



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.



Identifying Long Non-Coding RNA in the Chicken Transcriptome

By Richard Izen Kuo

Thesis Submitted for the Qualification Doctor of Philosophy, The University of Edinburgh, 2022

Abstract

The transcriptome remains a vast under explored space in genomics. Unlike the genome which is linear in nature, the use of alternative transcription start, end, and splicing sites in eukaryotes creates the possibility of near infinite differentially expressed RNA. While many expressed messenger RNA have been identified through the proteins that they produce, there is still very little known about the world of long non-coding RNA (lncRNA).

Long non-coding RNA are a vast unknown space and represent one of the largest frontiers of transcriptomics. While little is known about this class of RNA as a whole, there have been specific lncRNA which have been found to be crucial components of biological development. Given the characteristics of lncRNA there may also be a sub-class that is involved in cell differentiation and speciation. In order to explore lncRNA and generate high throughput predictions of their functions, I used the chicken as a model and applied comparative genomics using newly assembled genomes from other avian species.

Long non-coding RNA present the almost perfect scenario for evading detection from previous RNA discovery methods. They have been shown to be poorly conserved across species, with generally low expression levels and no downstream product that is immediately identifiable. Given these factors, previous RNA detection methods such as expressed sequence tags and RNA sequencing cannot provide reliable evidence for the mass identification of lncRNA.

In the first chapter I explore the characteristics of Iso-Seq (Pacific Biosciences long read RNA sequencing technology) and methods for processing the data to improve long non-coding RNA identification. I also explore the use of non-traditional cDNA library preparation methods including cDNA normalization and 5' cap selection. I found that the ability of long read RNA sequencing to provide full length transcript sequences allows for more robust methods of lncRNA prediction.

In the second chapter, I explore the data processing of long reads. I use a dataset generated by Pacific Biosciences using the Universal Human Reference RNA as an example of ideal long read data. By using data based on the human transcriptome, I was able to compare my results with information from one of the most well annotated and studied transcriptomes. I demonstrate the Transcriptome Annotation by Modular Algorithms (TAMA) software that I developed and how it can be used to explore the non-coding RNA within the transcriptome.

In the third chapter, I explore the transcriptome constructed from Iso-Seq data on different chicken tissue samples. I used the TAMA software along with other tools to make pipelines optimized for lncRNA discovery and to perform functional annotation. Using these methodologies I identified over 300,000 putative transcript models corresponding to over 50,000 genes. Of these over 100,000 transcript models appear to be lncRNA which correspond to over 38,000 gene loci. The majority of these are predicted as sense exonic and mono-exonic lncRNA. While it will require further investigation to produce sufficient evidence that these RNA are not the result of transcriptional noise, I have identified a subset of these which appear to have functional importance given their co-expression with known genes. I demonstrate that while lncRNA appear to be generally lowly expressed, they often express in a tissue-specific manner which suggests a possible role in tissue differentiation.

From these investigations, I have found that there are potentially thousands of unannotated lncRNA within the chicken transcriptome with characteristics that require new technologies such as long read sequencing to identify. These novel lncRNA include a subset which could have functional roles in the regulation of cell differentiation.

Lay Summary

Ribonucleic acids, more commonly referred to as RNA, are the direct biochemical products which are produced from the genome (DNA). While DNA is often thought of as a blueprint for life, the information in the genome must be converted in RNA to produce any functional units. Most of the currently known RNA are used as sequence templates to create proteins. This occurs by converting the sequence of an RNA molecule into amino acid chains which then fold into 3 dimensional structures known as proteins. These RNA are called messenger RNA (mRNA) are the most closely studied class of RNA. However, there is still a large frontier of undiscovered RNA of which we understand very little. The class of RNA with the greatest potential for new discoveries both in numbers and significance are known as long non-coding RNA (lncRNA).

lncRNA are special in that they do not get converted into proteins. They also have unique properties such as being very unique to each species and being produced at much lower levels than mRNA. We have already discovered a few lncRNA which serve vital functions to the development of humans and other mammals. It is possible that there are many more lncRNA which have important biological functions. Some lncRNA may even be responsible for the differences between individuals.

When studying RNA, researchers often focus on specific species known as model organisms. Model organisms are selected so that many groups of researchers can build knowledge of one species which can then be used to understand related species. This gives researchers a foundation of knowledge and commonality for comparing results. In this study, I use the chicken (*Gallus gallus*) as a model organism. The chicken is used as a representation and reference for other birds. The chicken provides an ideal model organism for studying lncRNA because there is already a great deal of information about its genome, biology and traits, with other closely related species also becoming better studied.

To discover new lncRNA in chickens, I used a new method of RNA detection called long read RNA sequencing. This new technology surpasses previous technologies by allowing for the identification of complete RNA sequences. Previous technologies only allowed for the identification of short fragments of RNA which then had to be pieced together like a puzzle to estimate the real RNA sequence. The ability to see full length RNA sequences on a genome-wide scale provides a huge advantage for lncRNA discovery. This is due to the generally unknown sequence characteristics of lncRNA which make them difficult to reconstruct from shorter pieces of information. Using the full length sequence information, I was able to develop analysis software and pipelines to distinguish lncRNA from mRNA. By comparing these predicted new lncRNA genes to other sources of information such as RNA expression data from short read RNA sequencing and previously published genome annotations, I was able to infer possible functions and/or relationships for the newly predicted lncRNA.

This study represents one of the first attempts to use long read RNA sequencing for high throughput lncRNA discovery and annotation. I demonstrate the power of long reads for this type of investigation as well as areas in both the laboratory processing and bioinformatic analyses that could be improved to make the overall methodology more powerful.

Using long read RNA sequencing with new RNA processing pipelines and bioinformatic tools, I was able to identify tens of thousands of possible novel lncRNA that are expressed from the chicken genome.

Author's Declaration

I declare that this thesis was composed by myself and that the work contained therein (including publications) is my own, except where explicitly stated throughout the thesis and within this declaration. This work has not been submitted for any other degree or professional qualification.

Acknowledgements

I would like to thank Dave Burt, Jacqueline Smith, and Alan Archibald. Dave Burt provided me with the opportunity and encouragement to pursue this degree. Jacqueline Smith helped me continue my programme through challenges of contract renewals, a pandemic, and new ventures. Alan Archibald stepped in to help guide me through the final hours of studentship. My PhD programme was certainly not a typical one and the extra effort that Dave, Jacqueline, and Alan put in to help me along is the primary reason I was able to submit this thesis.

I would like to acknowledge the contribution of the late Pete Kaiser, who was partially responsible for bringing me to the Roslin Institute and giving me a great opportunity with the National Avian Research Facility.

I would like to thank Mike McGrew, Lonneke Vervelde, and Zhiguang Wu for help accessing chicken samples and for showing me how amazing avian research could be.

I would like to thank Bruce Whitelaw, who supported my trio of activities as a student, staff, and entrepreneur.

I would like to thank the various members of my group including Kasia, Bob, Matt, Andrew, Lel, Le, and CK. Kasia and Bob were vital for all the wet lab work. Matt, Andrew, Lel, Le, and CK were helpful ears for discussing bioinformatic ideas.

I would like to thank Laura Glendinning who sent me a copy of her thesis so I could understand how to put together a fairly unorthodox composition.

I would like to thank others at The Roslin Institute including Amanda, Christelle, Deepali, Julia, Mazdak, Omar, and Kay who made up my social circle that helped me feel at home at the institute.

I would like to thank Liz Tseng, Jason Underwood, Jonas Korlach, and Gerrit Kuhn from Pacific Biosciences whose interest in my work and invitations to speak gave me much needed confidence to continue forward.

I would like to thank Rachel Hawken, Ron Okimoto, Jun Chen, and John Henshall from Cobb Vantress who made my visit to Cobb headquarters a wonderful experience and allowed me to see first-hand the applications to which hopefully my research will one day enable.

I would like to thank my family for supporting me through my studies and encouraging me to finally complete it.

And lastly, I would like to thank my partner Carrie who is the reason I came to Scotland and who supported me through all those long hours at my computer writing code or emails or papers or this thesis.

Publications

Mueller RC, Ellström P, Howe K, Uliano-Silva M, **Kuo RI**, Miedzinska K, Warr A, Fedrigo O, Haase B, Mountcastle J, Chow W, Torrance J, Wood JMD, Järhult JD, Naguib MM, Olsen B, Jarvis ED, Smith J, Eöry L, Kraus RHS. A high-quality genome and comparison of short- versus long-read transcriptome of the palaeartic duck *Aythya fuligula* (tufted duck). *Gigascience*. 2021 Dec 20;10(12):giab081. doi: 10.1093/gigascience/giab081.

Mueller RC, Mallig N, Smith J, Eöry L, **Kuo RI**, Kraus RHS. Avian Immunome DB: an example of a user-friendly interface for extracting genetic information. *BMC Bioinformatics*. 2020 Nov 12;21(1):502. doi: 10.1186/s12859-020-03764-3.

Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, et al. Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics*. 2020;21(1):1–22.

Yin Z, Zhang F, Smith J, **Kuo R**, Hou ZC. Full-length transcriptome sequencing from multiple tissues of duck, *Anas platyrhynchos*. *Sci Data*. 2019 Nov 21;6(1):275. doi: 10.1038/s41597-019-0293-1.

Chakraborty P, **Kuo RI**, Vervelde L, Dutia BM, Kaiser P, Smith J. Macrophages from Susceptible and Resistant Chicken Lines have Different Transcriptomes following Marek's Disease Virus Infection. *Genes*; 2019; 10, 74. Available from: <https://www.mdpi.com/2073-4425/10/2/74>

Borowska D, **Kuo R**, Bailey RA, Watson KA, Kaiser P, Vervelde L, Stevens MP. Highly multiplexed quantitative PCR-based platform for evaluation of chicken immune responses. *PLoS One*. 2019 Dec 3;14(12):e0225658. doi: 10.1371/journal.pone.0225658.

Elisabetta Giuffra, Christopher K. Tuggle, and the FAANG Consortium. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. *Annu. Rev. Anim. Biosci.* 2019. 7:65–88. Available from: <https://www.annualreviews.org/doi/full/10.1146/annurev-animal-020518-114913>

He C, Hu X, Jung RS, Larsson M, Tu Y, Duarte-Vogel S, Kim P, Sandoval NP, Price TR, Allan CM, Raney B, Jiang H, bensadoun A, Walzen RL, **Kuo RI**, Beigneux AP, Fong LG, Young SG. Lipoprotein lipase reaches the capillary lumen in chickens despite an apparent absence of GPIHBP1. *JCI Insight.* 2017;2(20):1–17.

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 Genomics.* 2017;18(1):1–19.

Psifidi A, Fife M, Howell J, Matika O, van Diemen PM, **Kuo R**, Smith J, Hocking PM, Salmon N, Jones MA, Hume DA, Banos G, Steven MP, Kaiser P. The genomic architecture of resistance to *Campylobacter jejuni* intestinal colonisation in chickens. *BMC Genomics;* 2016;17:293.

Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, Markovic C, Bouk N, Pruitt KD, Thibaud-Nissen F, Schneider V, Mansour TA, Brown CT, Zimin A, Hawken R, Abrahamsen M, Pyrkosz AB, Morisson M, Fillon V, Vignal A, Chow W, Howe K, Fulton JE, Miller MM, Lovell P, Mello CV, Wirthlin M, Mason AS, **Kuo R**, Burt DW, Dodgson JB, Cheng HH. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. *G3 (Bethesda).* 2017;7(1):109-117. [doi:10.1534/g3.116.035923](https://doi.org/10.1534/g3.116.035923)

Pengelly RJ, Gheyas AA, **Kuo R**, Mossotto E, Seaby EG, Burt DW, Ennis S, Collins A. Commercial chicken breeds exhibit highly divergent patterns of linkage disequilibrium. *Heredity (Edinb).* 2016;117(5):375–382. [doi:10.1038/hdy.2016.47](https://doi.org/10.1038/hdy.2016.47)

Wragg D, Mason AS, Yu L, **Kuo R**, Lawal RA, Desta TT, Mwacharo JM, Cho C-Y, Kemp S, Burt DW, Hanotte O. Genome-wide analysis reveals the extent of EAV-HP integration in domestic chicken. BMC Genomics 2015;16:784. doi:10.1186/s12864-015-1954-x

Cheng Y, Prickett MD, Gutowska W, **Kuo R**, Belov K, Burt DW. Evolution of the avian β -defensin and cathelicidin genes. BMC Evolutionary Biology; 2015;15:188. doi:10.1186/s12862-015-0465-3

Boschiero C, Gheyas AA, Ralph HK, Eory L, Paton B, **Kuo R**, Fulton J, Preisinger R, Kaiser P, Burt DW. Detection and characterization of small insertion and deletion genetic variants in modern layer chicken genomes. BMC Genomics; 2015;16:1–17. doi:10.1186/s12864-015-1711-1

Gheyas AA, Boschiero C, Eory L, Ralph H, Kuo R, Woolliams JA, Burt DW. Functional classification of 15 million SNPs detected from diverse chicken populations. DNA Res. 2015;22:205–17. doi:10.1093/dnares/dsv005

Contents

Table of Contents

Abstract	2
Lay Summary.....	4
Author's Declaration.....	5
Acknowledgements	6
Publications.....	8
Contents.....	11
Abbreviations	13
Chapter 1: Introduction.....	15
1.1 Chicken as a model for long non-coding RNA discovery	15
1.1.1 Chicken as a model organism	15
1.1.2 Transcriptome annotation as a concept	18
1.1.3 The application of transcriptome annotation	19
1.2 Long non-coding RNA	22
1.2.1 The transcriptome and long non-coding RNA	22
1.2.2 Long non-coding RNA background	24
1.2.3 Long non-coding RNA in chickens	26
1.3 Long read transcript sequencing	27
1.3.1 Background on RNA detection methods	27
1.3.2 Long read RNA sequencing.....	29
1.3.3 Normalization and 5'-cap selection.....	35
1.4 Bioinformatics.....	38
1.4.1 RNA sequencing bioinformatic pipelines	38
1.4.2 Long non-coding RNA detection methods	41
1.5 Objectives	43
Chapter 2: Paper - Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human.....	44
2.1 Introduction and aims	44
2.2 Research Paper	44
2.3 Discussion	61

