold using the estimate of the fraction of missing transcripts (see Methods). Using the pretrained model, Bambu is able to estimate the fraction of missing annotations accurately in ONT, PacBio and other species (Mouse and Arabidopsis) data (Supplementary Figure 1g, Supplementary Text Table 2). This dynamic default threshold is calibrated to ensure high levels of precision, however it can be changed for more sensitive transcript discovery. As the NDR is calculated on the multi-sample transcript probability score, it replaces sample-specific thresholds with a single, interpretable, and comparable parameter for transcript discovery.

Bambu achieves a higher dynamic range and accuracy

Next we benchmarked Bambu against FLAIR, TALON, StringTie2 and IsoQuant using the SG-NEx long read RNA-Seq samples. First we compared the impact of transcript discovery parameters on precision, confirming that read-count based parameters generate vastly different levels of precision for identical thresholds (Figure 2g). In contrast, the same NDR threshold provides a comparable precision across independent samples (Figure 2g). Next, we evaluated the precision and sensitivity to identify valid multi-exon transcripts when 50% of the transcript annotations were removed from chromosome 1 in the reference (see Methods). A comparison of the different methods demonstrates that Bambu achieves higher precision at a comparable sensitivity when samples are analysed individually (Figure 2h). StringTie2 provides a threshold parameter for read counts and gene proportion, therefore to compare all tools we applied a manual gene proportion threshold to the output of FLAIR and TALON, which resulted in improved performance for these tools in the tested sample, but did not outcompete Bambu (Supplementary Figure 1i). The same results are obtained when applied to spike-in transcripts and when no annotations were provided (Supplementary Figure 1j–m). Furthermore, Bambu enables transcript discovery with a wider dynamic range of precision compared to all other tools (Figure 2h).

A unique feature of Bambu is its ability to analyse results across multiple samples with a single, calibrated threshold. When all SG-NEx samples are jointly analysed, Bambu maintains the ability to perform transcript discovery across the full dynamic range of precision, while other methods that use non-calibrated thresholds applied to each sample showed a smaller range of precision or sensitivity (Figure 2i). This property is particularly relevant for well annotated genomes (e.g., human), where high sample numbers with high sequencing depth and using read-count-based thresholds would otherwise result in high numbers of novel transcripts that might impact downstream quantification. Together these results show that Bambu is more accurate, provides a wider range of precision and is the only method where the precision can directly be controlled with a single transcript discovery parameter.

Context-aware annotations improve transcript quantification

After transcript discovery, Bambu estimates transcript expression using the full length and partial length read classes for all samples. We first compared the quantification in Bambu without transcript discovery (NDR=0) to existing quantification-only methods Salmon[10], NanoCount[39], featureCounts[40], and LIQA[41] using the sequin spike-in RNAs with complete reference annotations. We find that quantification with Bambu on spike-in RNAs had comparable or better performance to the existing quantification-only tools (Figure 3a–b).

Next, to evaluate the impact of transcript discovery on quantification, we specified different NDR thresholds and compared the accuracy of abundance estimates for spike-in RNAs with partially missing annotations. Quantification after transcript discovery in Bambu (NDR>0) allows the abundance estimation for missing gene isoforms, reducing the overall estimation error (Figure 3c–d, Supplementary Figure 2a–c, Supplementary Figure 3a). More sensitive transcript discovery will increase the number of false positive transcripts, highlighting the importance of choosing a threshold that is appropriate for the analysis (Figure 3e).

Our results further suggest that transcript discovery also stabilises abundance estimates for isoforms which are already present in the reference annotation, leading to more consistent results on the same data and higher reproducibility with increasing NDR (Figure 3f–g, Supplementary Figure 2d–e, Supplementary Figure 3b). The same observation is made when the extended annotations from Bambu are used with other quantification-only tools, where we observe that transcript discovery reduces the quantification error of annotated transcripts (Supplementary Figure 4–5).

A comparison of quantification after transcript discovery in Bambu with other transcript discovery methods[17–20,42] shows higher variation across the tools (Supplementary Figure 2). These results reflect the differences quantification methods, but also in the extended annotations that differ for each tool and which are defined by tool-specific default parameters, precision, and sensitivity, further highlighting the impact of transcript discovery on quantification and emphasising the value of the single parameter in Bambu that enables the control of false positive transcripts and their impact on quantification.

Bambu estimates full-length and unique read support

Even in long read RNA-Seq data, reads that match multiple transcripts are still frequently present (*non-unique* reads: 13.8 – 49.5%, Supplementary Figure 6a). Similar to existing methods, Bambu uses an EM algorithm to probabilistically assign non-unique reads to transcripts. While this approach has been previously demonstrated as effective[39], there is no guarantee that transcripts which are only supported by non-unique reads are indeed expressed in the samples of interest. To address this, we calculate for each read if it matches a complete transcript (full-length), and if it can be uniquely assigned to a single transcript (unique) (see Methods). Bambu then provides estimates of the full-length and unique read support in addition to the total abundance estimation in counts per million (CPM) for each gene isoform. When we compared transcript expression estimates across biological replicates, we found that transcripts with a higher number of unique or full-length reads show higher correlation, suggesting that they provide additional information that is not captured by the total CPM (Figure 4a, Supplementary Figure 6b–e).

To evaluate if CPM, unique, and full-length read counts can be used as evidence that a transcript is expressed in the samples of interest, we included artificial isoforms containing a unique splice junction (exon skipping) in the reference annotation prior to quantification (Figure 4b). These artificial isoforms are not expressed, however, due to probabilistic read-assignment and approximate read matching in Bambu, they can still be estimated to be active (Figure 4c, Supplementary Figure 6f). Increasing the CPM threshold reduces the number of such transcripts, indicating that read-misassignment mostly affects transcripts with lower overall read count, and suggesting that a basic expression filter already provides some evidence that a transcript is expressed (Figure 4c). Using the presence of unique reads or full-length reads further improved the ability to identify expressed transcripts, achieving a higher precision at the same level of sensitivity when compared to the basic CPM expression filter (Figure 4c, Supplementary Figure 6g). While Bambu considers all reads to obtain the CPM estimate, the ability to track full-length and unique reads enables Bambu to quantify transcript abundance in the presence of unexpressed reference annotations, as well as providing an additional layer of evidence to inform downstream analysis or follow up experiments.

Quantification of retrotransposon-derived transcripts

Among the most difficult genes to quantify are those that are derived from retrotransposons, as they are highly repetitive and often not accurately annotated. One of the most striking examples of retrotransposon expression is the HERVH-LTR7 (Human endogenous retrovirus subfamily H-Long Terminal Repeat) family which has been reported to be a highly specific marker of pluripotency in hESCs[43–46]. The human genome contains several thousand annotated HERVH-LTR7 elements. However, the number of expressed elements is expected to be much smaller[47]. To test[48,49] if Bambu can enable the detailed reconstruction of HERVH-derived transcripts purely from RNA-Seq data, we analysed the hESC samples from the SG-NEx data. We observed a significant enrichment of repetitive sequence[50] in novel genes ($p < 0.001$), with HERVH-LTR7 being the dominating repeat family (Figure 5a). In total, 242 genes (encompassing 464 transcripts) are transcribed from HERVH, 64 of them contributing to 90% of the HERVH RNA in hESCs, suggesting that only a small minority of HERVH elements are actually transcribed (Figure 5b, Supplementary Figure 7, and Supplementary Table 2). These HERVH-derived genes are supported by full-length reads, and they generate distinct transcripts with alternative splicing patterns that are unique to each locus, suggesting that they might be functionally distinct (Figure 5c–f). Quantifying transcript expression in hESCs without the extended annotations from Bambu results in an overestimation of existing transcripts such as *ESRG*, where the majority of reads originate from previously undescribed isoforms (Figure 5d, Supplementary Figure 7c). Together, these results illustrate how context-aware quantification with Bambu reduces quantification error while enabling the estimation of individual retrotransposon-derived isoform expression from long read RNA-Seq without any additional experimental or computational requirements.

Discussion

While static reference annotations simplify the quantification of transcript expression, they do not necessarily reflect the samples or analysis of interest, with unknown transcripts continuing to be discovered even in well annotated species[22,23,51]. To address this, we developed Bambu, a method that enables quantification of transcript expression with annotations that are inferred specifically for the context of interest using long read RNA-Seq.

The accuracy of transcript discovery is influenced by the parameter thresholds that are applied to identify novel transcripts[52,53]. Bambu employs a machine learning approach to control the false positive rate using a single transcript discovery parameter (the NDR). In contrast to parameters such as read count, relative expression and TPM, the NDR ranks transcript candidates by their probability of representing a valid full-length transcript. Therefore, unlike arbitrary thresholds for parameters used in other tools[17–19], a more stringent NDR threshold guarantees higher precision while providing an upper bound on the FDR. This is especially relevant for well annotated genomes or analyses which involve high numbers of samples where precision is more important than sensitivity to obtain accurate annotations and quantification results. However, even analyses where a high number of novel transcripts are expected will benefit from this property as the NDR identifies the most precise set of annotations even when more sensitive thresholds are selected.

In most cases, the expected number of novel transcripts is unknown before transcript discovery. To avoid arbitrary and often inappropriate default thresholds, Bambu estimates the fraction of annotations that are missing in the sample, which is then used to recommend a suitable NDR for the analysis. Therefore, unlike other tools that rely on fixed default thresholds[17–19], Bambu infers a threshold to achieve high precision in transcript discovery and accurate quantification.