Chapter 3: Paper - Illuminating the dark side of the human transcriptome with long read transcript sequencing	64
3.1 Introduction and aims	64
3.2 Research Paper	64
3.3 Discussion	85
Chapter 4: Identifying lncRNA in the chicken transcriptome	88
4.1 Introduction and aims	88
4.2 Methods	89
4.2.1 Creating cDNA libraries for Iso-Seq	89
4.2.2 Processing Iso-Seq data	90
4.2.3 Short read RNA sequencing	90
4.2.4 Short read RNA sequencing data processing	90
4.2.5 Coding predictions and lncRNA classification	91
4.2.6 Comparing Iso-Seq annotation to Ensembl	91
4.3 Results and Discussion	91
4.3.1 Samples, sequencing, and raw data processing	91
4.3.2 Identifying Gene and Transcript Models in Gal5	94
4.3.3 Coding RNA analyses	95
4.3.4 Non-coding RNA pipeline	96
4.3.5 Comparison of all genes to the Ensembl annotation	96
4.3.6 Single exon long non-coding RNA	101
4.3.7 Length of lncRNA	103
4.3.8 Additional evidence for lncRNA	105
4.3.9 Gene expression patterns for lncRNA	115
4.4 Conclusion	122
Chapter 5: Discussion	126
5.1 Improving long read sequencing for lncRNA discovery	126
5.2 Improving analysis of long read RNA sequencing	129
5.3 Future work for chicken lncRNA annotation	133
5.4 General conclusions	138
Chapter 6: Bibliography	139

Abbreviations

Bp : Base Pairs

Cap Norm Brain : 5' cap selected and cDNA normalized brain library

CCS : Circular Consensus Sequence

cDNA: Complementary DNA

CDS : Coding DNA Sequence

ddNTPs : Dideoxynucleotide triphosphates

DNA : Deoxyribonucleic acid

ENA : European Nucleotide Archive

EST : Expressed Sequence Tags

FAANG : Functional Annotation of Animal Genomes consortium

FLNC : Full Length Non-Chimeric

lincRNA : Long Intergenic Non-coding RNA

lncRNA : Long Non-Coding RNA

miRNA : Micro RNA

NCBI : National Center for Biotechnology Information

ncRNA : Non-Coding RNA

NFL : Non-Full Length

NGS : Next Generation Sequencing

NMD : Nonsense Mediated Decay

No Norm Brain : Non-normalized 5' cap selected brain library

ONT :Oxford Nanopore Technologies

PacBio : Pacific Biosciences

RNA : Ribonucleic acid

RNAseq : RNA Sequencing

ROI : Read of Insert

rRNA : Ribosomal RNA

siRNA : Small Interfering RNA

snRNA : Small Nuclear

TAMA : Transcriptome Annotation by Modular Algorithms software

TPM : Transcripts per Million

tRNA : Transfer RNA

VGP : Vertebrate Genome Project

Chapter 1: Introduction

1.1 Chicken as a model for long non-coding RNA discovery

1.1.1 Chicken as a model organism

Since the beginning of modern biology, scientists have observed that there is a connection between the many different species in the natural world. This understanding was first formalized in the use of taxonomy. Taxonomy is a system of organizing organisms based on their traits into groupings that signify the closeness between each species. The origin of taxonomy in the western world is believed to be ancient Greece [1]. However, the modern form of taxonomy was pioneered by Carl Linnaeus in the 18th century [2, 3]. So even before the theory of evolution was proposed by Charles Darwin [4], there was a deep understanding that there is a special relationship between life forms. As taxonomy developed in modern biology, it started to serve as a way for scientists to segment the living world for the purpose of understanding the limits of generalization with respect to biological mechanisms [5].

This understanding that fundamental biological processes are shared across species led to the use of specific species to conduct experimentation [6], [7]. In the early 20th century scientists were focusing on experimentation using a smaller number of species as opposed to the more broadly based observational studies that were popular previously. In the mid-20th century, scientists chose particular species to study with the idea that discoveries in these species could be generalized to species that were in the same or similar taxonomic groupings. This understanding was essentially the birth of the model organism as a concept [8]. This was a significant step in biology because it opened up much easier paths for scientists to design experiments for investigating basic biological functions. For instance, if we wanted to understand how cells replicate in humans, we could study how they replicate in mice because the basic cell biology will be the same between these two species.

As the use of model organisms became more popular, scientists began to scour the taxonomic charts to find ideal organism for studying their biological field of interest. These species would usually be selected for a combination of features such as ease of handling, ease of breeding, life spans, accessibility of tissues or developmental states, and similarity to another organism with commercial or medical value [7].

The first animal model organisms were *Mus musculus* (mouse) and *Drosophila melanogaster* (fruit fly)[6]. The mouse served as a convenient model to study mammalian biology due to its ease of rearing, ease of handling, short life span, and taxonomic relatedness to humans [9]. The fruit fly was chosen for the same reasons except instead of similarity to humans they were valued for their relation to insects [10]. These first two animal model organisms ,and selection of such, demonstrate an important aspect of the selection process which is that as a whole, model organisms are more useful when representatives are chosen to cover a large range of the taxonomic tree [11]. Thus, it was only natural for new animal model organisms to be selected so that they were spaced out across the entire animal kingdom. Species such as *Caenorhabditis elegans* (nematode worms)[12], [13] and *Danio rerio* (zebrafish)[14] were soon to follow as major model organisms [6].

The modern chicken has been domesticated over thousands of years with the closest wild ancestor being the red jungle fowl [15]. Given its long history in captivity, the chicken has also been the focus of various biological observations. For example, Aristotle described chicken embryos in his works and Darwin wrote descriptions of various chicken breeds [16]. The ease of breeding chickens, accessibility of their embryos, and differentiated breeds made chickens ideal model organisms to represent the avian species.

As phylogeny began to replace taxonomy and genomics became more sophisticated with the advent of DNA sequencing, the objective to map out the tree of life became a unified goal in the scientific community [5], [17]. Given the resources needed to sequence and assemble a

genome, researchers began stratifying the search by selecting the best candidates for their initial searches. The chicken became a vital species to initiate the genomic exploration of birds [18][19][20]. In 2004, the chicken genome became one of the first animal genomes to be sequenced and assembled. The chicken genome had several characteristics that also made it ideal for scientific discovery such as the genome being a third the size of the human genome and having long conserved syntenic blocks when compared to mammals.

The sequencing of the chicken led to major discoveries with respect to the evolution of vertebrates [21][22][23]. Large scale efforts to explore the chicken transcriptome took place around the same time using expressed sequence tags (EST) [24],[25]. EST sequencing is a method where RNA is reverse transcribed into cDNA, the cDNA are then incorporated into vectors for cloning, and libraries of hundreds and thousands of cloned cDNA inserts are subjected to single pass Sanger sequencing [26], [27]. At the time, this process was practically the only method of identifying the actual RNA sequences transcribed in living cells. Even then, there was debate as to how valuable EST sequencing truly was with the opposing faction claiming that RNA sequences could be reliably identified through *ab initio* predictions from the genome assembly[28]. However, EST sequencing became a prominent methodology for identifying transcript models for the use of annotating genomes [29]-[30]. While it was possible to attain full-length transcript information from EST sequencing, the process had low throughput and was expensive. Thus, the majority of EST information gathered were comprised of short fragments of transcript sequences. This was certainly true for the beginning of chicken transcriptome annotation [24], [25],[31], [32]. Given the difficulty in attaining this type of information, it was impressive that the chicken as a model had so much time and effort invested into the annotation of its genome. This attention highlights the shared understanding of the importance of chicken as a model and the need for further research into chicken and avian genomics as a whole.

1.1.2 Transcriptome annotation as a concept

Transcriptome annotation is the pursuit of identifying and understanding transcripts.

Transcripts are the RNA molecules expressed by cells. These RNA molecules are transcribed from regions of a genome. In prokaryotes, the production of RNA transcripts is comparatively simple with a single region of the genome typically producing a single RNA molecule. However, in eukaryotes the use of alternative transcription start and ends sites coupled with alternative splicing means that a single genomic locus can produce a myriad of RNA molecules based on the combination of the sequences present within that genomic locus. Alternative transcription start and end sites occur when the RNA polymerase has multiple locations it can start and end within a genomic locus [33]. Splicing is when internal sections of an RNA molecule are cut out with the flanking regions rejoined or spliced together. The sections of the transcript that are retained are labeled as exons on the genome and the sections that are spliced out are labeled introns in the context of the gene loci. Alternative splicing occurs when there are different combinations of splicing producing multiple different transcripts from the same locus. Each individual genomic loci from which transcripts can be produced is called a gene. Thus in eukaryotes a single gene can produce many transcripts. This feature has been proposed as one of the important features of eukaryotes that have allowed for the development of more complex organisms [34],[35].

Given that transcript sequences are RNA copies of regions of a genome, the most typical method of structuring this information is by representing genes as loci within a genome assembly. This is in accordance with our understanding of the relationship between the genome and the transcriptome. In this view, the genome can be thought of as a giant repository of nucleotide sequences that are templates for transcription into RNA to enact function. The unit of the transcriptome is the transcript, however the term gene is often used to denote a grouping of transcripts.

The standard method of ordering transcriptome information is by identifying the locations of a gene via coordinates based on a genome, and then identifying the unique transcripts that are produced from each gene locus. However, this definition can break down when we try to define what exactly a gene is [36]. The classical meaning of a gene was defined before we had significant knowledge of the genome or indeed of RNA. Thus for the purposes of this study, I will admit that the definition of a gene is vague and not entirely biologically accurate.

Essentially the term “gene” is somewhat abstract and can be interpreted in different ways. In contrast, the term transcript is much more biologically accurate in that we have evidence of specific RNA sequences that represent specific transcripts. The term gene in this case is simply a higher order class meant to help us group transcripts into similar functions. However, even in this definition we can find outliers. For example, within the bounds of a single gene other genes can exist within the intronic regions [37]. A single locus can produce transcripts with very different functions even if they share overlapping exon sequences. It is also important to note that we are typically referring to a single strand and not overlapping expressed regions on opposite DNA strands.

In the context of this study, the primary goal of transcriptome annotation was to define all expressed transcripts and assign as much function as currently possible using *in silico* analyses.

1.1.3 The application of transcriptome annotation

There are two ways that researchers identify transcripts: directly or indirectly. In the direct approach, researchers use tools and methods that are designed to identify the sequences of individual RNA molecules. For example, RNA sequencing is a direct method for identifying RNA. With the indirect approach, researchers compare non-RNA features to make predictions on what RNA might look like in a given transcriptome. For example, using comparative genomics to identify similar regions between two or more genomes and then projecting the transcript models from a well annotated genome to a lesser annotated genome.

With respect to the direct approach of annotation, these methods are all based on technologies that allow for the identification of at least a portion of the sequences from RNA that are expressed in the species of interest. While these technologies will be covered in more detail later on in this chapter, in general it can be said that earlier technologies had lower throughput for providing sequence information from RNA. These older technologies were also less capable of providing information on full length transcripts [38]. Given these constraints, the process of going from sequence data collection to transcriptome annotation was focused on assembling partial sequence information from RNA into the best estimate of the full-length transcripts. This idea was very similar to how genome assemblies were performed however, the transcriptome has intrinsic complexity that makes it more difficult to reconstruct from short pieces of information. While the genome can be thought of as practically linear with respect to the sequence composition, the eukaryotic transcriptome is comprised of a combinatorial array of different RNA sequences where a single genomic locus can produce a multitude of unique transcripts.

The innate complexity of the transcriptome means that its exploration requires far more sequencing than is required for genome assembly. This is due in part to differential splicing but it is also due to the ephemeral nature of RNA expression where different cells at different times produce different RNA. This meant that the application of sequencing for transcriptome annotation was cost prohibitive for many research groups [39].

Given the high cost of RNA sequencing methods pre-2010, many researchers opted to use the information generated from the closest model organism to interpret the genomes of the newly studied species [40]. For example, if a researcher were looking into annotating the recently assembled genome for a bat species, they would compare their genome to that of the mouse, rat, or human to find genes by inference from sequence alignments. In fact there are many widely used bioinformatic software which rely on this type of strategy (Augustus, Maker2, etc.) and this strategy is used in part by the major public annotators NCBI [41] and Ensembl [42]. One

of the core assumptions behind the use of this method of annotating genomes is that we can use sequence indicators to gauge the appropriateness of adapting a gene model from one species' genome to another. The sequence indicators include looking at the percent alignment of the gene from the model organism to the genome of the species of interest as well as cues that are intrinsic to protein coding genes [40]. Since protein coding genes have been found to have in general fairly well conserved coding sequence (CDS) regions, the assumption is that the CDS regions of the alignment should have few mismatches and that the protein sequence should also be somewhat maintained. Thus, splice junctions can be predicted based on the assumption that the junctions should preserve the reading frame that would produce a protein sequence most similar to the model organism.

This comparative genomics approach to genome annotation was very attractive due to the method reducing the need for relatively expensive exhaustive RNA sequencing. However, in many cases there were no ground truths to which to compare the results. Thus, all these *in silico* predictions of gene model for newly studied organisms were built on a large number of assumptions which were not investigated further to ascertain the truth[43].

The danger of this approach really emerged as the snowball effect in annotation philosophy began building a momentum of thought that was seldom questioned as new researchers entered the fray. The main issue being that much of annotation had become a self-satisfying practice with little in the way of reality checks to verify the validity of published annotations. The result of which I believe has yet to surface completely but could potentially involve the misunderstanding of results from downstream studies that rely heavily on those annotations. This mentality has also heavily biased the opinions in the field that novel genes, which do not appear to be similar to what is seen in the closest annotated organism, are noise. Given the strong knock-on effect of the traditional comparative genomics method of annotation, we can see that the closest annotated organism mostly has an annotation which reflects some chain of annotations going all the way back to the closest model organism. Thus through this method we have systematically looked away from species-specific gene models as we go along.

1.2 Long non-coding RNA

1.2.1 The transcriptome and long non-coding RNA

The term transcriptome has been defined with various bounds over the course of its introduction. Given that the first RNA that were discovered were messenger RNA (mRNA) that encoded peptide sequences, the initial use of transcriptome referred to all mRNA that were expressed. However, as we discovered different classes of RNA the term has evolved to encompass the entire range of RNA that exist. This includes classes such as transfer RNA (tRNA)[44], ribosomal RNA (rRNA)[45], micro RNA (miRNA)[46], small interfering RNA (siRNA)[47], small nuclear RNA (snRNA)[48], and long non-coding RNA (lncRNA)[49] among others. In the context of this study, I will define the transcriptome as all naturally occurring RNA.

In general, we can breakdown RNA into two major classes: coding RNA (mRNA) and non-coding RNA (ncRNA). Coding RNA (otherwise known as messenger RNA) are simply any RNA that are translated in cells to produce peptide sequences. However, even this definition can be somewhat vague. For instance, nonsense mediated decay (NMD) RNA are translated into peptide sequences when they first encounter a ribosome but then are quickly degraded (along with the amino acid product) so that they cannot produce any more proteins [50][51]. So while NMD RNA may be technically a subset of mRNA because they do get translated, they are effectively ncRNA because they never produce any functional proteins [52][53].

On the other side we have non-coding RNA, which essentially encapsulates all other RNA. Within the ncRNA class, the sub-classes can be somewhat confusing and do not allow for a tree-like organizational structure. The naming of sub-classes is mostly a consequence of the order of discovery of different types of ncRNA and their ascribed functional roles. For example, ribosomal RNA function within ribosomes and transfer RNA are part of the translation mechanism. Then there are sub-classes that use size as the major trait such as small interfering

RNA (siRNA) and micro RNA (miRNA). However, overlaps between functions, size, and other traits have led to a chaotic naming system. To illustrate, ribosomal RNA are almost always referred to as such even though technically they are long non-coding RNA because of their size. In fact, it could be argued that rRNA were the first lncRNA to be discovered[54]. This issue with classification is a testament to the complexity of the transcriptome and highlights the need for care and nuanced descriptions to capture the biological significance of each molecule.

In this study, I define the class hierarchy as being comprised of two main categories: messenger RNA and non-coding RNA (ncRNA). I define NMD's as being part of the non-coding class since they do not produce viable proteins. Likewise, any other RNA that do not encode functional proteins will be under the non-coding class.

Thus, ribosomal RNA, transfer RNA, and all other RNA classes that do not fall under the mRNA class are ncRNA. Prior to the discovery of any non-coding RNA, scientists believed that only mRNA were functional and significant. However, as we uncovered the world of non-coding RNA we realized how important certain types can be for basic biological functions. Explorations in non-coding RNA revealed that RNA has the ability to act as a genome (as in viruses), as protein instructions (mRNA), and function similarly to proteins by forming structural units (rRNA). Thus, it is important to remember that the class of non-coding RNA, although termed due to a lack of a certain type of function, can have other functions some of which we are yet to understand.

As we dig deeper into understanding non-coding RNA, we have picked out sub-classes that we were able to confirm as having specific functions (tRNA, rRNA, etc.). However, we still have a class that exists only as a catch-all for non-coding RNA of which we have little understanding. This is the class of long non-coding RNA (lncRNA). Simply put, lncRNA are defined as any non-coding RNA greater than 200 base pairs (bp) in length. However, this method of classification is generally reserved as a term for RNA that do not exist in other sub-classes such as the 23S ribosomal RNA[55] (in prokaryotes) which is around 2,900 bp in length. Thus, the class of

lncRNA essentially represents the largest frontier of under-annotated RNA within the entire transcriptome.

1.2.2 Long non-coding RNA background

The first lncRNA to be discovered (and the most well-known) which fits the current definition of lncRNA's is XIST. The XIST gene was discovered by Carolyn J. Brown, Andrea Ballabio, and Huntington F. Willard in 1991 [56][57]. The XIST gene is integral to the inactivation of the additional X chromosome in the females of mammalian species including human and mouse. The RNA produced from this gene in humans contain multiple splice junctions, are polyadenylated and are about 17 kb in length [58][59]. The fact that XIST is processed in much the same manner as mRNA, opened up the field of lncRNA discovery as those features were previously thought to be exclusive to mRNA. It also highlighted some of the major challenges for identifying lncRNA, namely that they could be misrepresented as mRNA if not analyzed properly.

Since the discovery of XIST a small number of genes have been confirmed as functional lncRNA in various species but overall the search has been slow and prone to controversy. One other notable lncRNA is the HOX Transcript Antisense Intergenic RNA or HOTAIR gene discovered in 2007[60]. Since its discovery, studies have shown the HOTAIR lncRNA gene to be involved in the development of various types of cancer [61][62][63]. While the exact mechanism by which HOTAIR is involved in tumorigenesis is still unclear, it demonstrates how crucially functional lncRNA can be.

The main reasons for the difficulty of identifying functional lncRNA is somewhat philosophical in nature. It is easier to prove that a specific RNA is translated into a protein than to prove that it is not. Also unlike mRNA that have well known associated features, lncRNA have not been characterized to an extent where we can apply useful general rules for their identification. In

addition, even if we can confirm that a specific RNA is non-coding and that we have the correct transcript model, it requires in-depth biochemical testing to identify a function.

To make this search even more difficult, researchers have identified two general characteristics of lncRNA which are a lack of sequence conservation across species [64][65] and generally low expression levels [66]. This makes them difficult to find using comparative genomic approaches and with standard sequencing methods.

There may also exist a type of blind spot in the field that is the result of the vast amount of work done on annotated model organisms like mouse, human, fruit fly, etc. This has to do with the only method of sub-classifying lncRNA which is based on the location with respect to the nearest protein-coding gene and the bias against non-spliced transcript models [67].

The location-based classes are intergenic, intronic, sense exonic, and antisense [68]. Intergenic lncRNA (lincRNA) are lncRNA that occur in genomic regions with no coding genes. LincRNA are the most widely studied as they are separated from genes making it easier to distinguish their transcripts from transcriptional noise associated with other genes of whatever class. Intronic lncRNA occur within the intronic region of a protein-coding gene. Sense exonic lncRNA have exons that overlap on the same strand with a protein-coding gene. Antisense lncRNA occur on the opposite strand of a protein-coding gene. There are also two mutually exclusive splicing classes termed mono-exonic (single exon) and multi-exonic (2 or more exons). The positional classes and the splicing classes can be used in conjunction for finer categorization.

Of these classes, the most difficult to find and least studied are the sense-exonic mono-exonic lncRNA. This is due to the difficulty in distinguishing them from the protein-coding genes that they overlap. This is where the blind spot can occur, since the model organisms tend to have well annotated genomes with almost complete listings of the protein-coding genes. Thus, all regions that are spanned by protein-coding genes are essentially dark zones where the search

for lncRNA is limited. Actually this extends to any of the positional classes for mono-exonic lncRNA which overlap with a protein genic region either on the same strand or opposite strand.

Thus, while well-studied model organisms may at first seem to be ideal for the extension of annotations into the realm of lncRNA, they are actually in some ways encumbered by their very strengths.

1.2.3 Long non-coding RNA in chickens

While there have been numerous efforts to identify and annotate lncRNA in chicken, these studies relied heavily on short read RNA sequencing [69][70][71] and in some cases comparison with mammal lncRNA [72]. The issue with using mammal models to compare with the chicken genome for the identification of lncRNA is that lncRNA sequences are poorly conserved even across species in the same family[64]. Thus, mammalian genomes are simply too disparate to avian species with respect to lncRNA. The logic behind this methodology follows a common assumption in comparative genomics that functional sequences including genes should be conserved across species. While this is certainly true for a set of protein-coding genes that are required for basic cellular functions, it seems intuitive that it cannot be true for all genes given that at least some set of genes must be responsible for the differentiation of phenotypes that we see across species. Thus relying heavily on conservation for the identification of lncRNA in chickens results in a very small number of lncRNA that are likely to have little or no role in species differentiation (i.e. no chicken-specific lncRNA).

The use of short read RNA sequencing for chicken lncRNA discovery shares many of the issues that plagued short read based identification in mammalian species. Certain classes of lncRNA were more identifiable using short read technologies such as intergenic lncRNA and multi-exonic lncRNA. However, this left a number of other classes in the dark. The inherent noisiness of short read data coupled with the difficulty in distinguishing real transcript models for lowly expressed genes meant that the compromise between sensitivity and specificity made it

virtually impossible to have good performance on both ends. Basically, it is a choice of including a large number of non-real transcript models in the predictions or discarding a large number of potentially real transcript models.

Another issue with both the conservation and short read methods is that they lack the ability to characterize full-length lncRNA models. Thus, even if they were able to identify some part of the transcript model they were not very strong at producing the full length of the model. This was already true for protein-coding genes but is exacerbated in lncRNA where open reading frames cannot be used to judge completeness.

The result of these efforts are lncRNA predictions that are usually far too conservative in terms of numbers of lncRNA as well as not accurate with respect to the full length of lncRNA transcript models.

1.3 Long read transcript sequencing

1.3.1 Background on RNA detection methods

The sequencing of RNA and DNA has changed massively since its inception. While DNA sequencing is more well known today and so ubiquitous that RNA sequencing is often mislabeled as DNA sequencing in popular news articles, it was actually RNA which was first used in nucleotide sequencing development [73]. The reason for this is that RNA offered the easiest material to attain in large and relatively pure quantities at the time and its single stranded nature made it seem more accessible. The first complete nucleic acid sequence identified was that of alanine tRNA from *Saccharomyces cerevisiae* in 1965 by Robert Holley and his collaborators [74][73]. This was done using a ribonuclease-based method. Around the same time Fred Sanger was also sequencing RNA using radiolabeled partial-digestion fragments (not to be confused with what is commonly called Sanger sequencing today)[75].

In the 1970's, researchers began developing DNA sequencing methods with the most famous method (Sanger sequencing) being developed in 1977 which is also known as a chain termination method [76]. This method revolutionized DNA sequencing by making it much easier to attain long sequence information. It was also used to indirectly sequence RNA through cDNA copies of RNA species (commonly referred to as cDNA libraries). This is probably the start of the common practice of developing DNA sequencing techniques which are then adapted for RNA sequencing via cDNA libraries. Sanger sequencing or dideoxy chain-termination sequencing is based on the use of radiolabeled dideoxynucleotides (ddNTPs) and polyacrylamide gel. The DNA would be placed in the gel in 4 different lanes for electrophoresis after undergoing a DNA extension reaction with the incorporation of radiolabeled ddNTPs and dNTPs. Each lane would be assigned a different nucleotide type to be radiolabeled (i.e. A,T,G, or C). Since the radiolabeled ddNTPs halted further extension bands would appear at each length where a termination occurred. Thus, by looking at the different band positions from each of the 4 lanes one could determine the sequence of the original DNA template.

The next big leap began what is commonly referred to as second generation or next generation sequencing. This started with pyrosequencing that was commercialized by 454 Life Sciences. 454's pyrosequencing method involved the immobilization of DNA templates on beads within individual wells. Single dNTPs were then passed over the flowcell for incorporation causing a reaction from the released pyrophosphate products.

However, the technology that rapidly changed the DNA/RNA sequencing world which is currently commonly labelled as Illumina/Solexa sequencing was developed by Shankar Balasubramanian and David Klenerman[77][78]. Solexa sequencing involved the use of a flow cell where DNA is immobilized by oligonucleotides bound to the surface. Bridge amplification is used to create densely packed clusters of each DNA template. During sequencing, fluorescent dNTPs are added in a sequencing by synthesis method. The clusters of identical DNA templates create a stronger fluorescent signal that can be detected consistently which results in very high accuracy of base calling. This method also had the advantage of enabling paired end sequencing

where the other side of the DNA template could also be sequenced in order to get both ends of each DNA fragment. This method was relatively inexpensive with the potential for massive throughput and high accuracy. One major downside however, was the limitation on read length that was about 500 bp. This limitation is caused by the cluster of DNA getting out of synchronization leading to erroneous base calls. Thus, another characteristic of Solexa/Illumina sequencing is that the quality of the reads always deteriorate toward the end leading to a bias toward the start of the read.

Another technology that was commonly used to detect and quantify RNA is the microarray. While this method is useful for cheaply identifying relative gene expression it cannot be used to identify novel transcripts and thus does not have any practical usage for genome annotation where transcriptome discovery is the primary objective.

1.3.2 Long read RNA sequencing

Long read RNA sequencing, also known as third generation sequencing, was developed after what is commonly referred to as Next Generation Sequencing (NGS). Unlike its predecessors the focus of long read sequencing was to produce the longest usable reads possible. The main motivation for this was to solve issues with genome assembly where short reads were found to create problematic scenarios due to repeats and structural variations found in the genomes of many species.

The two leaders in long read sequencing are Pacific Biosciences (PacBio)[79] and Oxford Nanopore Technologies (ONT)[80]. Both technologies are single molecule sequencing platforms that differentiate them from many of the technologies in the second generation sequencing era. PacBio sequencing uses a flow cell with microscopic wells termed Zero Mode Waveguides. These are designed so that when a laser is directed toward a well, the light is guided into a single wavelength that can be distinguished between wells to allow for the basecalling from a single molecule reaction from each well. This method of basecalling is coupled with the rolling

amplification of a circular single-stranded DNA. The circular single-stranded DNA is actually comprised of a typical blunt end double stranded DNA fragment with hairpin structures ligated on both ends. This allows the DNA fragment to be sequenced twice with every full pass through the circular single strand template. Since the rolling amplification can create very long templates (currently an average of 100 kb), the original DNA template can be sequenced many times. This multiple sequencing of a single molecule allows the use of intra-read sequence comparison to improve overall basecalling accuracy (Figure 1.1). Note that the longer the polymerase read (entire length of read generated) and shorter the DNA template (sometimes referred to as Read of Insert (ROI)) the higher the overall accuracy of the read will be.

So while the raw error rate for PacBio basecalling is about 10-15%, the overall error rate after intra-read correction can be lower than 0.0001%. Another way of putting it would be to say that through the use of this method, one can achieve accuracy higher than that of Illumina sequencing.