While Bambu uses annotations for training the transcript prediction model, splice site correction and NDR calibration, it also works well on poorly annotated genomes and those without any annotation. In these scenarios a pretrained model predicts the transcript probability score without the need for reference annotations. Bambu provides a model trained on nanopore RNA-Seq data from human cell lines, however, for samples with vastly different genomes or alternative sequencing technologies, Bambu includes the option to pre-train a model on a similar dataset. Without reference annotations the NDR is not estimated, and transcripts are instead ranked by the uncalibrated transcript probability score from the pretrained model. While the transcript score is not calibrated to be comparable across analysis, its ability to rank transcript by precision is identical to the NDR, making it possible to discover novel transcripts with low false discovery rates even in the absence of annotations.

Very sensitive transcript discovery provides more novel transcript candidates, however, we observe that the increasing complexity of annotations can impact quantification due to reads which are assigned to multiple transcripts. Long-read RNA-Seq is able to generate reads which match full-length transcripts, thereby providing evidence that transcripts are present in the samples of interest[22,52–54]. In Bambu we use all reads for quantification, while providing the full-length read count estimate for each transcript. However, the classification of reads as full-length in Bambu is influenced by both sequencing and alignment errors as well as RNA degradation, with non-unique full-length reads still being ambiguous. Approaches have been developed to enrich full-length reads experimentally[55], or selectively analyse full-length reads during quantification[42]. These approaches are more effective in identifying which transcripts are truly expressed, however, as they significantly reduce the number of reads used for quantification, the overall abundance estimates are likely to be less accurate. Here we find that non-unique full-length reads only represent a minor fraction of reads, suggesting that for most genes, Bambu’s approach provides accurate quantification while still preserving the key information about full-length reads for each transcript.

However, there are a few limitations in Bambu. One challenge with long read RNA-Seq data is the presence of incomplete reads, which limits the ability to identify precise start and end positions for transcripts. Reference annotations can not be relied upon to overcome this challenge as they often incorrectly describe longer or shorter first and last exons. Therefore to minimise the impact of imprecise start and end coordinates in reads and annotations, Bambu heavily relies on splice junctions for transcript discovery and quantification. While this increases the robustness for most transcripts that have unique splicing patterns, transcripts which only differ due to alternative start or end coordinates will currently not be identified with Bambu. Furthermore, even though single exon transcripts can be identified and quantified, their predicted start and end coordinates might be incorrect, which particularly impact genomes that are rich in single exon transcripts such as yeast or microbes. Finally, this also impacts transcripts that are subsets of other transcripts and which are indistinguishable from degradation products. Bambu by default excludes subset transcripts to mitigate high numbers of false positives originating from truncated reads. While this procedure ensures high precision in transcript discovery, Bambu may potentially overlook valid novel subset transcript candidates, and thus leading to biased quantification results towards non-subset transcripts. These limitations can potentially be addressed with statistical modelling of transcript degradation and truncation, resulting in more precise inference of start and end coordinates, identification of alternative start or end sites from long read RNA-Seq data, and improved quantification for single exon transcripts, subset transcripts, and non-subset transcripts. Doing so will form a key component in the future direction of improving Bambu.

The ability to perform context-specific quantification is particularly relevant for applications where novel transcripts are expected. When applying Bambu to hESCs, we discovered many retrotransposon-derived genes that are missing from current annotations, with full-length reads providing an additional layer of evidence. Furthermore, Bambu enables applications in other areas, including identifying and quantifying individual fusion transcripts when reads are aligned to the breakpoint-corrected genome, or combined with tools for detection of RNA modifications from direct RNA-Seq data, Bambu can provide insights into epitranscriptomic changes at novel transcripts. Yet, Bambu is not just limited to long read RNA-Seq. With a small set of representative long read samples, Bambu can be used to generate context-specific annotations to further improve quantification from large-scale short read data sets. In conclusion, Bambu simplifies transcript discovery and quantification across multiple samples to a single command with a single parameter, while improving accuracy and interpretability of expression estimates, suggesting that quantification with context-specific annotations could become a routine approach to analysing transcript expression with long reads.

Methods

Transcript Discovery and Quantification with Bambu

Bambu performs both transcript discovery and quantification with the following steps:

(1) Construction of Read classes

1.1 Alignment error correction

Firstly, to utilise long reads which can have high noise during basecalling resulting in errors in splice-junction alignments, we trained a machine learning model with scalable tree boosting to effectively predict the probability of a true splicing junction. This was performed for every 5' and 3' splice site considering information from splicing junctions within 15 base pair (bp) distance. The model is trained using R XGBoost[31] function with default parameters and uses 5' and 3' distances between read alignments and annotation, strand information, the presence of splice motifs, and read support relative to the reference splicing junctions as features. Splicing junctions that are predicted to be true will be noted as high confidence junctions and maintained, while junctions which are predicted to originate from mis-alignments are noted as low confidence junctions and will be corrected to the closest reference splice site within 10 bp, or kept with the original alignment if no reference splice site is within this range.

1.2 Construction of Read Classes

Here we assume that reads with identical splice patterns originate from the same transcripts. Under this assumption, we summarise all reads spanning the same (error-corrected) junctions into *read classes (RC)*. The start coordinate of the read class is defined as the position that includes 80% of all read start positions for this read class, the end coordinate is similarly defined to include 80% of the read end positions for each read class. Only read classes with high confidence junctions (and optionally read classes which are unspliced) are retained for transcript discovery, whereas all read classes are used for quantification. Novel read classes are matched to genes on the basis of exon overlap with reference annotations. Read classes which do not overlap with reference annotations are classified as belonging to a novel gene, assigned a new gene id, and grouped together with other read classes which they overlap.

(2) Feature extraction and Model Training

To predict the probability that a read class represents a valid transcript, Bambu uses a supervised machine learning approach using XGBoost. Firstly, Bambu extracts 9 features from the read classes: read counts-per-million (referred to as read count), the proportion of contribution to the total read count of its corresponding gene (referred to as gene proportion), the proportion of reads mapping to the strand with higher read count, the standard deviation of supporting reads' 5' and 3' ends positions, and the frequency of A/Ts in the first and last 20 bp of the read class as features. (See Supplemental Text Section 1 for additional details). Bambu then represents each read class i from sample j with a feature vector $x_{i,j}$ and an associated binary class label $y_{i,j}$ that indicates if a read class is a valid transcript:

$$RC_{i,j} = (x_{i,j}, y_{i,j})$$

with $x_{i,j} = \{read\ count,\ relative\ read\ count,\ \ldots.\}$

and $y_{i,j} = \{1\ \mathit{if}\ RC_{i,j} \mathit{is\ a\ valid\ transcript};\ 0\ \mathit{otherwise}\}$.

The probability that a read class represents a valid transcript (the Transcript Probability Score) for read class $RC_{i,j}$ is then defined as

$$TPS_{RC_{i,j}} = P(y_{i,j} = 1) = f(x_{i,j})$$

Here, $f(x)$ represents the sample-specific model that is trained for each sample j using scalable tree boosting with the XGBoost (with default parameters and 50 stumps used in the decision trees). During training, $y_{i,j}$ is defined using the set of all reference transcript annotations T:

$$y_{i,j} = \left\{1\ \mathit{if}\ \mathit{introns}_{RC_{i,j}} = \mathit{introns}_t\ \mathit{for\ any}\ t \in T,\ 0\ \mathit{otherwise}\right\}$$

To reduce the noise during training, we exclude any read class with only single read count and read classes that do not overlap with any reference annotation (novel genes). Because Bambu relies on splice-junction coordinates to define $y_{i,j}$ during training, single exon read classes are excluded. However, Bambu optionally trains a separate model for read classes which consist of only a single exon, and which can predict the TPS for these transcript candidates (option min.txScore.singleExon = 0 in Bambu argument opt.discovery, see Supplementary Text Section 4).

2.1 The pre-trained model in Bambu

Training of $f(x)$ requires annotations to define class labels. When less than 1000 annotated transcripts are expressed in the sample (very poorly annotated genomes), training a sample-specific model is not supported due to insufficient training data from annotated transcripts. In this case, a pre-trained generic model can be used. Model pre-training can be done with a well annotated genome that is closely related to the genome of interest, or a pre-trained model from Bambu can be used. The default pre-trained model is based on the SGNex_HepG2_directRNA_replicate5_run1 sample from SG-NEx resource[33] which is applicable to any species with similar characteristics as the human genome (Supplementary Text Figure 1.b). The pre-trained model is also used to recommend a transcript discovery threshold, see Methods (below), Supplementary Text Section 5 and the Bambu online documentation for additional details.

2.1.1 Automatic recommendation procedure to use a pre-trained model—

More complete annotations will result in a more accurate transcript discovery model learned by Bambu. As the pre-trained model is assumed to be learned on very well annotated genomes, it shows higher accuracy compared to a sample-specific model when annotations are poor or the data is very noisy. If the estimated fraction of missing transcripts is below 50% (see below for details on the estimation of the fraction of missing transcripts), Bambu will recommend to use a pre-trained model, resulting in a mean improved performance for these samples even when the minimum annotation threshold to train a sample-specific model is satisfied (Supplementary Figure 1h).