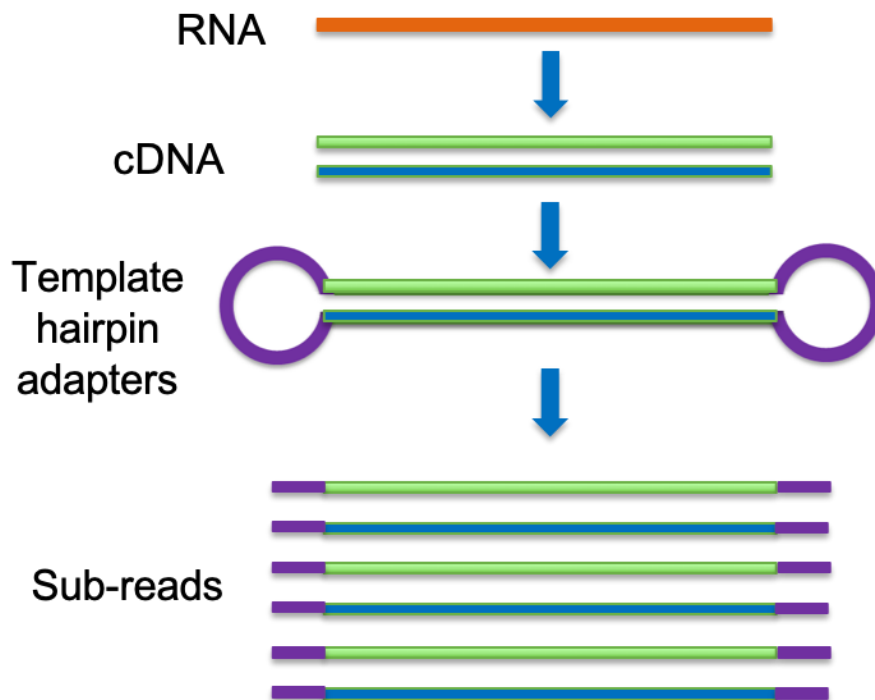
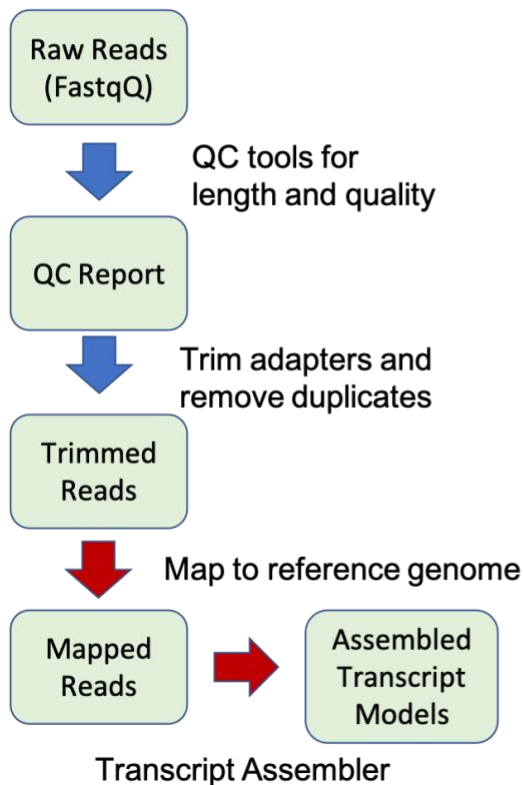


Figure 1.1 : A diagram illustrating how PacBio sequencing works. RNA is converted to cDNA. The cDNA has hairpin adapters ligated to both ends. During sequencing, a polymerase

attaches to the circular template and begins a rolling amplification while situated in the Zero Mode Waveguide well. This creates a very long polymerase read which passes through the template multiple times. Each pass is split up into sub-reads.

Short Read Pipeline (Illumina)



Long Read Pipeline (PacBio)

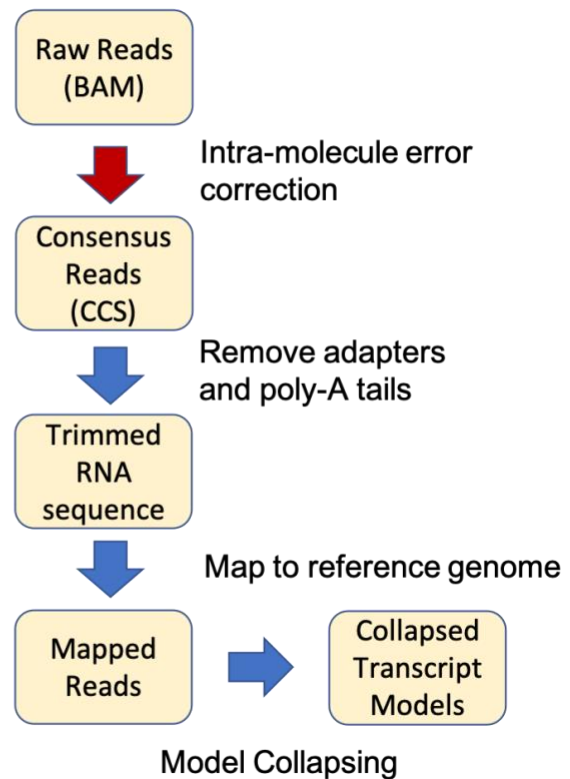


Figure 1.2 : A diagram illustrating the differences between the basic bioinformatic pipelines of short read and long read RNA sequencing. Red arrows denote bottleneck steps due to computational requirements.

Oxford Nanopore Technologies uses a very different approach to produce long reads. As the company name would suggest they use nano sized pores set in a thin film. These nanopores consist of a special protein that sit inside a hole. The proteins change their conformation as different nucleic acids pass through. When there is a voltage potential applied across the film, this change in conformation can be detected as localized changes in current. Thus, ONT use these current changes to call the bases of the nucleotide sequences as they pass through.

This method of sequencing allows for very compact sequencing machines, such as the MinION, and can be used for the direct sequencing of RNA molecules. Note that most third and second generation methods require cDNA as templates. In general, nanopore sequencing only passes through a template once which means that intra-read error correction methods are not used thus lowering the overall accuracy. However, the accuracy for ONT sequencing is dependent on the basecalling from the current readings that means that the current limitation lies more on the bioinformatic side of interpreting the raw signals (also known as squiggles). However through improvements in the bioinformatics tools, the accuracy for ONT has been steadily increasing over the past few years. They have also produced new flow cells with two aligned pores so that they can read twice from the same molecule. Thus, there are a number of potential means of improving the accuracy.

Another important consideration is that (at least at the time of this writing), all ONT sequencing requires the generation of double-stranded nucleotide sequences. Even for direct RNA sequencing, a second cDNA strand must be produced in order to sequence properly. This is because the second strand helps to prevent secondary structures that would change the kinetics of the nucleotide sequences travelling through the pores. However, it may be possible to account for this in some other way in the future. This limitation however, has important effects on RNA sequencing because all ONT RNA and cDNA sequencing still relies on the limits of reverse transcription, i.e. the length of RNA molecule that can be reverse transcribed before the reverse transcriptase falls off the template.

Another consideration with long read RNA sequencing is that typically oligo-dT primers are used to prime RNA with poly-A tails. This is to ensure that the 3' end of the transcripts are captured. However this excludes any RNA without poly-A tails and the use of oligo-dT's can also result in internal priming which leads to 3' truncation of the transcript sequences. There are methods to ligate a poly-A tail to all RNA within a sample however this is rarer and allows for the inclusion

of 3' degraded RNA with no robust way of discriminating between 3' complete and 3' degraded sequences.

Around the start of high throughput long read RNA sequencing, the early adopter researchers had to devise their own strategies for analyzing this novel type of data. While the general upstream processing to prepare reads for mapping to the genome were provided by the sequencing companies. The tools for filtering and interpreting the results after mapping were lacking. Many of the early users (and many current users) defaulted to the tools provided by PacBio which at the time was called ToFU[81]. The primary tool within the ToFU package was called Collapse and its purpose was to remove redundant read information using the read mapping information. The output of ToFU was a collection of transcript models that were grouped into genes based on genomic overlap between the different models. This concept of collapsing was key to developing a deeper understanding of the requirements in downstream analysis of long read RNA sequencing data. However, there were still more aspects of the data that had not yet been fully understood and were not utilized in ToFU Collapse. ToFU Collapse was also designed to be a generalist tool with minimal parameterization control which meant that it sufficed for a wide range of experiments but did not exactly optimize for the specific needs of each. Despite this generalist methodology, many of the first papers to be published based on Iso-Seq data used the ToFU pipeline [82][83][84][85][86].

The room for improvement on this foundational tool was not lost on the community and several groups developed their own tools to solve for the specific challenges they encountered their experimental designs. These include tools like TAPIS [87][88], FLAIR[89], Stringtie2[90], and TALON[91]. These tools have proved useful for their creators as well as other researchers who shared similar ideas as to the objective of their transcriptome annotation. However, they all make use of algorithms that are based on certain assumptions which may not be biologically accurate or be relevant to the possible situations that occur during long read sequencing.

For example, Stringtie2 was very much developed to be a long read successor to Stringtie[92] which is arguably the most popular short read transcriptome assembler. Stringtie2 seems to make the assumption that like short reads, long reads should be scaffolded together to make the longest

contiguous sequence possible and that this conglomerate longer sequence is what should represent a transcript model. This behavior is quite baffling considering the purpose of long reads is to provide sequence information on the full length of a transcript. Thus stitching together reads in the way that is typically done for short reads is counter to the primary benefit of the technology. The typical effect of this behavior is a reduction in the ability to discover novel isoforms and an over lower sensitivity of alternative transcripts. This characteristic of Stringtie2 can be seen in recent studies[93] where there have been comparisons of different processing pipelines.

In general, the tools that have been developed so far have had a greater emphasis on matching the more conventional ideas of transcriptome annotation. While this may be good for creating results that mirror what has been seen before, this philosophy closes off opportunities to dig further into the unknown territory of the transcriptome and especially the non-coding RNA.

Indeed there were many analytical choices that had to be made in order to cope with limitations of short read RNA sequencing. Since there was no way of connecting the start of transcription with the end of transcription at the transcript level, any alternative transcripts that had similar exonic structure to a longer transcript would essentially be lost by assignment of the reads originating from those shorter transcripts into the longer transcript model. This was the case with early work by GENCODE when annotating the human genome[94][95]. This choice in transcriptome assembly meant that there was a bias toward predicting longer transcript models with more exons. Also the amount of noise created from short read data make it difficult to determine when a single exon transcript model is indeed real or rather just the result of some transcriptional or processing noise. These biases created from these limitations of short read data are evident in the GENCODE lncRNA study where they report that the predominant number of exons for lncRNA is 2 and that there is strong expression correlation between anti-sense lncRNA and the protein coding genes they overlap on the other strand. They state the lncRNA seem to have fewer exons in general than protein coding genes and yet instead single exon transcripts are not identified as the predominant class. Thus these choices have shaped our view of non-coding RNA in perhaps an inaccurate way.

1.3.3 Normalization and 5'-cap selection

One of the major drawbacks to using long read RNA sequencing as compared to short read are the cost per read, also referred to as throughput. Especially at the advent of Iso-Seq (the first high throughput long read RNA sequencing) the cost per read was orders of magnitude higher than a comparable Illumina experiment.

Most RNA sequencing projects fall under what is known as bulk RNA sequencing. This is when a sample is comprised of some tissue or other mixture of cell types. Given the heterogeneity of cells in these experiments, large differences in the abundance of unique RNA often occur. These differential abundances can be caused by either a high or low expression across all cells in the sample and/or by the difference between the number of each cell type within the sample. The highly expressed genes can make finding lower expressed genes difficult. This is because when sequencing for discovery we are essentially applying a random sampling of the original RNA library. Thus, more of the reads produced will come from highly expressed genes as compared to low expressed genes. In many cases the highly expressed genes make up more than half of the RNA library.

Combining gene differential expression with the high cost of long read sequencing creates a major challenge for identifying low expressed genes. Given that lncRNA are typically expressed at very low levels this means that even though long read RNA sequencing allows for much more robust identification of full length RNA, it suffers from not being able to sample lncRNA with sufficient efficiency.

Thus in order to address this challenge, a cDNA normalization step can be implemented in preparation of the cDNA library prior to sequencing. The objective of cDNA normalization is to

create a cDNA library that contains all the unique transcript sequences of the original RNA library but with the same relative abundance across all unique transcripts [96]-[97]. This means that there is a uniform level of abundance across all unique genes/transcripts. Making the cDNA library uniformly distributed maximizes the sampling efficiency during sequencing for identifying all transcript sequences present within the original RNA sample. Thus this method could theoretically provide significant improvement in the use of long read RNA sequencing for the identification of lncRNA.

Another issue pertaining to RNA sequencing in general is the propensity of RNA to degrade rapidly[98][99]. Usually this occurs from the 5' end and runs toward the 3' end in a linear fashion. This phenomenon can be easily observed when looking at mapped short read RNA sequencing data where the read coverage across a gene is typically lower at the 5' end as compared to the 3' end [100][101]. In long read RNA sequencing, this phenomenon is manifested as mapped reads that show varying 5' ends but matching 3' ends (Figure 1.2).

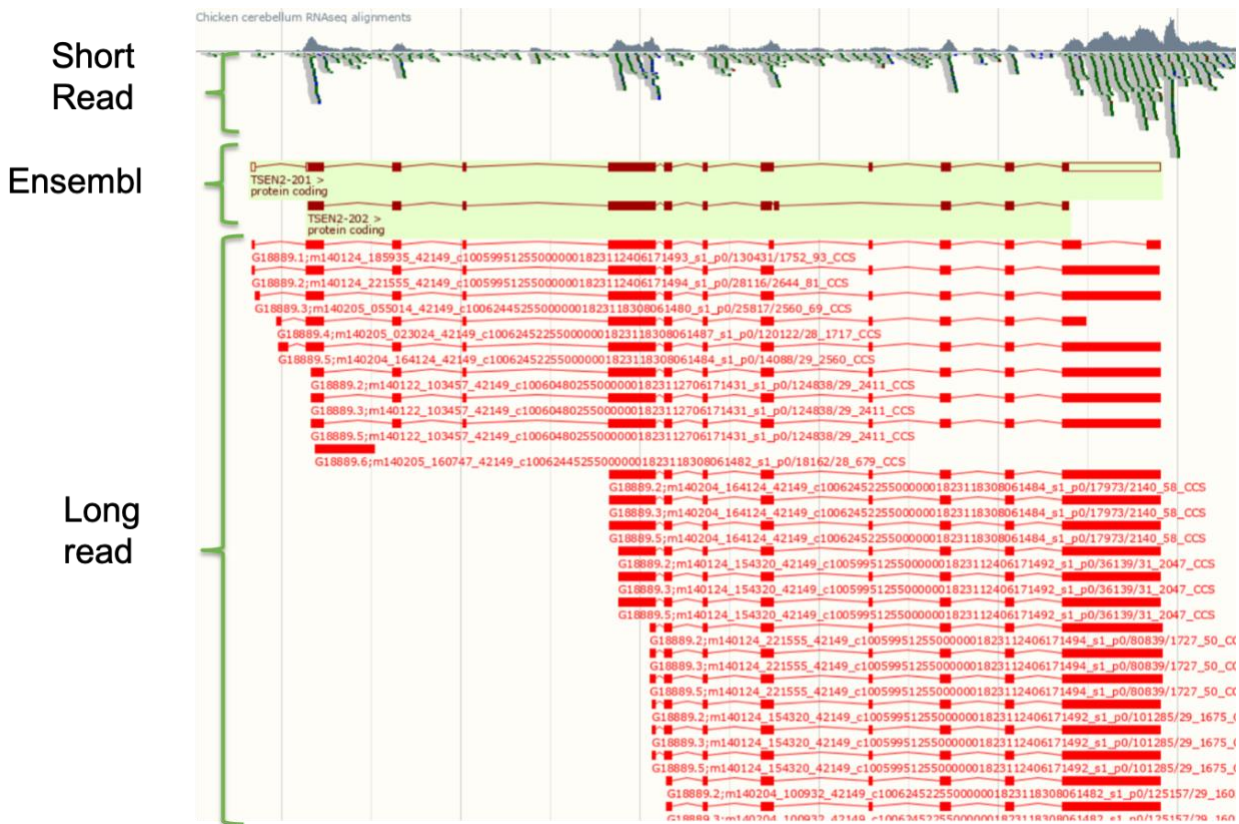


Figure 1.3 : Example of long reads from degraded RNA. The top track shows read coverage from short read RNA sequencing. The middle track shows transcript models from the Ensembl annotation. The bottom track shows Iso-Seq reads for this genomic locus. The reduction in representation on the 5' end is visible in both the short and long read data. The long read data track does not show all reads but rather a representative of each type of read regardless of its abundance.