2.2 Multiple Sample Transcript Discovery

In Bambu, transcript discovery is performed across all samples $j = 1, \ldots, N$. Here we define the Transcript Probability Score for a read class i as the probability that it is a valid transcript in at least 1 sample:

$$TPS_i = P\left(\left(\sum_{j=1}^{N} y_{i,j}\right) > = 1\right) = 1 - P\left(\sum_{j=1}^{N} y_{i,j} = 0\right)$$

To avoid that technical replicates inflate the multi-sample TPS, we provide a greatest lower bound using the maximum TPS for read class i across all samples:

$$TPS_i = 1 - P\left(\sum_{j=1}^{N} y_{i,j} = 0\right) = 1 - \prod_{j=1}^{N} \left(P\left(y_{i,j} = 0\right)\right) = 1 - \prod_{j=1}^{N} \left(1 - P\left(y_{i,j} = 1\right)\right) > = 1 - min_{j=1,\ldots,N}\left(1 - P\left(y_{i,j} = 1\right)\right) = max_{j=1,\ldots,N}\left(P\left(y_{i,j} = 1\right)\right) = max_{j=1,\ldots,N}\left(TPS_{RC_{i,j}}\right)$$

For a comparison of alternative definitions of the multi-sample TPS, see the Supplementary Text 3. In the case of a single sample, $TPS_{RC_i} = TPS_{RC_{i,j}}$, therefore we use TPS to refer to the multi-sample TPS for simplicity.

(3) Prediction of novel transcripts using the Novel Discovery Rate (NDR)

3.1 Transcript discovery threshold using the TPS

To identify novel transcripts, the TPS for each read class is estimated. Novel transcripts are then identified based on a TPS threshold p:

$$\widehat{y_i} = \{1 \ if \, TPS_i > \ p, \ 0 \, otherwise\}$$

For each threshold p, the number of read classes predicted to be valid transcripts is defined as

$$R_p = \sum_{i=1}^{M} \widehat{y_i}$$

3.2 Limitations of sample-specific thresholds

Refer to Table 1 for definition of error rates in the following section. The TPS can be used to rank transcripts by their probability of being valid. The rank of read class i is defined as the number of read classes with less or equal TPS:

$$K_i = \sum_{r=1}^{M} 1\left(TPS_r \le TPS_i\right)$$

with the function r providing the map between i and K_i:

$$r(K_i): K_i \rightarrow i, \text{ i.e., } RC_i \Leftrightarrow RC_{r(K_i)}.$$

The number of true positive (valid) novel transcripts associated with threshold p is then defined as

$$V_p{}^{TP} = \sum_{K_i = 1}^{R_p} y'_{r(K_i)}$$

with y'_i indicating the unknown true class label of read class i. The number of false positive novel transcripts associated with threshold p is defined as

$$V_p{}^{FP} = \sum_{K_i = 1}^{R_p} (1 - y'_{r(K_i)})$$

The False Discovery Rate (FDR) associated with p is then defined as

$$FDR_p = E\left[V_p{}^{FP} / R_p\right]$$

Furthermore, we define the Valid Discovery Rate (VDR) as the expected number of valid novel transcripts associated with p:

$$VDR_p = E\left[V_p{}^{TP} / R_p\right]$$

The VDR can be interpreted as the expected fraction of missing transcripts in the annotations. Since the scale of the TPS can differ with each new analysis based on the samples and annotations that are provided during model training (i.e. $f_j(x) \neq f_{j'}(x)$’ for any two different analysis samples j and j'), the same threshold p can lead to results with a different false discovery rates ($FDR_{p_j} \neq FDR_{p_{j'}}$) making the selecting of a meaningful and consistent threshold p difficult in practice (the same limitation applies to all sample-specific parameters such as read count). To address this in Bambu, users can specify the expected novel discovery rate. Bambu then selects the optimal threshold p for the specified NDR.

3.3 The Novel Discovery Rate: definition and calculation

To obtain a calibrated score that provides comparable and interpretable thresholds, we define the Novel Discovery Rate (NDR) as the expected number of non-annotated transcripts:

$$\begin{aligned} NDR = E[V/R] = E\left[(V^{FP} + V^{TP})/R\right] = \ E[V^{FP}/R] + E[V^{TP}/R] \\ = FDR + VDR \end{aligned}$$

Therefore, the NDR can be interpreted as an upper limit of the FDR for very well annotated genomes (FDR>VDR) or as an upper limit of the VDR for poorly annotated genomes ($FDR \ll VDR$) (see below for additional information on interpretation of the NDR). To select the threshold p that maximises the number of novel transcripts for the specified target NDR, we first calculate the number of novel transcripts associated with each candidate threshold p' as:

$$V_{p'} = \sum_{K_i = 1}^{R_{p'}} \left(1 - y_{r(K_i)}\right)$$

With y_i indicating if read class RC_i matches any annotation that is provided (see above). The observed $NDR^{o}{}_{p'}$ for p' is then calculated as:

$$NDR^{o}{}_{p'} = V_{p'}/R_{p'}$$

We then select the threshold p which corresponds to the largest number of novel transcripts such that the observed NDR is below or equal the target NDR:

$$p = argmax_p\left(NDR^{o}{}_{R_p}\right)\ such\ that\ NDR^{o}{}_{R_p} \le NDR$$

The maximum number of novel transcript candidates for the target NDR is then

$$R_{NDR} = R_p$$

Each read class is then associated with an optimal threshold $p \le TPS_i$ such that $NDR_P \le NDR_i$, and the NDR associated with each read class is then estimated as

$$NDR_i = min_p\left(NDR_p\right)\ such\ that\ p < TPS_i$$

Therefore, $NDR_i \le NDR_{i'}\ if\ TPS_i \ge TPS_{i'}$

3.4 Properties and Interpretation of the NDR

The interpretation of the NDR differs for very well annotated genomes (transcript discovery) and poorly annotated genomes (genome annotation):

3.4.1 Transcript discovery for well annotated genomes (FDR>VDR)—

For well annotated genomes such as the human genome, it is expected that most transcripts are annotated (FDR>VDR). In this scenario, the NDR provides an upper limit of the FDR:

$$FDR_p \le NDR_p$$

Here, the NDR enables the selection of a threshold p that maximises the number of novel transcripts for a desired maximum FDR. Unlike sample-specific threshold, the NDR ensures that the maximum FDR is comparable across samples or analyses: $FDR_{p_j} \le NDR \ge FDR_{p_{j'}}$ with p_j and $p_j{}'$ corresponding to the optimal threshold for the target NDR in sample j and j'

respectively. Since VDR>0 in most cases, the NDR is a conservative estimate of the FDR, with the true $FDR \ll NDR$.

3.4.2 Genome annotation (transcript discovery for poorly annotated genomes)—

For poorly annotated species (for example from newly assembled genomes), it is expected that a large fraction of novel transcripts are missing, in which case FDR<<NDR. In this scenario the NDR does not provide a meaningful upper limit of the FDR. However, the FDR is expected to increase dramatically once the NDR exceeds the expected VDR. To avoid this when Bambu is used for genome annotation, the NDR can be used as an a upper limit on the expected fraction of novel transcripts:

$$VDR_p \leq NDR_p$$

In this scenario, the NDR enables users to select a threshold for transcript discovery that ensures that the maximum number of transcripts is selected such that the expected VDR for high quality annotations is not exceeded, thereby preventing over-annotation with false positive transcripts. Here the NDR is also a conservative estimate of the VDR to ensure that the number of false positive transcripts is minimised.

3.5 Automatic estimation of fraction of missing transcripts

Bambu automatically estimates the fraction of missing annotations (1- completeness of annotations) at a maximum FDR of 0.1 to (1) recommend a default NDR threshold and (2) suggest the use of a pre-trained model for increased accuracy (see above for additional details). To estimate the completeness of the annotations that are pvided, the TPS from the pre-trained model is used. During model training, it is assumed that a complete set of annotations is provided, which ensures that 1 - NDR provides an approximation of the precision. Based on this assumption, Bambu identifies the TPS threshold ($\widehat{TPS}$) from the pre-trained model corresponding to a precision of 0.9 (i.e., NDR = 0.1). Using the default pretrained model across the SG-NEx data, the mean TPS at this NDR threshold is 0.891 with a SD of 0.065 suggesting the relationship between the TPS and NDR is robust. Bambu then infers the completeness of annotations as the fraction of novel transcripts among all transcripts with $TPS > \widehat{TPS}$ when the pre-trained model is applied. While the Bambu default pre-trained model is based on human RNA-Seq data, any pre-trained model can be used to estimate the completeness of annotations.

3.6 Estimation of a dynamic default NDR

The completeness of the reference annotations should be considered when choosing a NDR threshold: For well annotated genomes, a stringent NDR is recommended, whereas in the case of largely unannotated genomes, a larger fraction of novel transcripts is expected, and a more sensitive NDR threshold is recommended. To moderate the impact of the number of reference annotations on transcript discovery when the NDR is not explicitly specified, Bambu uses a dynamic default NDR which corresponds to an expected maximum False Discovery Rate of 10% for the samples and annotations of interest. To achieve this, Bambu uses the estimated fraction of missing annotations at a FDR of 10% as the default NDR.

We would like to note that this is a conservative estimate to minimise the number of false positives. For use cases when sensitivity is most important and higher false discovery rates are acceptable (e.g., genome annotation), this will be an underestimate, and we recommend manually specifying the desired NDR.

(4) Additional filtering and final annotation output of transcript discovery

4.1. Additional filtering

Bambu provides additional user-defined filters for which read classes will be considered as novel transcripts. In particular, read classes that (1) are possible degradation or sequencing artefacts (remove.subsetTx, default: TRUE), (2) have a read count below a minimum threshold (min.readCount, default = 2), or (3) have a gene read proportion below a minimum threshold (min.readFractionByGene, default = 0.05) can be removed. Furthermore, the minimum number of samples in which a read class has to pass the min.readCount and min.readFractionByGene filters can be adjusted. Bambu also allows the specification of the minimum distance to annotated transcripts for read classes (min.exonDistance) to be considered as novel, and the minimum overlap (min.exonOverlap) to merge novel unspliced transcript candidates with annotations. Read classes that pass all the filters, contain high confidence junctions and which have a NDR below the (user defined or default dynamic) NDR threshold will then be retained as novel transcript candidates (novel read class candidates). All read classes that pass and that do not pass these filters will be retained and used for quantification.