This phenomenon (known as 5' degradation) is particularly problematic when using long read RNA sequencing for three main reasons. The first is that, due to the much lower throughput of long read sequencing, there is lower read coverage across the transcriptome that reduces the chances of sequencing a full length transcript. In comparison, short read sequencing with higher throughput produces much greater coverage per transcript thus increasing the probability of sequencing at least one full length RNA. This could result in the improper annotation of a truncated transcript model as full length. If the source RNA had been degraded past the start

codon, then this would change the coding prediction for the transcript model effectively creating an erroneous annotation.

The second reason is, depending on the method of bioinformatic analysis, that even though the full-length transcript model was identified, the 5' degraded models could be classified as separate unique transcript models. Thus, the outcome is similar to the first reason but perhaps slightly better because the full-length model is at least attained.

The third reason is that essentially all reads produced from degraded RNA are practically useless since the primary aim of long read RNA sequencing is for the identification of accurate full-length transcript models. Thus, degraded RNA effectively lowers the useful throughput of the already low throughput of long read sequencing.

To address this issue, several methods of 5' cap selection have been developed. In eukaryotes, mature RNA are typically modified to have a methylated G at the 5' end (known as a 5' cap). This modification protects the RNA from rapid degradation once exported to the cytosol. 5' cap selection methods are based on using different strategies for targeting RNA that have the methylated G on the 5' end [102][103][104]. The purpose of this method is to enrich for full length RNA that in turn allows for more efficient usage of long read RNA sequencing capacity or platforms.

1.4 Bioinformatics

1.4.1 RNA sequencing bioinformatic pipelines

The general pipeline for processing sequencing reads to creating transcriptome annotations involves four major steps: read preparation, read mapping, transcriptome assembly, and further annotations. However, these steps can vary greatly between processing short read versus long read sequencing data. There are also methods for producing transcriptome

annotations in the absence of a reference genome assembly[105][106]. Using RNAseq reads to produce transcript models without the use of a reference genome is termed *de novo* transcriptome assembly[107][108]. This method is typically used when sequencing a species without a genome assembly or with a poorly assembled genome. However, when a good genome assembly is available it is almost always preferable to use the primary reference based method[109]. This method is also used to explore the parts of the transcriptome that are expressed from regions of the genome which are missing from the genome assembly[110].

The first step of any sequencing analysis is to understand the experimental conditions. This includes what the objectives of the experiment are and the experimental design. For short read sequencing there are different library preparation methods that can be used including the use of total RNA, rRNA depleted RNA, and poly-A tail selected RNA. The transcript libraries could be prepared to preserve strand information and different fragment sizes can be used. During sequencing, there is a selection of read lengths and between sequencing a single end or paired ends. All these variables are important to consider when processing the data.

The first step after receiving raw data from a sequencing centre is typically to run a quality control check on the reads. This step allows for the identification of the quality drop-off in Illumina reads and the presence of adapter sequences that may need to be trimmed from the reads. The next step is called trimming and involves the removal of the lower quality ends of reads as well as removing adapter sequences[111]. After trimming, the short reads are typically ready to be mapped to a reference genome assembly[112][113]. When mapping RNA sequencing reads to a reference genome a splice-aware aligner must be used. This basically means that the software for aligning the reads to the genome must account for the possibility that any given read may represent a region overlapping a splice junction. In these situations one half of the read may map thousands of base pairs away from the other half. In some cases a read can overlap more than one splice junction making this alignment even more challenging.

After mapping, the alignment file is then processed by a transcript assembly tool[114][112]. This tool will go through each mapped read and piece them together in the attempt to solve the puzzle of what the original RNA sequences looked like[115]. The resulting transcript models are represented as regions projected onto the genome assembly. Essentially this means that the genome assembly is used as a coordinate system for placing the different features of each transcript model.

After transcript assembly, the transcript/gene models can be used for a number of different downstream analyses to provide further annotation [42][116][117]. This can include removing models that are unlikely to be real, predicting coding potential, predicting open reading frames, and matching isoforms to known peptide sequences[118].

For all of these steps in the short read pipeline, there are a wide range of software tools developed with a select few attracting the most usage. The most widely used mapper for short read RNAseq data is Tophat2 [119] which actually uses the Bowtie [113] aligner but adds on splice awareness. The most wide used reference based transcript assembly tools are Cufflinks [120] and its successor Stringtie [114].

For long read RNA sequencing data processing the same general principles apply but there are differences which require special approaches and methods.

During the read preparation phase, adapters and poly-A tails are removed from the reads so that the remaining sequence represents the RNA sequence as it relates to the genome. However, in the case of Iso-Seq, there are some additional steps that occur before the trimming step and some optional steps that occur after. For Iso-Seq, the data provided by the sequencing centre are called subreads. The subreads represent each pass through the template from their circular sequencing. In PacBio's terminology the full read is called a polymerase read and each pass is called a subread. Since each polymerase read typically contains multiple subreads, the next step of processing uses these multiple passes of the same template to do an intra-read

error correction called circular consensus sequence (CCS). During CCS some reads can be filtered out based on the quality scores and number of passes. After CCS, the CS reads are then trimmed for adapters and poly-A tails. However, during this process the reads are checked for artificial concatemers that are discarded. The reads are also oriented 5' to 3' during this step. After trimming, the reads can either be mapped directly to the reference genome or they can be used for another round of error correction using inter-read sequence alignment. This step is known as Cluster/Polish. After mapping, we see a big difference in approach compared to short read methods. Since the reads are ideally supposed to represent full length RNA, the goal is not so much an assembly as it is a collapsing of redundant transcript models and some other data filtering. During the collapsing step some algorithms can be applied to remove transcript models that are likely to be erroneous. This includes removing models arising from transcriptional noise, models derived from degraded RNA, models from truncated cDNA, and models with incorrect splice junction coordinates. Since long read technologies produce reads with higher error rates, there is more noise around identification of the correct splice junctions. Incorrect splice junction predictions can cause changes to the open reading frame that have a significant impact on downstream analyses. Thus even after all these filtering steps, there is still a need to build in awareness of these issues into downstream methods.

1.4.2 Long non-coding RNA detection methods

Prediction of lncRNA ultimately comes down to answering two questions: is the transcript model real and does it produce a protein? With respect to the first question, this strategy is usually to use different pieces of evidence to filter out transcript models that look like they may be the result of some noise in the full sequencing pipeline. In short read pipelines, the prevailing philosophy for removing noise models is to focus on read coverage and bias toward multi-exonic models. The reasoning for this is that noise is expected to have a lower read coverage than real transcript models and splice junction mapping is unlikely to occur for transcriptional noise. This mentality is also the consequence of short read sequencing limitations. Given that short reads have a tendency to multi-map across the genome and most

reads do not contain a splice junction, there is a tendency for short read data to produce a large amount of low coverage mono-exonic models. However, this strategy also means that there is a substantial bias against the identification of lowly expressed mono-exonic lncRNA. Even before the transcript models are assessed for their coding potential, many possibly real lncRNA could be removed from the dataset.

All lncRNA prediction tools are in essence trying to answer the second question of whether or not a transcript model represents an mRNA. Thus, all these methods are simply looking for any evidence that the transcript in question has coding potential. There are four main ways of identifying coding potential: directly matching the sequence of a transcript to a known gene or transcript, matching the transcript sequence to a known peptide sequence, identifying k-mer patterns that occur frequently in coding genes, and identifying open reading frames. At the core of these methods is the idea that if the transcript is protein coding it will have some similarity to known protein coding genes. However, a major issue that can complicate this assumption is the presence of an incorrectly mapped splice junction that in turn could cause a frame shift in the open reading frame. Depending on where this frameshift occurs, a real protein-coding transcript could appear to be non-coding due to the disruption of the open reading frame. Given the relatively low error rate of short reads and their greater read coverage for splice junctions, frame shifts from erroneous splice junction predictions are not as great a concern. The lower throughput of long reads leading to lower coverage per gene and the higher error rate, erroneous frame shifts are far more abundant in the long read based transcript models.

On the other hand, short read transcript models suffer from issues with exon chaining and determining transcript start and end sites. These limitations result in models that often represent a conglomeration of features that belong to different transcripts and as previously mentioned, these issues can cause a significant bias in the types of lncRNA that can be identified.

While both forms of sequencing have their strengths and weaknesses for identifying lncRNA, there are a number of ways that splice junction errors can be dealt with and of course these types of errors do not affect mono-exonic models which are more easily identified from long read data. Thus, long read RNA sequencing data provides unique advantages over short read for the identification of novel lncRNA.

1.5 Objectives

The objectives for this piece of work can be divided into 4 aspects: investigate new technologies for lncRNA discovery, apply this technology to the chicken transcriptome, develop pipeline and software for lncRNA discovery in chicken, identify ways to annotate predicted novel lncRNA genes.

- Investigate new technologies for lncRNA discovery
 - To understand how PacBio Iso-Seq sequencing works both from a biochemical perspective and a data analysis perspective.
- Apply this technology to the chicken transcriptome
 - To understand how PacBio Iso-Seq can be used for the annotation of the chicken transcriptome. This includes what problems can be solved with this new type of sequencing data and what issues arise.
- Develop pipeline and software for lncRNA discovery in chicken
 - To create or establish software and pipelines that allow for the processing from raw Iso-Seq sequencing reads to a transcriptome annotation.
- Identify ways to annotate predicted novel lncRNA genes
 - To identify different methods for using the Iso-Seq based annotations along with other sources of information for the identification of novel lncRNA. This includes finding ways to further classify lncRNA and establish a foundation for future work.

Chapter 2: Paper - Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human

2.1 Introduction and aims

In this chapter I explore the use of long read RNA sequencing along with different cDNA library preparation methods for annotation the chicken transcriptome. I analysed Pacific Biosciences Iso-Seq sequencing data produced from chicken embryo and chicken brain. I looked at the effects of cDNA normalization on transcriptome sampling as well as the use of 5' cap selection for identifying 5' intact transcripts sequences. I also performed some analyses for identifying lncRNA and comparing the predicted lncRNA to the predicted coding genes.

In this study, all work was performed by myself except as further specified: sample collection and preparation were performed by Ian R. Paton, Lel Eory contributed to the short read data processing, and sequencing was performed by GATC Biotech (Konstanz, Germany).

2.2 Research Paper

This research was published as “Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human” in BMC Genomics [121]. BMC Genomics does not require authors to obtain permission to include these papers in their thesis, provided the original work is properly cited.

RESEARCH ARTICLE

Open Access



Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human

Richard I. Kuo¹ , Elizabeth Tseng², Lele Eory¹, Ian R. Paton¹, Alan L. Archibald¹ and David W. Burt^{1,3*}

Abstract

Background: Despite the significance of chicken as a model organism, our understanding of the chicken transcriptome is limited compared to human. This issue is common to all non-human vertebrate annotations due to the difficulty in transcript identification from short read RNAseq data. While previous studies have used single molecule long read sequencing for transcript discovery, they did not perform RNA normalization and 5'-cap selection which may have resulted in lower transcriptome coverage and truncated transcript sequences.

Results: We sequenced normalised chicken brain and embryo RNA libraries with Pacific Bioscience Iso-Seq. 5' cap selection was performed on the embryo library to provide methodological comparison. From these Iso-Seq sequencing projects, we have identified 60 k transcripts and 29 k genes within the chicken transcriptome. Of these, more than 20 k are novel lncRNA transcripts with ~3 k classified as sense exonic overlapping lncRNA, which is a class that is underrepresented in many vertebrate annotations. The relative proportion of alternative transcription events revealed striking similarities between the chicken and human transcriptomes while also providing explanations for previously observed genomic differences.

Conclusions: Our results indicate that the chicken transcriptome is similar in complexity compared to human, and provide insights into other vertebrate biology. Our methodology demonstrates the potential of Iso-Seq sequencing to rapidly expand our knowledge of transcriptomics.

Keywords: Iso-Seq, PacBio, Single molecule long read sequencing, Transcriptome sequencing, RNAseq, Chicken, Avian, Gallus gallus, Genome annotation, Coding RNA, Non-coding RNA

Background

Transcriptome annotation is crucial for a wide array of biological research areas, including genomics, proteomics, epigenetics, immunology, and phylogenomics [1]. The identification of the full repertoire of transcribed elements provides information on the functional roles and relationships of genomic loci which in turn can be compared to understand a vast array of biological mechanisms. However, due to the complexity of transcript splicing and the limitations of previous technologies, researchers had to choose between low-throughput, costly methods to generate accurate full-length transcript

models, such as cDNA cloning [2] or high-throughput, cheaper methods to generate imprecise transcript models, such as short read RNA sequencing [3, 4]. The current status of chicken annotation represents a prime example of this trade off.