4.2. Integration with reference annotations

Bambu compares all novel read class candidates with the reference annotations. Read classes which have similar exon junctions compared to reference annotations will not be considered novel transcripts (see online documentation for additional distance thresholds that can be specified). Remaining novel read class candidates will then be included as new transcripts in the reference annotations that are provided (if any are provided). Novel transcripts which overlap with any existing gene are assigned to this gene id, and novel transcripts which overlap with any other novel transcript that is assigned to an annotated gene are iteratively assigned to the same gene id. Bambu then classifies novel transcripts of annotated genes according to the overlap with reference annotations as containing new first exons, new last exons, or new internal exons. Overlapping novel transcripts which are not assigned to any annotated gene will be grouped as novel genes and assigned a novel gene id. Bambu returns the combined reference annotations and novel transcript (referred to as the extended annotation).

(5) Quantification

5.1 Read class to transcript assignment

Using the extended reference annotations, Bambu then assigns each error-corrected read class to a set of reference transcripts based on the compatibility of splice junctions and the distance to the most similar reference transcripts. To avoid mis-assignments due to incorrect start and end sites, the first and last exons are optionally excluded. Quantification is performed separately for each sample using the same set of extended annotations. In the following we refer to transcripts as t_i, with $i = 1, .., N$, using j as the index for reads, read classes or equivalent read classes (see below for definition).

5.1.1 Compatibility of read classes with transcripts and genes—

Read classes are assigned to all transcripts and genes which are *compatible:*

5.1.1.1 Transcript compatibility: For a read class RC_j to be considered compatible with transcript t_i, Bambu requires that all splice junctions of RC_j are similar to a continuous set of splice junctions from t_i. By default, Bambu allows for up to 35 bp distance for each exon such that a read class is considered compatible with a transcript. Read classes which are not compatible with any transcript are considered *incompatible* and not considered for transcript quantification. All incompatible read classes are still assignable to genes, which allows us to use them for gene quantification.

5.1.1.2 Gene compatibility: For a read class RC_j to be considered compatible with gene g_i, Bambu requires that any exon of RC_j overlaps by at least 35 bp with any exon from gene g_i. Read classes which are incompatible with all transcripts, but compatible with genes are still used to obtain more accurate gene expression estimates (see below for additional details).

5.1.2 Definition of full length (equal) and partial read classes—

Read classes for which (a) all splice junctions are present in a transcript and (b) which are compatible with this transcript are considered to be *equal* to that transcript. All reads that correspond to equal read classes are counted as full-length reads. Compatible read classes which are not equal to any transcript are considered to represent non-full length (*partial*) reads.

5.1.3 Definition of equivalence read classes (equiRCs)—

Read classes which are compatible with the same set of transcripts will be summarised into *equivalent read classes* (equiRCs). To leverage the advantage of long read RNA-Seq to generate full-length reads, equal and partial read classes are summarised as different equiRCs. For each equiRC j, we observe the number of reads that form this read class (n_j), and the set of transcripts I_j that are compatible.

With this definition, equiRCs can be summarised into five categories (Table 1, Supplementary Figure 8): 1) Full Intron Match equiRC (FIM): equiRCs equally aligned to a unique transcript; 2) Subset Intron Match equiRC (SIM): equiRCs partially aligned to a unique transcript; 3) Multiple Full Intron Match equiRC (MFIM): equiRCS equally aligned to multiple transcripts (due to the existence of very similar transcripts where only first or last exons are different); 4) Full and Subset Intron Match equiRC (FSIM): equiRCS equally aligned to a transcript while partially aligned to one or more longer transcripts; this occurs when transcripts are subsets of other (longer) transcripts, and; 5) Multiple Subset Intron Match equiRC (MSIM) : equiRCs of fragmented reads that aligned partially to multiple transcripts.

5.2 Generative model

We followed the conventional generative model used for transcript quantification with specific changes to allow quantification of total expression and full length and unique reads from long read RNA-Seq data[7,9,10]:

Bambu estimates the read count τ_i for each isoform $i = 1, \ldots, M$ in the sample. We assume that each read $r = 1, \ldots, N$ originate from a single transcript i, which we describe as $R^t{}_{ri}$:

$$R^t{}_{ri} = \{1 \ if\ read\ r\ originates\ from\ transcript\ i,\ 0\ \ else\}$$

with $P\left(R^t{}_{ri} = 1\right) = \theta_i$ describing the probability that a read r originates from transcript i, which we also refer to as the relative transcript abundance, and $S^t{}_i$ describing the set of reads that originate from transcript i. Then τ_i is defined as:

$$\tau_i = \sum_{r=1}^{N} R^t{}_{ri} = \left|S^t{}_i\right|$$

with the expected value for τ_i obtained as:

$$E[\tau_i] = E\left[\sum_{r=1}^{N} R^t{}_{ri}\right] = \sum_{r=1}^{N} E\left[R^t{}_{ri}\right] N P\left(R^t{}_{ri} = \quad 1\right) = N\theta_i$$

Bambu summarises reads that have the same set of compatible transcripts into equivalent read classes (equiRCs). The assignment of a read r to equiRC j is then described as:

$$R^{rc}{}_{rj} = \{1 \ if\ read\ r\ matches\ equiRC\ j,\ 0\ else\}$$

with $j = 1 \ldots K$. Here we refer to the set of reads that are assigned to equiRC j as $S^{rc}{}_j$ with

$$\left|S^{rc}{}_j\right| = \sum_{r=1}^{N} R^{rc}{}_{rj} = \sum_{r \in S^{rc}{}_j} R^{rc}{}_{rj} = n_j$$

The set of transcripts that are compatible with equiRC j is described as $T^{rc}{}_j$, and the set of equiRCs which are compatible with transcript i are described by $RC^t{}_i$. The observed value of $R^{rc}{}_{rj}$, is always conditional on the (unknown) true read to transcript assignment $R^t{}_{rj}$. Here we define the conditional probability that a read r matches read class j given that it originates from transcript i as a_{ij}:

$$a_{ij} = P(R^{rc}{}_{rj} = 1 \middle| R^t{}_{ri} = 1)$$

with $\sum_{j=1}^{K} a_{ij} = 1$ for each transcript i. Without prior information, a_{ij} is assumed to be equally distributed among the set of equiRCs that are compatible with transcript i ($RC^t{}_i$ with $\left|RC^t{}_i\right| = K_i$), and 0 otherwise: $a_{ij} = \left\{\frac{1}{K_i} if\ j \in RC^t{}_i; 0\ \ if\ j \notin RC^t{}_i\right\}$.

For each read r, $R_{rij}^{rc,t}$ denotes if a read that is assigned to equiRC j originates from transcript i:

$$R_{rij}^{rc,t} = R^{rc}{}_{rj} \times R^{t}{}_{ri} = \left\{1 \ \ if \ R_{rj}^{rc} = 1 \ \ and \ R^{t}{}_{ri} = 1, \ 0 \ \ else\right\}$$

We can then calculate the probability as

$$P\left(R^{rc,t}{}_{rij} = \ 1\right) \ = \ P\left(R^{rc}{}_{rj} = 1 \ \cap R^{rc}{}_{rj} = 1\right) = P(R^{rc}{}_{rj} \ = 1 \middle| R^{t}{}_{ri} = 1)P\left(\ R^{t}{}_{ri} = 1\right) = a_{ij}\theta_{i}$$

given $R^{rc}{}_{rj}$ is dependent on $R^{t}{}_{rt}$.

The set of reads in equiRC j being generated from isoform i is then denoted as $S^{rc,t}{}_{ij}$, with the number of reads in equiRC j that originate from isoform i defined as n_{ij}:

$$\left|S^{rc,t}{}_{ij}\right| = n_{ij}$$

As n_{ij} is a sum of i.i.d. Bernoulli random variables ($R_{rij}^{rc,t}$), n_{ij} follows a binomial distribution:

$$n_{ij} = \sum_{r=1}^{N} R_{rij}^{rc,t} \sim Binom\left(N, p_{ij}\right)$$

with $p_{ij} = P\left(R_{rij}^{rc,t} = 1\right) = a_{ij}\theta_{i}$.

As Bambu works at the level of read classes for increased computational efficiency, first n_{ij} is inferred (see below) and the read count for each isoform i, τ_{i} is then calculated as

$$\tau_{i} = \sum_{j=1}^{K} n_{ij}$$

5.3 Parameter estimation using Expectation Maximisation (EM)

To estimate the transcript abundance, we use an Expectation Maximisation (EM) algorithm, that iteratively optimises the likelihood of the relative transcript abundance parameter ($\theta = \{\theta_{1}, \theta_{2}, \ \ldots \ , \ \theta_{M}\}$) given the observed read-to-read class assignments ($R'^{rc} = \left\{R'^{rc}{}_{rj}, \ r = 1\ldots N, \ j = 1\ldots K\right\}$, with $\left|S'^{rc}{}_{j}\right| = n'{}_{j}$) and the latent data that describes for each read r the transcript that generated this read ($R'^{t} = \left\{R'^{t}{}_{ri} : r = 1\ldots N, \ i = 1\ldots M\right\}$, with $\left|S'^{t}{}_{i}\right| = \tau'{}_{i}$). In practice Bambu works at the level of read classes, therefore the latent data used in Bambu’s EM is the read assignment within each read class ($R'^{rc,t} = \{R'^{rc,t}_{rij}, \ r = 1\ldots N, \ i = 1\ldots M, \ j = 1\ldots K\}$, with $n'{}_{ij} = \left|S'^{rc,t}{}_{ij}\right|$ and $\tau'{}_{i} = \sum_{j=1}^{K} n'{}_{ij}$).