The Ensembl chicken annotation (release 83), built primarily on short read RNAseq and comparative data, contains 17,108 genes with 17,954 transcripts [5]. These numbers stand out for two major reasons. The first reason is that the number of genes is far lower than that found for other vertebrate organisms, for example, the Ensembl (release 83) human annotation contains 60,675 genes (including coding and non-coding genes). The difference in the number of genes annotated in the chicken and human genomes is heavily influenced by lack of long non-coding gene predictions in the chicken annotation. While it can be argued that this may

* Correspondence: d.burt@uq.edu.au

¹The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK

³The University of Queensland, St. Lucia, Canberra, QLD 4072, Australia

Full list of author information is available at the end of the article



represent differences between mammals and birds, evidence that many more genes exist in birds can be seen in the cDNA support track on Ensembl. The second reason is that the current chicken annotation is almost entirely comprised of protein coding genes for which a single transcript is described. Again this is contrary to what we know from other vertebrates with the human annotation (Ensembl release 83) containing 199,184 transcripts (i.e. an average of 3.3 transcripts per gene). These discrepancies highlight major limitations to using short read RNA sequencing and comparative data for building gene and transcript models.

With short read RNAseq data there are three major transcription characteristics that are difficult to determine [6]: (i) transcript start sites (TSS) and transcript termination sites (TTS), (ii) exon chaining, and (iii) transcriptional noise. If multiple TSS or TTS exist for a transcribed locus, then interior TSS and TTS can go undetected due to combinations of inconsistent read coverage, overlapping exons, and overlapping splice junctions. Thus for any transcript model produced via short read data, we often cannot determine if there are alternative TSS and TTS which have not been detected. Similarly, the process of chaining exons and splice junctions together to reconstruct full-length transcript sequences can be problematic. Since a single short read cannot usually span all splice junctions within a multiple splice junction transcript, transcript assemblers must predict which exons are linked to reconstruct the full length sequence. However, non-uniform transcript coverage can obscure the underlying model by suggesting different splicing events. Even with uniform read coverage, there are scenarios where the problem of identifying the correct exon chaining model is intractable (Fig. 1). The third issue with short read RNAseq data arises from transcriptional noise. Transcriptional noise becomes problematic when it occurs within intronic or intergenic regions. The origin of these reads is unclear and transcript assemblers have taken different approaches to minimizing the influence of transcriptional noise [7–9]. Despite these efforts, there are some clear implications

of the phenomena. For example, due to the possible occurrence of transcriptional noise within intronic regions, it is difficult to determine if a transcript model should include a retained intron or not. When transcriptional noise occurs in intergenic regions it can be erroneously predicted as a gene or it can be fused with a neighbouring gene. When combining these three issues, the uncertainty of short read assembled transcript models becomes restrictive.

The annotation of most vertebrate genome sequences, except human and mouse, has been hampered by the lack of full length cDNA/transcript sequences for the species of interest and has instead had to largely rely upon Expressed Sequence Tags (ESTs) and their abundant successors, short read RNA-seq. As a result, the complexity of transcription of the chicken genome is underrepresented in the current genome annotation and constrains some analyses. For example, many differential expression analysis experiments rely on the annotation to define transcription events. Since a large number of alternative transcript models are likely missing in many vertebrate annotations, alternative transcription dependent mechanisms may have been unknowingly omitted from these studies.

While these issues are common in short read RNAseq data, they are practically eliminated with long read sequencing where the full-length of a transcript may be sequenced in a single read. With full-length sequencing, TSS and TTS can be easily defined since the reads span the entire length of the transcript. Similarly, predicting exon chaining from probabilistic models is not necessary. Transcriptional noise is reduced and in the cases where it does occur, it is more easily identified.

With the recent development of Pacific Biosciences (PacBio) SMRT Iso-Seq sequencing [10], it is now possible to attain high throughput, full-length transcript sequencing. While this technology has huge potential for transcriptome annotation, it still requires development for both library preparation and data analyses. Iso-Seq has been used in previous studies to identify transcript sequences [11–14], however, there are two main issues



Fig. 1 Short read transcript modelling problem. Example of transcript model that is impossible to resolve using short read data. Given the read support in yellow, it is impossible to determine which splicing model is real

with these earlier approaches. The first issue is that normalization of the RNA libraries was not performed, thus many low abundance transcripts may not have been sequenced due to the higher probability for attaining reads from high abundance transcripts. The second issue is that transcription start sites could not be confirmed due library preparation protocols lacking 5'-cap capture, thus the identified transcript sequences are not guaranteed to be full-length.

To address these concerns, we generated PacBio SMRT Iso-Seq sequencing data from chicken brain and embryo RNA. Both RNA libraries were normalized to reduce over-represented transcripts, however we only performed 5' cap selection on the embryo library. We also performed Illumina short read RNA sequencing on 20 tissue types to both verify transcribed loci and compare transcript models.

We identified important considerations for Iso-Seq sequencing and data analyses. Using this understanding of the data limitations, we surveyed the chicken transcriptome to discover transcriptional complexity similar to the human annotation. This complexity is comprised of the type and number of alternative transcription events, previously unannotated biotypes in chicken, and transcriptional sequence variance between species. We have also identified two classes of long non-coding RNA that are under-represented in all mammalian annotations.

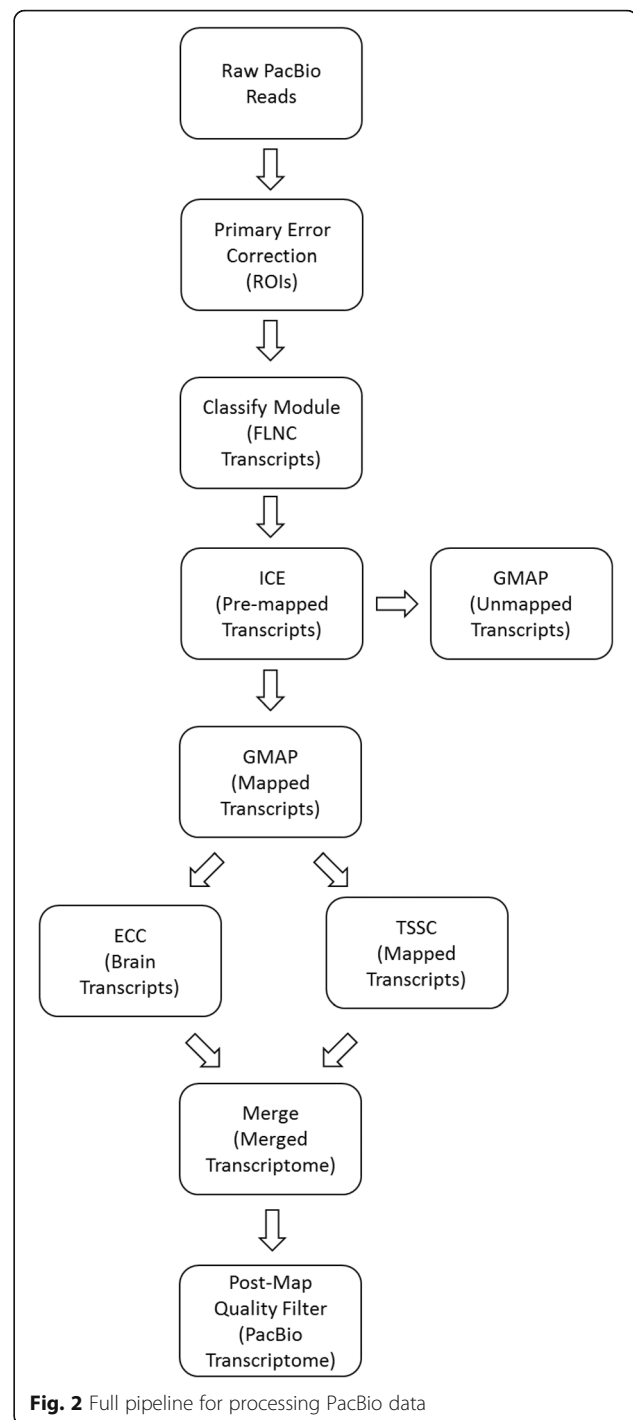
Our results provide guidance for future Iso-Seq studies as well as insight into chicken and all vertebrate transcriptomes. The data from this study were submitted to the European Nucleotide Archive (ENA) and used by Ensembl for their future chicken annotations.

Results

Processing PacBio data to create a high quality non-redundant PacBio transcriptome

Strategy for processing of PacBio Iso-Seq reads

Analysing PacBio Iso-Seq data requires a very different approach as compared to short read RNAseq data. Initial processing of this type of data focuses on reducing the final error rate of the acquired transcript sequences. While the raw error rate of PacBio sequencing is around 11-14% [10], the use of circular sequencing and computational error correction can greatly increase the final quality score. The software for achieving this is still in an early stage of development and evolving rapidly. We have adopted methodology supported by the PacBio development team known as the Iso-Seq pipeline, also known as the *pbtranscript-tofu* analysis suite [15], and incorporated it into our own pipeline (Fig. 2). The methods used to error correct Iso-Seq reads can have major implications for the limitations of downstream analyses. We have identified some major considerations when processing this type of data.



Raw data error correction

We attained 805,606 reads-of-insert (ROIs) from the brain and 247,626 ROIs from the embryo libraries. The lower yield for the embryo project was a result of issues with loading SMRT cells with a size selection of lower than 1-kb. Graphs for the read lengths for each size selection are available in additional files (see Additional file 1). Since every ROI sequence should begin with the

adapter sequence, we calculated the quality scores for each ROI by aligning the known adapter sequence to the adapter sequence within each ROI sequence and dividing the number of matches in the alignment by the length of the adapter sequence (see Methods). ROI average quality scores were 96.1% for brain 1-kb, 95.4% for brain 2-kb, 84.7% for embryo 0.8-kb, and 85.9% for embryo 2-kb. We ran *pbtranscript-tofu pbclassify* [15] with the ROIs as input to attain 515,175 full-length, non-chimeric (FLNC) transcripts for brain and 138,266 FLNC transcripts for embryo. After a further round of error correction using Iso-Seq iterative clustering for error correction (ICE) tool, from the *pbtranscript-tofu* analysis suite [15], we attained 211,292 transcripts for brain and 14,776 transcripts for embryo.

We mapped the resulting transcripts sequences to the *Gallus_gallus_4* genome assembly using GMAP [16]. 199,560 transcripts from brain and 11,881 transcripts from embryo mapped to this genome assembly. This left 11,732 brain and 3028 embryo transcripts unmapped. The unmapped transcripts are most likely a combination of transcripts which should map to the unassembled regions of the genome and transcripts which contain large errors missed in the previous filtering steps.

Collapsing transcript models to reduce redundancy

In previous studies [11–14], no 5' cap selection was performed, thus possible 5' degradation was ignored. In order to understand if the absence of 5'-cap selection in the library preparation would result in significant loss of TSS in our final transcript models, we collapsed the mapped transcript sequences from both the brain and embryo libraries using two methods from the PacBio *pbtranscript-tofu* analysis suite [15]. Both methods assume that the 3' end is intact, thus any transcript models with unique TTS are not collapsed. In the first method, termed Transcription Start Site Collapse (TSSC) (Fig. 3), transcripts with identical splice junctions and 3'-ends but

varying TSS are collapsed so that only the longest transcript is kept. While this method is inappropriate for libraries which were not 5'-cap selected, we still use the results from TSSC as a comparison. The second method, termed Exon Cascade Collapse (ECC) (Fig. 3), is identical to the first except that transcripts that are missing 5' end exons are included in the collapsing group. ECC is a more aggressive form of collapsing than TSSC and all transcripts that would be collapsed in TSSC would also be collapsed in ECC.

We looked at the ratio of the pre-collapsed transcripts to the collapsed transcript for each library from each collapsing method as an indicator of 5'-sequence loss. Since there should be no 5' sequence loss for the embryo library, the ratio of pre-collapsed to collapsed transcripts in the embryo library is used as a baseline for this comparison.

From 199,560 brain transcripts after running ICE, we attained 80,814 TSSC and 55,932 ECC models. From 11,881 embryo transcripts after running ICE, we attained 9368 TSSC and 8468 ECC models. Thus the number of transcript models drop by 59.5 and 72.0% for TSSC and ECC methods with the brain data, whereas the embryo transcript number only decreased by 21.2 and 28.7%, respectively. It is possible that these differences are caused by real biological differences in transcription start sites, however, 5'-sequence loss seems more likely and should not be ignored in the downstream analyses.

To reduce redundancy in our dataset we used the TSSC method for the embryo sequences and ECC for the brain sequences. This resulted in 55,932 transcripts from brain and 9368 transcripts from embryo after collapsing. Although it might seem strange for there to be a significant amount of collapsing within the embryo data there are biological reasons for this to occur. For instance, TSS are known to be variable so that otherwise identical transcripts can have different TSS, as shown by the evidence of wide promoter regions from cap analysis

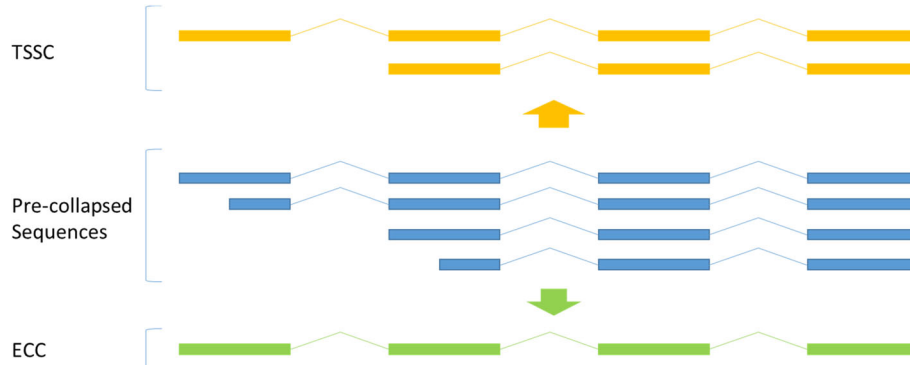


Fig. 3 Iso-Seq mapped read collapsing methods. Two methods for collapsing PacBio mapped sequences to remove redundant models: Transcription Start Site Collapse (TSSC) and Exon Cascade Collapse (ECC). ECC is more aggressive in collapsing

of gene expression (CAGE) studies [17]. While it is possible that using the TSSC method for the embryo data can result in the loss of unique transcript models that represent different TSS, due to the low coverage (compared to short read RNAseq data) for each transcript, it is impossible to identify whether differences in the observed TSS are a result of a single wide promoter region or from multiple distinct promoter regions. For the purposes of this study, we chose to follow a conservative approach which meant removing possibly redundant information at the cost of filtering out some real biological information.

Filtering out low quality models using post-mapping quality estimates

Mapping the transcript sequences to the genome can be seen as a final error correction step. The differences between the pre-mapped and post-mapped sequences provide an indication of the error rate for the sequences after all prior error correction and allows for the filtering of erroneous models that are a result of poor mapping. Post-mapped sequences are defined by using the genomic nucleotides for the predicted genomic coordinates of the mapped transcripts. True sequence variation can contribute to sequence differences, however the primary purpose of this transcriptome annotation is to identify models based on the reference genome. For each transcript sequence, we aligned the pre-mapped sequence with the post-mapped sequence and counted the number of mismatches within the alignment. We calculated the error rate by dividing the number of mismatches by the length of the transcript. Using this method, we selected for transcript sequences with less than 10% error rate. This resulted in 55,315 brain transcripts and 9206 embryo transcripts, greater than 98% retention for both sample types (Fig. 4a).

While a 10% error rate may seem high in comparison to short read data, the distribution of transcripts based on quality scores/error rates shows that the mode is 99% quality score. In addition, we are able to attain unique mappings due to the length of the sequences. Thus while the error rate makes this data unsuitable for polymorphism detection, it is low enough to provide accurate transcript models. There are several possible explanations for the occurrence of transcripts that did not meet our 10% error rate threshold. The observed discrepancy between the mapped and pre-mapped sequences could have resulted from sequencing error, errors in the reference genome assembly, and/or biological differences due to the difference between the genomes of the reference assembly specimen (red jungle fowl) and the chickens sampled in this project. Due to the ambiguity of the source of error, we chose to remove these sequences from our downstream analyses. Previous studies

did not report using this method of error correction which may indicate that some of their transcript models contained erroneous models [13, 14].

Identifying possible transcript truncation due to internal poly-A regions

Poly-A tail selection is a commonly used method of attaining transcript sequences with intact 3'-ends [11, 12]. However, internal stretches of A's can bind to oligo-dT primers thus resulting in a 3' truncated transcript sequence. In short read RNAseq sequencing this usually does not pose a large problem because many inserts will be sequenced and the farthest downstream 3'-end will usually be selected as the TTS. However, with single molecule long read sequencing, we make the assumption that each sequence has an intact 3'-end. To assess whether this is an appropriate assumption, we investigated the possible rate of occurrence of poly-A truncation by looking at the 3'-genomic sequence of each predicted transcript. The primers used for poly-A tail selection were designed to bind to a minimum of a stretch of 20 A's. Due to the prevalence of insertion/deletion sequencing errors in PacBio sequencing we used a 30 bp window. Thus for each PacBio transcript, we extracted the 30 bp downstream genomic sequence and looked for stretches of A's. If a PacBio transcript model were a result of internal poly-A truncation, we should see a stretch of at least 20 A's within this region. We allowed one mismatch within a string of A's and used the longest string of A's for our calculations. From the 64,277 identified transcripts, only 700 had a stretch of at least 20 A's immediately following their putative TTS in the genome sequence. Thus, around 1.1% of the deduced transcripts may be artificially truncated (Fig. 4b).

We also looked at the length of poly-A's within the respective ROI's as a comparison (The poly-A tails within the ROI sequences are removed during the *pbtranscript-tofu pbclassify* error correction step). If the length of poly-A's in the ROI's (Fig. 4c) are much longer than the length of internal poly-A's (Fig. 4b) then the prevalence of internal poly-A truncation is likely to be minimal. The ROI's have an average of 39.6 consecutive A bases in their poly-A tails and a peak at about 27 bp (Fig. 4c). This matches a previous study that found a peak of TAIL-seq tags with poly-A tails of about 20 nt in length [18]. Due to the dominance of true poly-A tails at around 27 bp it is non-trivial to differentiate transcripts with real poly-A tails from those that may be truncated due to internal poly-A stretches. However, since only a maximum of 1.1% of the transcript models could have internal poly-A truncation, this issue seems to have a limited effect on Iso-Seq data. This methodology for identifying possibly truncated transcripts can be used in future Iso-Seq studies to flag and/or filter transcript models.

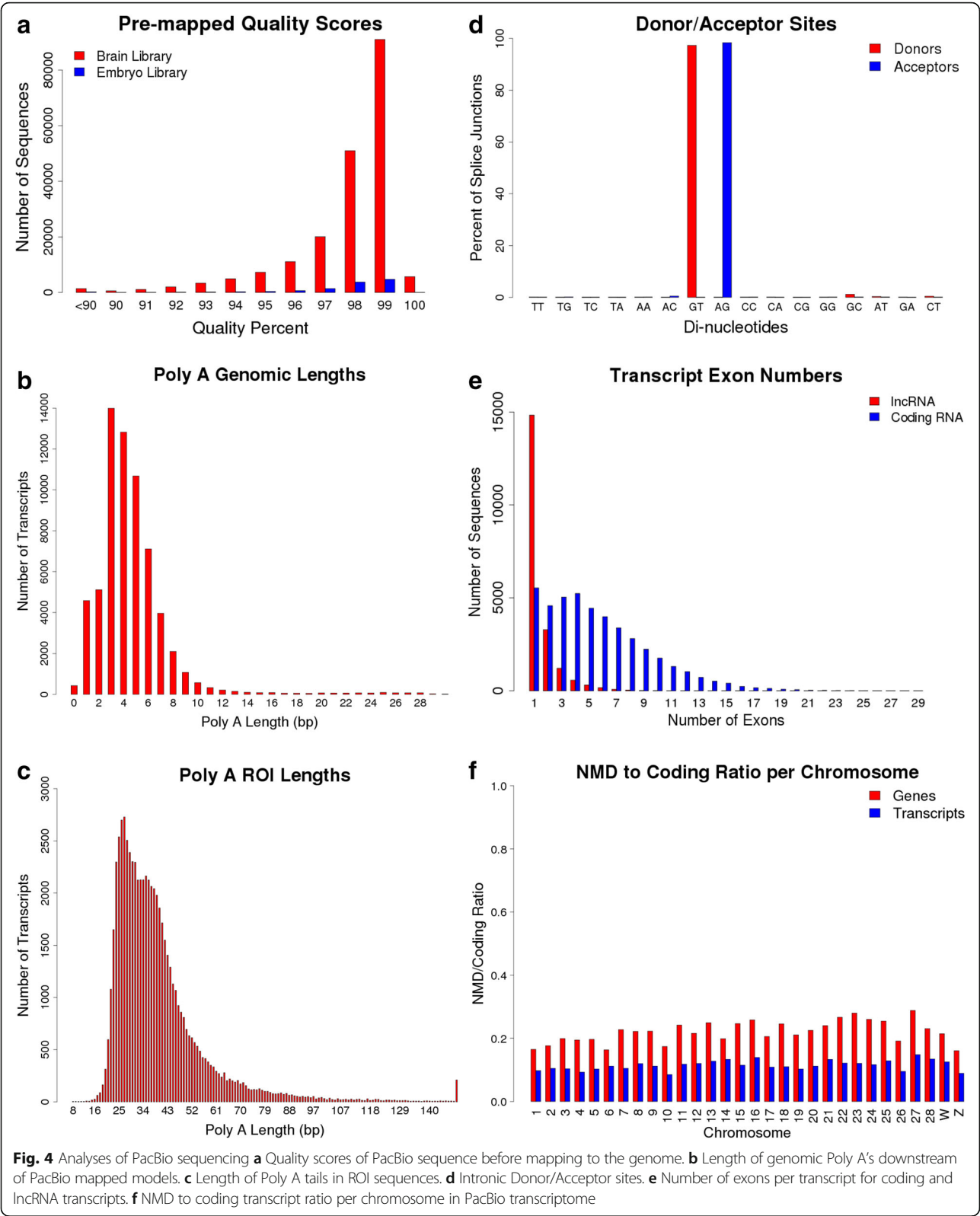


Fig. 4 Analyses of PacBio sequencing **a** Quality scores of PacBio sequence before mapping to the genome. **b** Length of genomic Poly A's downstream of PacBio mapped models. **c** Length of Poly A tails in ROI sequences. **d** Intronic Donor/Acceptor sites. **e** Number of exons per transcript for coding and lncRNA transcripts. **f** NMD to coding transcript ratio per chromosome in PacBio transcriptome

Verifying canonical splice sites

We looked at splice site donor and acceptor intronic dinucleotides to see if they conformed to canonical observations [19]. The GT-AG donor-acceptor sequence was used by 97.0% of brain transcripts and 98.0% of embryo transcripts, both very similar to the reported 98.7% in mammals [19]

(Fig. 4d). While GMAP does have a bias for mapping splice sites to conform to the canonical GT-AG scenario, the similarity between the canonical splice site percentages suggests that the splice site predictions are generally accurate.

Merging the transcriptomes from each sample

We merged the brain and embryo transcripts to form a unified transcriptome annotation to be used for further analyses. Merging was performed with the criteria that transcripts were merged if they had the same exon structures with less than a 10-bp difference for each exon start/end and less than 20-bp difference for the transcript start/end. We allowed for slight differences in exon start/end and transcript start/end to account for possible mapping errors as a result of insertion/deletion sequencing errors which are the most prevalent in Iso-Seq sequencing. When merging, the transcript with the earlier start site was used as the new representative transcript. This merging process resulted in a total of 64,277 distinct transcripts. Only 488 transcripts between the two sets were merged which translates to 244 shared transcripts. Of the 244 shared transcripts, 176 had the brain transcript as the new representative model, meaning that for this set, the brain transcripts had an earlier TSS.

We grouped the transcripts into putative gene models by clustering transcripts that had at least a one nucleotide overlap. This resulted in 29,013 putative genes which we will refer to simply as genes. Of these, 4579 genes had transcripts from both brain and embryo libraries with only 621 genes having only transcripts from embryo libraries. This indicates that while most genes are transcribed across sample types, the resulting transcripts differ. Thus providing more support for the observation that alternative transcription plays a significant role in tissue differentiation [20]. However, due to the lack of 5' cap selection for the brain dataset, there may be more shared transcripts than we observed simply because we lacked the 5' end of the brain transcripts. Since short read data is generally inaccurate with respect to isoform level quantification, this biological phenomenon would be very difficult to detect without long read sequencing.

Estimating gene numbers for unmapped transcripts

Since we were unable to use genomic locations to group the unmapped transcripts, we instead used the BLASR [21] mapper to find hits between the unmapped reads. Reads were grouped if they had same stranded BLASR hits. 11,732 unmapped reads from the brain and 3028 unmapped reads from the embryo were clustered into 8812 groups. This indicates a significant number of genes that are not currently represented in

the Chicken annotations due to gaps in the genome assembly. We excluded these unmapped transcripts from further analyses due to the uncertainty of the sequence quality and the effects that would have on the prediction methods we used.

Comparison with previous chicken PacBio transcriptome sequencing studies

In order to estimate the benefit of library normalization with respect to the efficiency of transcriptome coverage for each SMRT cell used, we compared our data to a previous study [11] where PacBio Iso-Seq long read sequencing was performed on RNA from chicken embryonic hearts. The embryonic heart study yielded 1,566,465 reads that mapped to the Galgal4 genome assembly. While the exact number of unique transcripts was not reported, 9221 novel isoforms were identified. We estimated the maximum number of unique transcripts that they could have acquired to be 31,081, which was calculated by adding their number for novel isoforms with the number of publicly annotated isoforms reported in their paper, 21,860 (16,743 from Ensembl and 5117 from RefSeq). Dividing their total possible number of unique transcripts by the number of reads they produced shows that, at most, only 2% of their reads were unique. While out of 482,325 mapped reads from our brain library, we found 55,315 (11.5%) unique transcripts. Thus the normalization method appears to have provided a transcriptome coverage efficiency of more than 5 times that of the previous study [11]. This means that for every SMRT cell used with the normalization method, 5 SMRT cells would be required without normalization to achieve the same amount of transcriptome coverage.

This transcriptome coverage efficiency calculation assumes that the previous study did not achieve full coverage of the transcriptome for their sample. While it is possible that they reached full coverage of their sample transcriptome, it seems unlikely since we found evidence for 44,898 transcripts from our chicken heart short read RNAseq data.

Exploring the PacBio transcriptome of the chicken reference genome

Protein coding and noncoding RNA genes and transcripts

We used three methods to find evidence for protein coding potential: Blastx [22] with the Uniprot Uniref 90 protein database [23], the Coding Potential Calculator (CPC) software [24], and the Coding Potential Assessment Tool (CPAT) [25]. Combining the results from the three methods, we found 43,738 putative protein coding transcripts from 14,421 genes and 20,539 putative noncoding RNA transcripts from 17,178 genes (Table 1). Within the noncoding RNAs (ncRNAs), we found that 23 transcripts were shorter than 200 bp which means the rest were classified as long noncoding RNAs.

Table 1 Classification of biotypes for PacBio transcriptome

# of Transcripts	Biotype
43,738	Coding RNA
20,516	LncRNA
23	Short ncRNA
4735	NMD transcript
13,873	Antisense Exonic
2139	Antisense Intronic

We noticed a difference in the number of exons between coding and noncoding transcripts. There were 14,831 noncoding single exon transcripts (72.2%) and only 5533 protein coding single exon transcripts (12.7%) (Fig. 4e). Thus within this dataset single exon transcripts make up the majority of noncoding RNAs.

We classified the lncRNAs by positional relationship to the predicted protein coding transcripts. There were 12,999 long intergenic noncoding RNAs (lincRNAs), 2675 antisense lncRNA, and 4967 sense overlapping lncRNA.

Nonsense mediated decay products

Nonsense mediated decay (NMD) products are transcribed alternative splice variants from protein coding genes that are not translated into proteins [26–28]. NMD products have similar sequences to protein coding transcripts but typically have been spliced so that there is an early stop codon [27]. We used the NMD prediction rules outlined in NMD studies [27, 29–31], which state that a premature termination-translation codon occurring at least 50–55 base pairs upstream of a splice junction provides strong evidence for NMD. Using this criteria, we identified 4735 putative NMD transcript candidates within our PacBio data (Table 1).

Although Ensembl did not make NMD predictions for the Ensembl (release 83) chicken annotation, they did have predictions for the human and mouse annotations. We ran our NMD method on Ensembl (release 83) human and mouse annotations to compare our methods. Out of 13,401 Ensembl human NMD transcripts, 13,263 were predicted to be NMD using our method (99% agreement). From our NMD predictions, out of 79,901 Ensembl annotated human protein coding transcripts, only 909 were predicted to be NMD which equates to a false discovery rate (FDR) of 6.4%. Similarly for 5229 NMD transcripts in the mouse annotation, 5152 were predicted to be NMD using our method (99% agreement). Out of 50,706 Ensembl mouse protein coding transcripts, we predicted only 341 transcripts to be NMD which equates to an FDR of 6.2%.