The complete likelihood function can be written as:

$$\begin{aligned}
&L(\theta | R'^{t}, R'^{rc}) \quad = L(\theta | n', n') \\
&= \prod_{i=1}^{M} \prod_{j=1}^{K} P(n_{ij} = n'_{ij} \mid \theta_i) \\
&= \prod\nolimits_{i=1}^{M} \prod\nolimits_{j=1}^{K} \frac{N!}{n'_{ij}!(N - n'_{ij})!} \prod\nolimits_{r \in S^{rc,t}{}_{ij}} P(R_{rij}^{rc,t} = 1 \mid \theta_i \\
&) \prod\nolimits_{r \notin s^{rc,t}{}_{ij}} P(R_{rij}^{rc,t} = 0 \mid \theta_i) \prod\nolimits_{i=1}^{M} \prod\nolimits_{j=1}^{K} \frac{N!}{n'_{ij}!(N - n'_{ij})!} (a_{ij}\theta_i)^{n'_{ij}} (1 - a_{ij}\theta_i)^{N - n'_{ij}}
\end{aligned}$$

Expectation Step: In the k^{th} expectation step (E-step), the latent data $\hat{n}_{ij}^{(k)}$ is estimated as the conditional expected value of n_{ij} given the observed data R'^{rc} and $\hat{\theta}_i^{(k-1)}$:

$$\hat{n}_{ij}^{(k)} = E[n_{ij} | R^{rc} = \hat{\theta}^{(k-1)}]$$

Since we observe the read to read class assignment (R'^{rc}), we estimate the conditional expected value for n_{ij} as

$$E[n_{ij} | \ R^{rc} = \ R'^{rc}] = \sum_{r=1}^{N} P(R^{rc,t}{}_{rij} = 1 \mid R_{rj}^{rc} = R'^{rc}_{rj}) = \sum_{r \in S^{rc}{}_{j}} P(R^{rc,t}{}_{rij} = 1 \mid R_{rj}^{rc} = R'^{rc}_{rj})$$

Using the following 2 relations (1) $P(R^{rc,t}{}_{rij} = 1) = P(R^{rc,t}{}_{rij} = 1 \mid R_{rj}^{rc} = R'^{rc}_{rj}) P(R_{rj}^{rc} = R'^{rc}_{rj})$ and (2) $P(R_{rj}^{rc} = R'^{rc}_{rj}) = \sum_{i=1}^{M} P(R^{rc,t}{}_{rij} = 1)$ we obtain:

$$E[n_{ij} | R^{rc} = \ R'^{rc}] = \sum_{r \in S^{rc}{}_{j}} \frac{P(R^{rc,t}{}_{rij} = 1)}{\sum_{i=1}^{M} P(R^{rc,t}{}_{rij} = 1)}$$

Therefore

$$E[n_{ij} | R^{rc} \ = \ R'^{rc}] = n_j \frac{\theta_i a_{ij}}{\sum_{i=1}^{M} \theta_i a_{ij}}$$

During the E-Step, we use the estimated values $\hat{\theta}_i{}^{(k-1)}$ from the previous ($(k-1)^{th}$) iteration (with $\widehat{\theta_i}{}^{(1)} = \frac{1}{M}$), the observed read count for equiRC $j(n'{}_j)$, and the predefined values for a_{ij} to calculate the estimated value $\widehat{n_{ij}}{}^{(k)}$ for the k^{th} iteration:

$$\widehat{n_{ij}}{}^{(k)} = n'{}_j \frac{\widehat{\theta_i}{}^{(k-1)} a_{ij}}{\sum_{i=1}^{M} \widehat{\theta_i}{}^{(k-1)} a_{ij}}$$

With the estimated transcript abundance for iteration k being calculated as $\hat{\tau}_i{}^{(k)}$:

$$\hat{\tau}_i^{(k)} = \sum\nolimits_{j=1}^{K} \hat{n}_{ij}^{(k)}$$

Maximisation Step—In the Maximisation Step (M-Step), we obtain the maximum likelihood estimate for the unknown parameter θ given the estimate for the latent data ($\widehat{n_{ij}}^{(k)}$):

$$\hat{\theta}^{(k)} = argmax_{\theta}\left(log(L(\theta \mid \widehat{n_{ij}}^{(k)}))\right)$$

The maximum likelihood estimate can be obtained as follows:

$$log\left(\left(L(\theta | \widehat{n_{ij}}^{(k)}\right)\right) = log\left(\prod_{i=1}^{M}\prod_{j=1}^{K}\frac{N!}{\widehat{n_{ij}}^{(k)}!\left(N - \widehat{n_{ij}}^{(k)}\right)!}(a_{ij}\theta_i)^{\widehat{n_{ij}}^{(k)}}(1 - a_{ij}\theta_i)^{N - \widehat{n_{ij}}^{(k)}}\right)$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{K}\left((\widehat{n_{ij}}^{(k)}log(\theta_i a_{ij}) + \left(N - \widehat{n_{ij}}^{(k)}\right)log(1 - \theta_i a_{ij}) + log(\frac{N!}{\widehat{n_{ij}}^{(k)}!\left(N - \widehat{n_{ij}}^{(k)}\right)!})\right)$$

Taking derivative with respect to θ_i:

$$\frac{\vartheta l\left(\left(L(\theta \mid \widehat{n_{ij}}(k)\right)\right)}{\vartheta\theta_i} = \sum_{j=1}^{K}\left(\frac{\widehat{n_{ij}}(k)a_{ij}}{\theta_i a_{ij}} - \frac{\left(N - \widehat{n_{ij}}(k)\right)a_{ij}}{1 - \theta_i a_{ij}}\right)$$

$$= \sum_{j=1}^{K}\frac{\left(\widehat{n_{ij}}(k)(1 - \theta_i a_{ij}) - \left(N - \widehat{n_{ij}}(k)\right)\theta_i a_{ij}\right)}{\theta_i a_{ij}(1 - \theta_i a_{ij})}$$

$$= \sum\nolimits_{j=1}^{K}\frac{\left(\widehat{n_{ij}}(k) - N\theta_i a_{ij}\right)}{\theta_i a_{ij}(1 - \theta_i a_{ij})}$$

By setting the above equation to 0, therefore, we have

$$\hat{\theta}_i(k) = \frac{\sum_{j=1}^{K}\widehat{n_{ij}}(k)}{\sum_{j=1}^{K}Na_{ij}} = \frac{\sum_{j=1}^{K}\widehat{n_{ij}}(k)}{N} = \frac{\sum_{j \in RC_i^t}\widehat{n_{ij}}(k)}{N}$$

5.4 Zero-count equiRCs

Since long read RNA-Seq does not include a fragmentation step, most equivalent classes that are theoretically possible based on overlapping transcript intervals will not be observed in practice. For most transcripts (with moderate to high expression) the set of possible read classes is equal to the set of observed read classes. However, for transcripts which are very lowly expressed or inactive, the equiRCs representing the full-length transcript may not be observed, which can lead to an overestimation when such transcripts share read classes with highly expressed transcripts. To address this, we define the set of possible read classes as the set of observed read classes plus the set of full length read classes which are not observed. We set the read count for non-observed read classes to zero (zero-count, or empty read classes). This set of read classes will be used for quantification and to obtain the parameter a_{ij}.

5.5 Full-length and unique support

To keep track of full-length reads, we define a (observed) indicator variable $I^{f}{}_{ij}$, which takes a value of 1 if reads in read class j that originate from transcript i are full length reads and 0 otherwise. We can use $1 - I^{f}{}_{ij}$ to denote the indicator variable when reads in read class j that originate from transcript i are partial (non-full-length) reads. We then obtain the full-length transcript estimate for each transcript ($\widehat{\tau^{f}{}_{i}}$) as:

$$\widehat{\tau^{f}{}_{i}} = \sum_{j=1}^{K} \widehat{n_{ij}} I^{f}{}_{ij}$$

And we have the partial transcript estimate for each transcript ($\widehat{\tau^{p}{}_{i}}$) as

$$\widehat{\tau^{p}{}_{i}} = \sum_{j=1}^{K} \widehat{n_{ij}} (1 - I^{f}{}_{ij})$$

Similarly, to keep track of unique read support, we define a (observed) indicator variable $I^{u}{}_{ij}$, which takes a value of 1 if reads in read class j that originate from transcript i are uniquely mapped to transcript i and 0 otherwise. We then obtain the unique transcript estimate for each transcript ($\widehat{\tau^{u}{}_{i}}$) as

$$\widehat{\tau^{u}{}_{i}} = \sum_{j=1}^{K} \widehat{n_{ij}} I^{u}{}_{ij}$$

Implementation of Bambu was done in R using Rcpp[28,29], Bambu is available through Bioconductor.

Long read RNA-Seq data

For this analysis, we have used SG-NEx core cell lines include A549, K562, Hct116, HepG2, MCF7, HEYA8 and human embryonic stem cell (hESC) cell line generated using cDNA, direct cDNA, and direct RNA protocols. Processed fastq and genome alignment bam files were used for different methods.

Transcript discovery evaluation

Details on transcript discovery evaluation can be found in Supplementary Notes 1.

Transcript quantification with context-specific annotations

Details on transcript quantification evaluation can be found in Supplementary Notes 2.

Full-length and unique read support

Details on full-length and unique read support evaluation analysis can be found in Supplementary Notes 3.

Quantification of retrotransposon-derived isoforms

Details on quantification of retrotransposon-derived isoforms can be found in Supplementary Notes 4.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by funding from the Agency for Science, Technology, and Research (A*STAR) and the National Medical Research Council (NMRC). M.L. was supported by R01 HG009937.

Data Availability

The SG-NEx samples are available through github (https://github.com/GoekeLab/sg-nex-data), ENA (PRJEB44348), and AWS open data (https://registry.opendata.aws/sgnex/). Processed data associated with figures and tables are available on code ocean[56]. The PacBio data is available through SRA (SRP036136). The Arabidopsis data is available through ENA (PRJEB32782).

References

1. Matlin AJ, Clark F & Smith CWJ Understanding alternative splicing: towards a cellular code. Nat. Rev. Mol. Cell Biol 6, 386–398 (2005). [PubMed: 15956978]
2. Blencowe BJ Alternative splicing: new insights from global analyses. Cell 126, 37–47 (2006). [PubMed: 16839875]
3. Pan Q, Shai O, Lee LJ, Frey BJ & Blencowe BJ Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet 40, 1413–1415 (2008). [PubMed: 18978789]
4. Ben-Dov C, Hartmann B, Lundgren J & Valcárcel J Genome-wide analysis of alternative pre-mRNA splicing. J. Biol. Chem 283, 1229–1233 (2008). [PubMed: 18024428]
5. Graveley BR Alternative splicing: increasing diversity in the proteomic world. Trends Genet. 17, 100–107 (2001). [PubMed: 11173120]
6. Nilsen TW & Graveley BR Expansion of the eukaryotic proteome by alternative splicing. Nature 463, 457–463 (2010). [PubMed: 20110989]
7. Li B & Dewey CN RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011). [PubMed: 21816040]
8. Trapnell C et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol 31, 46–53 (2013). [PubMed: 23222703]
9. Bray NL, Pimentel H, Melsted P & Pachter L Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol 34, 525–527 (2016). [PubMed: 27043002]
10. Patro R, Duggal G, Love MI, Irizarry RA & Kingsford C Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419 (2017). [PubMed: 28263959]
11. Wang D et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Mol. Syst. Biol 15, e8503 (2019). [PubMed: 30777892]
12. Gonzàlez-Porta M, Frankish A, Rung J, Harrow J & Brazma A Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. 14, R70 (2013). [PubMed: 23815980]
13. Conesa A et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 17, 13 (2016). [PubMed: 26813401]
14. Deschamps-Francoeur G, Simoneau J & Scott MS Handling multi-mapped reads in RNA-seq. Comput. Struct. Biotechnol. J 18, 1569–1576 (2020). [PubMed: 32637053]
15. Sarkar H, Srivastava A, Bravo HC, Love MI & Patro R Terminus enables the discovery of data-driven, robust transcript groups from RNA-seq data. Bioinformatics 36, i102–i110 (2020). [PubMed: 32657377]

16. Pardo-Palacios F et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. Research Square (2021) doi:10.21203/rs.3.rs-777702/v1.
17. Tang AD et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat. Commun 11, 1438 (2020). [PubMed: 32188845]
18. Wyman D et al. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. bioRxiv 672931 (2020) doi:10.1101/672931.
19. Kovaka S et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 20, 278 (2019). [PubMed: 31842956]
20. Prjibelski AD et al. Accurate isoform discovery with IsoQuant using long reads. Nat. Biotechnol (2023) doi:10.1038/s41587-022-01565-y.
21. Soneson C, Matthes KL, Nowicka M, Law CW & Robinson MD Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. Genome Biol. 17, 12 (2016). [PubMed: 26813113]
22. Workman RE et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat. Methods 16, 1297–1305 (2019). [PubMed: 31740818]
23. Kuo RI et al. Illuminating the dark side of the human transcriptome with long read transcript sequencing. BMC Genomics 21, 751 (2020). [PubMed: 33126848]
24. Li H Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018). [PubMed: 29750242]
25. Wick RR, Judd LM & Holt KE Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 20, 129 (2019). [PubMed: 31234903]
26. Dempster AP, Laird NM & Rubin DB Maximum likelihood from incomplete data via theEMAlgorithm. J. R. Stat. Soc 39, 1–22 (1977).
27. Huber W et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods 12, 115–121 (2015). [PubMed: 25633503]
28. Eddelbuettel D et al. Rcpp: Seamless R and C++ integration. J. Stat. Softw 40, 1–18 (2011).
29. Eddelbuettel D Seamless R and C++ Integration with Rcpp. (Springer, New York, NY, 2013).
30. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing (2021).
31. Chen T & Guestrin C XGBoost: A Scalable Tree Boosting System. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (Association for Computing Machinery, 2016).
32. Hardwick SA et al. Spliced synthetic genes as internal controls in RNA sequencing experiments. Nat. Methods 13, 792–798 (2016). [PubMed: 27502218]
33. Chen Y et al. A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. BioRxiv (2021) doi:10.1101/2021.04.21.440736.
34. Pertea G & Pertea M GFF Utilities: GffRead and GffCompare. F1000Res. 9, 304 (2020).
35. Aken BL et al. The Ensembl gene annotation system. Database 2016, (2016).
36. Parker MT et al. Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. Elife 9, e49658 (2020). [PubMed: 31931956]
37. Berardini TZ et al. The Arabidopsis information resource: Making and mining the ‘gold standard’ annotated reference plant genome. Genesis 53, 474–485 (2015). [PubMed: 26201819]
38. Tilgner H, Grubert F, Sharon D & Snyder MP Defining a personal, allele-specific, and single-molecule long-read transcriptome. Proc. Natl. Acad. Sci. U. S. A 111, 9869–9874 (2014). [PubMed: 24961374]
39. Gleeson J et al. Accurate expression quantification from nanopore direct RNA sequencing with NanoCount. Nucleic Acids Res. (2021) doi:10.1093/nar/gkab1129.
40. Liao Y, Smyth GK & Shi W featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930 (2014). [PubMed: 24227677]
41. Hu Y et al. LIQA: long-read isoform quantification and analysis. Genome Biol. 22, 182 (2021). [PubMed: 34140043]

42. Tian L et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. Genome Biol. 22, 310 (2021). [PubMed: 34763716]

43. Zhang Y et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. Nat. Genet 51, 1380–1388 (2019). [PubMed: 31427791]

44. Lu X et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat. Struct. Mol. Biol 21, 423–425 (2014). [PubMed: 24681886]

45. Kelley D & Rinn J Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 13, R107 (2012). [PubMed: 23181609]

46. Göke J & Ng HH CTRL+INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. EMBO Rep. 17, 1131–1144 (2016). [PubMed: 27402545]

47. Berrens RV et al. Locus-specific expression of transposable elements in single cells with CELLO-seq. Nat. Biotechnol (2021) doi:10.1038/s41587-021-01093-1.

48. Semenick D TESTS AND MEASUREMENTS: The T-test. Strength & Conditioning Journal 12, 36 (1990).

49. Massey FJ The Kolmogorov-Smirnov Test for Goodness of Fit. J. Am. Stat. Assoc 46, 68–78 (1951).

50. Smit AFA, Hubley R & Green P RepeatMasker. RepeatMasker http://repeatmasker.org (1996).

51. Wyman D et al. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. 48 (2019).

52. Soneson C et al. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. Nat. Commun 10, 3359 (2019). [PubMed: 31366910]

53. Troskie R-L et al. Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome. Genome Biol. 22, 146 (2021). [PubMed: 33971925]

54. Garalde DR et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods 15, 201–206 (2018). [PubMed: 29334379]

55. Mulroney L et al. Identification of high confidence human poly(A) RNA isoform scaffolds using nanopore sequencing. RNA (2021) doi:10.1261/rna.078703.121.

56. Chen Ying , Sim Andre , Lee Joseph, Goeke Jonathan. Bambu [Source Code]. https://codeocean.com/capsule/3893005/tree/v2 (2023).

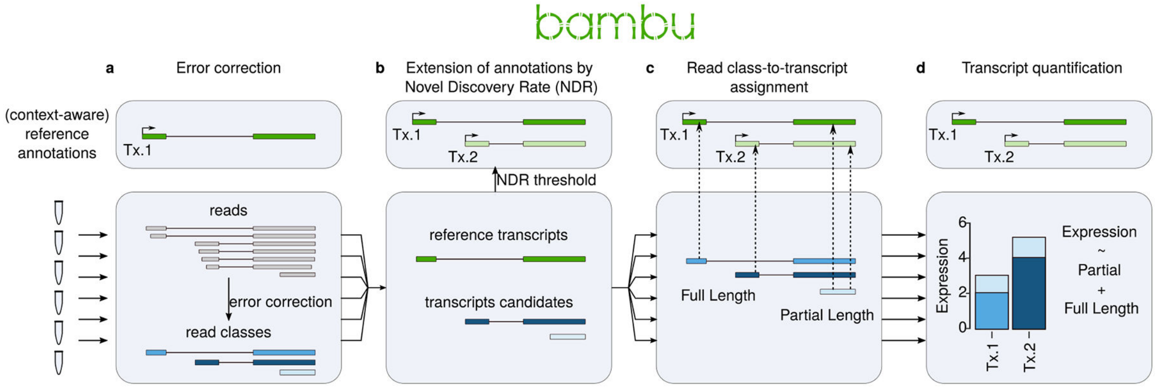
bambu
a
Error correction
b
Extension of annotations by Novel Discovery Rate (NDR)
c
Read class-to-transcript assignment
d
Transcript quantification
(context-aware) reference annotations
Tx.1
Tx.2
NDR threshold
reads
error correction
read classes
reference transcripts
transcripts candidates
Full Length
Partial Length
Expression
6
4
2
0
Tx.1
Tx.2
Expression ~ Partial + Full Length

Figure 1. Bambu enables simultaneous transcript discovery and quantification from Nanopore RNA-Seq data

Schematic illustration on how Bambu performs transcript discovery and quantification on Nanopore RNA-Seq data in four steps **(a)** For each sample, Bambu performs error correction on splice junctions of the aligned reads using input annotations **(b)** Performs transcript discovery jointly across samples at a given novel discovery rate (NDR) threshold and extends the input annotations with the retained novel transcripts **(c)** Assigns the read classes to transcripts in the extended annotation and categorises them as having full-length or partial overlaps **(d)** Performs probabilistic transcript quantification based on the read class to transcript assignment

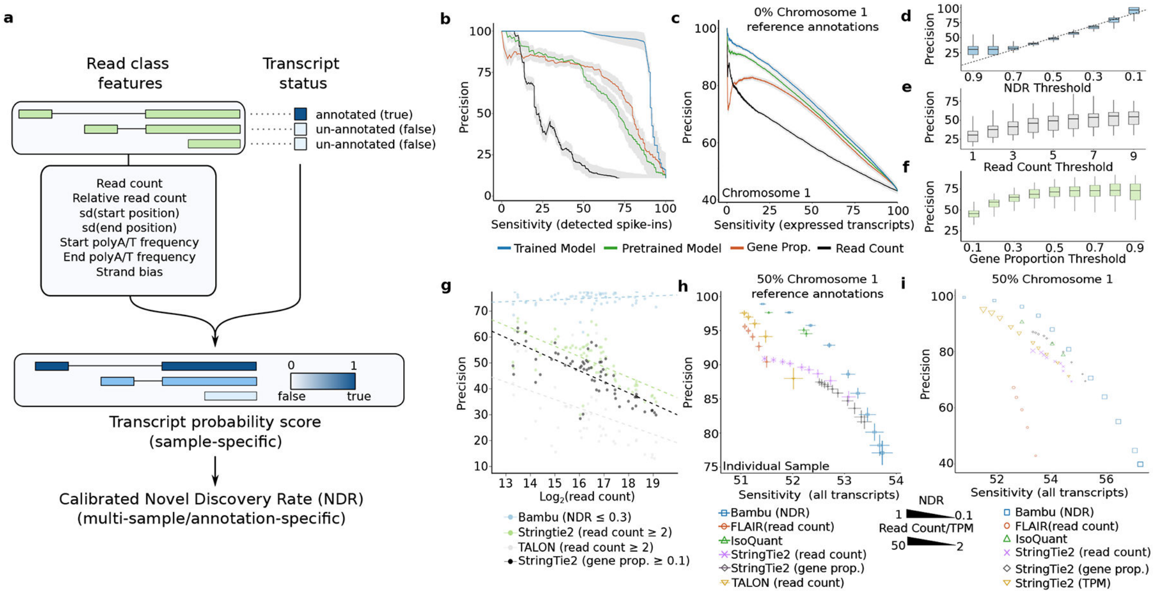

Figure 2. A calibrated machine learning full-length transcript classifier improves transcript discovery accuracy

(a) The schematic of transcript discovery steps performed by Bambu where 1) a machine learning model is trained on nine different features from read classes features to predict if a read class represents a full-length transcript, 2) the transcript probability score predicted in the first step is re-calibrated to a novel discovery rate across multiple samples **(b-c)** Average precision recall curves for the performance of transcript discovery for (b) SG-NEx spike-in data (n=8) and (c) all core SG-NEx data (n = 76)The model is evaluated on a (b) subset of the spike-in transcripts or (c) chromosome 1 after being trained on the other transcripts (blue), predictions from the generic model (green), or when read count (black) or gene proportion (red) is used alone as a classifier. The grey shaded area represents the mean +/− SE of the precision for each line. Sensitivity is measured as the percentage of all detected known transcripts. **(d-f)** Boxplot of the precisions of chromosome 1 read classes passing varying (d) NDR, (e) Read Count and (f) Gene Proportion thresholds across all core SG-NEx data (n = 76) all SG-NEx samples. **(g)** Each dot colour triplicate represents the precision from the same SG-NEx sample processed by either: Bambu with a NDR threshold of 0.3 (blue), StringTie2 with a read coverage threshold of 2 (green), TALON with a read count threshold of 2 (grey) and StringTie2 with a gene proportion threshold of 0.1 (black). Dotted lines represent a fitted linear regression for each tool **(h-i)** The average sensitivity and precision on (h) core SG-NEx samples (n = 76) and (i) when combining HepG2 SG-NEx samples (n=14) with 50% of human chromosome 1 annotations randomly removed. Each tool is displayed at several different parameter thresholds: Bambu (blue) with NDR thresholds, FLAIR (red), Stringtie 2 (purple) and (h) TALON (yellow) with read count/coverage thresholds. StringTie2 was also run with gene proportion thresholds (black) and (i) a varying TPM threshold (yellow). IsoQuant (green) points are run with varying "--model_construction_strategy". Error bars represent the mean +/− SD of the sensitivity and precision.

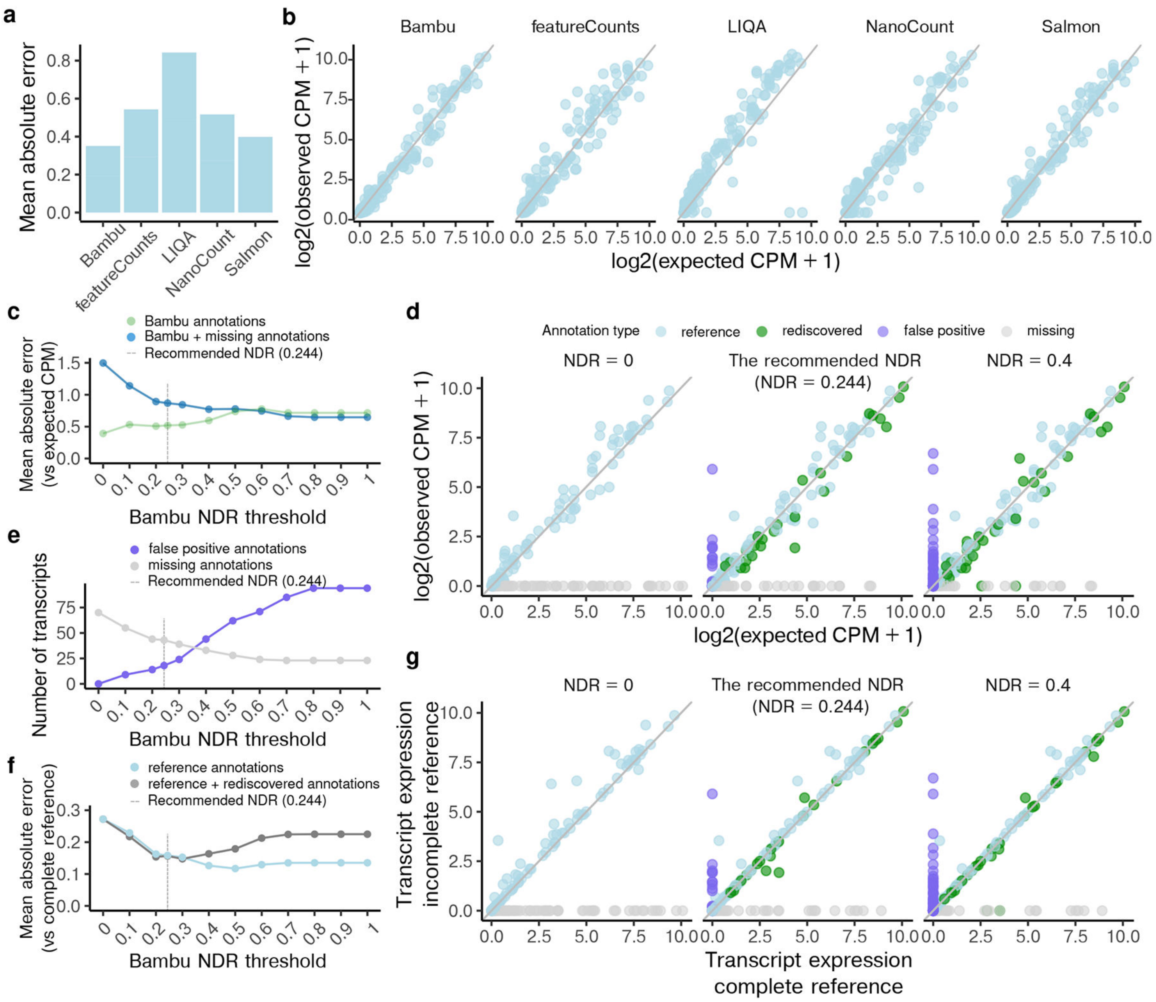

Figure 3. Transcript quantification on spike-in data shows improvement with varying novel discovery rates

(a) The mean absolute error (MAE) and **(b)** the scatterplots between log2 normalised transcript abundance estimates and expected spike-in abundance when applying Bambu, featureCounts, LIQA, NanoCount, and Salmon using full sequin annotations **(c)** The MAE between the log2 normalised transcript abundance estimates and expected spike-in abundance when applying Bambu using partial sequin annotation and with varying novel discovery rate (NDR) thresholds, for Bambu annotated transcripts, including annotations that are present in the reference (partial) sequin annotations, the annotations that have been artificially removed and rediscovered by Bambu, and also the false positive annotations discovered by Bambu (green), plus the annotations that are artificially removed from the partial annotation and remained missing after transcript discovery, i.e., missing annotations (blue) **(d)** The scatterplots between log2 normalised transcript abundance estimates and expected spike-in abundance when applying Bambu using partial annotations: without transcript discovery (NDR = 0), with default recommended NDR (0.244), and with a more sensitive NDR (0.4) **(e)** The number of missing (grey) and false positive (purple) transcripts when applying Bambu using partial sequin annotations with varying NDR thresholds **(f)** The MAE between log2 normalised spike-in transcript abundance estimates when applying

Bambu using full sequin annotation and using partial sequin annotation with varying NDR thresholds, for transcripts that are present in reference (light blue), transcripts that are present in reference and those rediscovered annotations (dark grey) **(g)** The scatterplots between log2 normalised transcript abundance estimates when applying Bambu using full sequin annotation and applying Bambu using partial sequin annotations: without transcript discovery (NDR = 0), with default recommended NDR (0.244), and with a more sensitive NDR (0.4) For (d) and (g), light blue dots represent transcripts that are present in the partial sequin annotations, green dots represent transcripts that have been artificially removed from the reference and rediscovered by Bambu, purple dots represent false positive transcripts, grey dots represent transcripts that have been artificially removed from the reference and remained missing after Bambu discovery. For (c), (e), and (f), the grey dotted line indicates the recommended NDR by Bambu (0.244)

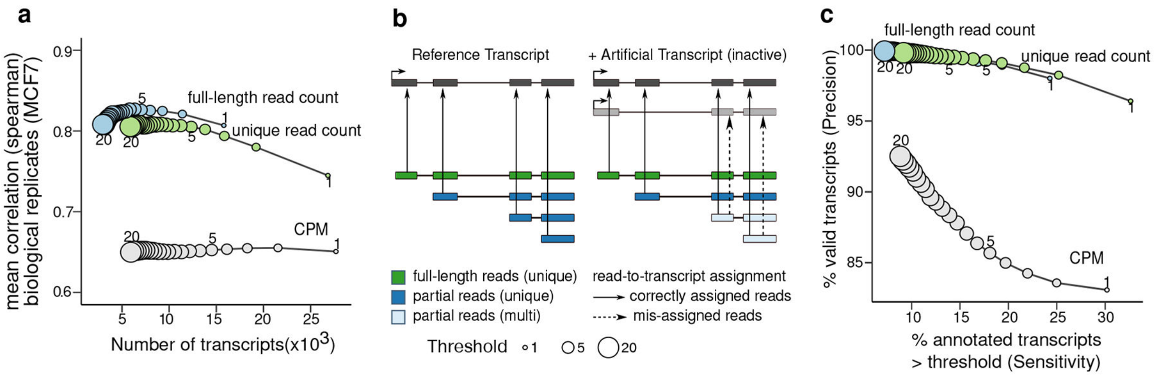

Figure 4. Full-length and unique read support provide evidence on expressed transcripts

(a) The average spearman correlation between the transcript abundance estimates for MCF7 replicates generated using direct cDNA and the number of expressed transcripts when different filtering methods and thresholds were applied to transcripts. Filtering is based on mean CPM or unique read count support being greater than the given threshold across the replicates **(b)** Full-length, unique read and partial reads and the potential read-to-transcript assignment for each of these read types **(c)** The sensitivity and precision of using full-length, unique read and CPM thresholds to filter out false positive transcripts overlapping with highly abundant isoforms that have no unique or full-length reads support at varying filtering thresholds from 1 to 20 on Hct116 samples. Filtering was based on the average values across replicates being not lower than the threshold.

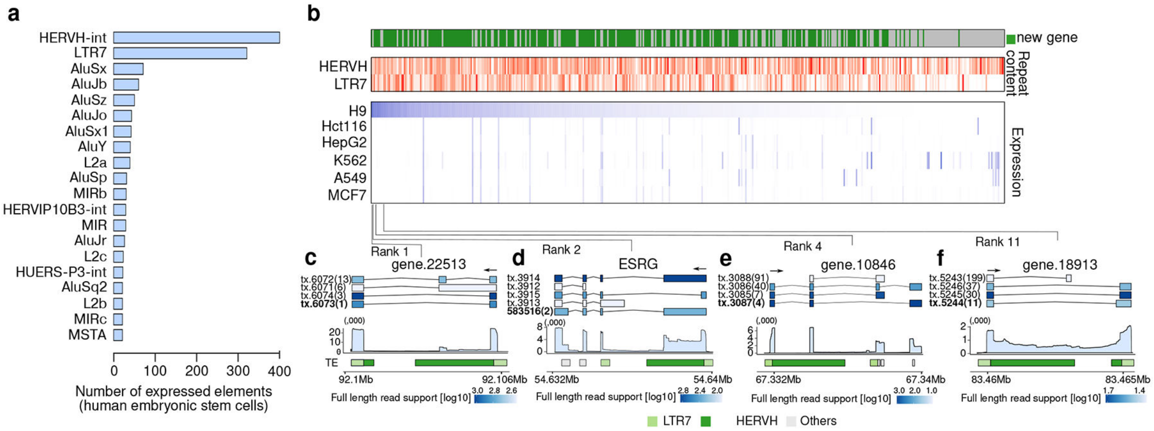

Figure 5. Bambu enables the discovery and quantification of highly repetitive genes

(a) Repeat families ranked by the number of expressed elements identified in the human embryonic stem cell cell line (H9) **(b)** Overview of the top 100 expressed novel and annotated transcripts that overlap with the HERVH-LTR7 retrotransposon in hESC cell line. For novel transcripts, we only include those with high overlap in novel exons. Top: Fraction of overlap with HERVH and LTR7. Bottom: Expression estimate in H9 hESCs and in the 5 SG-NEx cancer cell lines **(c-f)** Illustrations of highly expressed transcripts in H9 hESCs that originate from HERVH-LTR7 repeats that show distinct splicing patterns and transcript sequences. Top: transcript annotation colored by estimated full-length read support, with ranks in expression highlighted inside the bracket. Middle: mean read coverage for the specified genomic ranges for each selected gene. Bottom: show repeat masker colored by HERVH (green), LTR7 (light green), and all other repeat types (light grey).

Table 1:

Definition of error rates during transcript discovery

<table>
<tr><th></th><th>Invalid transcript</th><th>Valid transcript</th><th></th></tr>
<tr><td>Predicted valid transcript</td><td>$V^{FP} = V - V^{TP}$</td><td>$S + V^{TP}$</td><td>R</td></tr>
<tr><td>Predicted invalid transcript</td><td>$U + U'$</td><td>$T + U'$</td><td>$N - R$</td></tr>
<tr><td></td><td>$n_0 - n'$</td><td>$n_1 + n'$</td><td>M</td></tr>
<tr><td colspan="4"><b>observed:</b>
V: number of non annotated read classes with $TPS_i > p$
S: number of annotated read classes with $TPS_i > p$
U: number of non annotated read classes with $TPS_i < p$
T: number of annotated read classes with $TPS_i < p$
R: number of read classes with $TPS_i > p$
M: Total number of read classes
n_0: total number of non-annotated read classes
n_1: total number of annotated read classes</td></tr>
<tr><td colspan="4"><b>Not observed:</b>
V': number of non annotated valid read classes with $TPS_i > p$
U': number of non-annotated valid read classes with $TPS_i < p$
n': total number of non-annotated valid read classes</td></tr>
</table>

Table 2.

Definition of equivalence read class types

equiRC type	***Description***	***Equal (full-length)***	***Partial***	***Unique***
FIM	*Full Intron Match*	*X*	-	*X*
MFIM	Multiple Full Intron Match	X		-
MSIM	Multiple Subset Intron Match	-	X	-
SIM	*Subset Intron Match*	-	*X*	*X*
FSIM	*Full and Subset Intron Match*	*X*	*X*	-

Table 3.

Description of mathematical notations for transcript quantification

Mathematical notations	Descriptions
r	read index
i	transcript index
j	equivalence read classes (equiRC) index
M	total number of transcripts
K	total number of equivalence read classes (equiRC)
N	total number of reads, i.e., sequencing depth
$\tau_i\left(\tau_i{}^f, \tau_i{}^p, \tau_i{}^u\right)$	true read count for isoform i,superscripts f, p, u, represent true full-length, partial length, and unique alignment read count respectively
$R^t{}_{ri}$	the event that whether a read r originates from transcript i: takes a value of 1 when it happens and 0 otherwise (the realisation value)
$R^{rc}{}_{rj}$	the event that whether a read r matches a equiRC j
$R^{rc,\,t}{}_{rij}$	the event that whether a read that is assigned to equiRC j originates from transcript i
$S^t{}_i$	the set of reads that originate from transcript i
$S^{rc}{}_j$	the set of reads that are assigned to equiRC j
$S^{rc,\,t}{}_{ij}$	the set of reads in equiRC j being generated from isoform i
$T^{rc}{}_j$	the set of transcripts that are compatible with equiRC j
$RC^t{}_i$	the set of equiRCs which are compatible with transcript i
a_{ij}	the conditional probability that a read r matches to a equiRC $j$conditioning on the read r originates from transcript i
θ_i	global true relative expression of isoform i
n_{ij}	unobserved number of reads in equiRC j originate from isoform i
n_j	number of reads in equiRC j
$I^f{}_{ij}\left(I^u{}_{ij}\right)$	an indicator variable takes a value of 1 when reads in read class j that originate from transcript i are full length reads (unique reads) and 0 otherwise

The realisations of these variables are denoted with a prime symbol ($'$) on top.