When overlapping the PacBio NMD transcripts with the Ensembl annotation, we found that 4137 NMD transcripts overlapped with 2517 Ensembl genes. We looked

at the ratio of the number of NMD to coding for both gene level and transcript level per chromosome and found that all chromosomes had a ratio at gene level between 0.16 and 0.29 (Fig. 4f). We ran the same analysis on the Ensembl (release 83) human and mouse annotations and found similar ratios ranging from 0.12–0.5 and 0.03–0.26 respectively.

Identification and classification of antisense transcripts

In the chicken PacBio transcriptome, there are 13,873 transcripts that are exonic antisense overlapping to at least one transcript and 6446 genes that are antisense to at least one gene. We looked at the numbers of coding and noncoding transcripts with respect to these antisense transcripts and found that there were 7107 transcripts involved in a protein coding to noncoding antisense overlap, 4765 transcripts involved in protein coding to protein coding antisense overlap and 2001 transcripts for noncoding to noncoding antisense overlaps (Table 1). When converting these to gene antisense overlap pairs we found 1571 protein to noncoding gene pairs, 1329 protein coding to protein coding gene pairs and 1036 noncoding to noncoding gene pairs.

Looking at intronic antisense overlap, where transcripts have at least one exon that overlaps with the intron of an antisense transcript, we found 2139 transcripts and 1115 genes with at least one antisense intronic overlap (Table 1). When considering coding predictions we found 354 protein coding to noncoding, 298 protein coding to protein coding, and 140 noncoding to noncoding gene pairs.

Transcriptional complexity in the chicken genome

Due to the large number of unique transcripts that were identified, we were able to make a general assessment of transcriptional complexity in the chicken genome. We looked at the ratio of transcripts to genes, retained introns, skipped exons, alternative exon starts/ends, alternative TSS and TTS, and single exon transcripts.

We found a ratio of 2.22 for transcripts to genes. This low number is due to the abundance (19,120 genes of which 13,265 are lncRNA genes) of genes with single transcripts many of which are single exon genes. If these single transcript genes are removed then the ratio increases to 4.56 transcripts to genes (Fig. 5a). These numbers are likely an underestimation as we have only characterised two, albeit transcriptionally complex, tissue types and some lowly expressed transcripts may have been missed in our brain and embryo libraries.

For assessing alternative TSS we only used the transcript sequences from the embryo library since this library had 5'-cap selection and therefore should have intact 5'-sequences. We removed all genes with only one representative transcript since these would by default have only one TSS. There were 2037 genes that

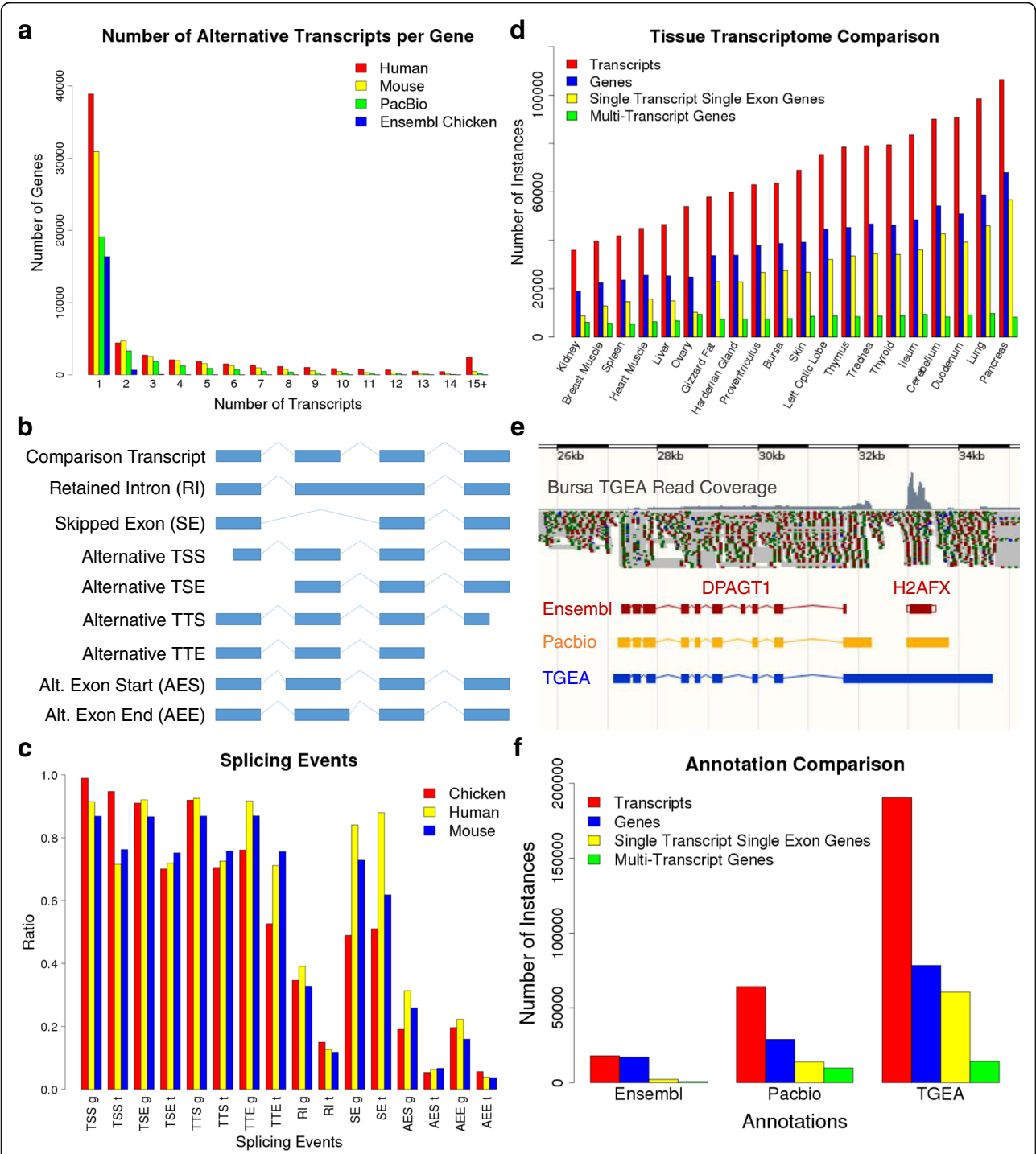


Fig. 5 Alternative splicing. **a** Comparison of number of alternative transcripts per gene between Ensembl annotations and PacBio transcriptome. **b** Classifications for alternative transcripts. **c** Comparison of rate of occurrence for the different classes of alternative transcripts between Ensembl human, mouse, and the chicken PacBio Transcriptome. Abbreviations for x-axis labels explained in Fig. 5b. **d** Comparison between STSE and MT genes for TGEA transcriptome. **e** Example of overhang event. **f** Comparison of the number genes and transcripts for Ensembl, PacBio and TGEA

matched these criteria and only 73 had only one TSS. Thus 96.4% of these genes had multiple TSS. The high rate of multiple TSS genes is presumed to be a combination of transcription factor binding wobble and alternative transcription start exons (TSE). If we ignore TSS caused by wobble and only look at TSE, 594 genes have a single TSE which means 70.8% of these genes have multiple starting exons.

For TTS we use both brain and embryo transcript sequences since both libraries had been selected for poly-A tails. Again all single transcript genes were removed which resulted in 9893 genes. Only 801 genes had a single TTS which means that 91.9% had multiple TTS. However, if we look at alternative transcription termination exons (TTE), we find that 2365 genes have a single TTE which means 76.1% have multiple TTE.

We also looked at occurrences of retained introns and skipped exons within both brain and embryo transcripts using only multi-transcript genes. We define retained introns as exons which overlap an entire intron (Fig. 5b) from another transcript. There are 3429 multi-transcript genes which have retained introns which equates to a rate of 34.7%. We define skipped exons as exons which are completely overlapped by an intron in another transcript (Fig. 5c). There were 4939 genes with at least one occurrence of skipped exons which equates to a rate of 49.9%.

We looked at alternative exon start (AES) and ends (AEE). For this set we used both brain and embryo transcripts but only assessed internal exons so that we did not include TSE and TTE. We also excluded retained intron exons from this set. There were 8006 genes with no AES which equates to a rate of 19.1%. There were 7952 genes with no AEE which equates to a rate of 19.6%. So the rates of AES and AEE are quite low as compared to other alternative splicing events.

We were interested to see if there were any alternative splicing differences between protein coding and lncRNA genes. Out of 14,421 protein coding genes, 6597 had only one transcript which gives a multi-transcript rate of 54.3%. Out of 17,178 lncRNA genes, 15,162 had only one transcript which gives a multi-transcript rate of 11.7%. Thus lncRNA genes are much less likely to contain alternative transcripts. We also noticed that lncRNA transcripts were much more likely to have only one exon. Out of 20,539 lncRNA transcripts, 14,831 contained only a single exon. Whereas, out of 43,738 protein coding transcripts, there were only 5533 single exon transcripts. lncRNA transcripts had a rate of 72.2% for single exon transcripts as compared to 12.7% for protein coding transcripts. When adjusting the multi-transcript rate for only multiple exon genes, there is a rate of 67.0 and 37.0% for coding and lncRNA genes respectively. Thus even after accounting for the high number of lncRNA single exon genes, coding genes are more likely to have alternative transcription.

Comparison of transcriptome assemblies derived from short and long read RNA sequencing data

We created a tissue gene expression atlas (TGEA) derived from the assembly of short read RNAseq data of

Table 2 Number of transcripts and genes by tissue type for TGEA transcriptome

Tissue	# of Transcripts	# of Genes
Kidney	35,867	18,916
Breast Muscle	39,649	22,357
Spleen	41,831	23,546
Heart Muscle	44,898	25,520
Liver	46,523	25,253
Ovary	53,933	24,787
Gizzard Fat	57,922	33,670
Harderian Gland	59,873	33,791
Proventriculus	62,954	37,824
Bursa	63,644	38,673
Skin	68,982	39,211
Left Optic Lobe	75,457	44,567
Thymus	78,491	45,312
Trachea	79,103	46,730
Thyroid	79,440	46,285
Ileum	83,541	48,446
Cerebellum	90,088	54,212
Duodenum	90,665	50,902
Lung	98,514	58,762
Pancreas	106,430	68,006

20 tissue types from J-line layer chickens (Table 2) to compare and independently validate the PacBio transcriptome. We merged the identified transcripts from each short read RNAseq tissue dataset into a single transcriptome annotation to create the TGEA.

The TGEA transcriptome predicts 78,351 genes with 190,474 transcripts. Thus the TGEA has 2.7 times the number of genes and 2.96 times the number of transcripts as compared to the PacBio transcriptome. While this difference is most likely explained by the inclusion of many more tissue types in the TGEA as compared to the PacBio transcriptome, there are also some fundamental differences in the proportion of multiple transcript and single transcript single exon (STSE) genes (Fig. 5d). When only comparing multiple transcript genes, the PacBio transcriptome has 9893 genes while the TGEA transcriptome has 14,220 genes. However, for STSE genes, PacBio has 13,824 genes while TGEA has 60,576 genes. Thus the ratio of STSE genes to multiple transcript genes is 1.40 in PacBio and 4.26 in TGEA. Of the total STSE genes in the PacBio transcriptome, 12,603 are classified as lncRNA.

We looked at genomic overlap between PacBio and TGEA transcripts to estimate the transcript coverage for each dataset. There were 9368 PacBio transcripts

which had no overlap with TGEA transcripts. However, when we guide the transcript assembly for the TGEA dataset using the PacBio transcripts, we find that only 18 PacBio transcripts have no coverage. This indicates that despite the high depth of sequencing and wide tissue coverage of the TGEA dataset, a large number of transcripts were not predicted even though there were data to support their existence. This may be a result of the difficulty in differentiating transcriptional noise from true transcripts. Thus the PacBio transcripts missing in the TGEA transcriptome were difficult to distinguish from noise using short read data.

There were 108,651 transcripts from 15,633 genes in the TGEA which overlapped the PacBio transcriptome. So 43% of the TGEA transcripts and 80% of the TGEA genes are not covered by the PacBio transcriptome. However, of these TGEA models with no PacBio transcript overlap, 76.6% of the transcripts and 91.5% of the genes are from single transcript single exon genes. While these may represent true transcripts and genes, it is difficult to be sure that these are not the result of transcriptional noise using only short read evidence.

We noticed during manual inspection of the PacBio and TGEA overlaps that some transcript models in the TGEA transcriptome seemed to be a merging of two adjacent genes in the PacBio transcriptome. We call this event an “overhang gene” (Fig. 5e). To investigate the abundance of these events, we searched for all TGEA transcripts which overlapped two PacBio genes. We identified 2515 overhang events where an upstream and downstream PacBio gene is represented as one merged gene in the TGEA transcript model. Of these, 208 events occur where the downstream gene model has a confirmed start site due to the presence of transcript models from the embryo data. Out of these, 79 overhang events have external support from Ensembl (release 83) chicken annotation showing that each gene is a separate well annotated gene. To understand the more general problem of gene merging we looked at all gene merging events where one TGEA transcript merged two or more PacBio genes. We identified 4254 merged gene events involving 10,991 PacBio genes.

Comparison of the PacBio transcriptome with public annotation

Ensembl and NCBI are the two major sources of public annotation for the chicken genome. Since the NCBI chicken annotation contains far fewer transcripts and genes (6352 and 6027 respectively) as compared to the Ensembl chicken annotation (release 83) and 96.8% of the transcripts in NCBI are also contained within

Ensembl, we chose to focus our analyses on the Ensembl annotation. The Ensembl (release 83) chicken annotation contains 17,108 genes with 17,954 transcripts. There are 15,508 genes annotated as protein coding, 42 predicted as pseudogenes, 150 ambiguous RNA, and the rest are an assortment of short noncoding RNA. There are no annotated lncRNA. It has a ratio of 1.05 transcripts per gene model with only 745 multiple transcript genes. For these multiple transcript genes, the ratio of transcripts per gene is 2.14. The PacBio transcriptome has a ratio of 2.22 transcripts per gene when including the entire gene set and 4.56 transcripts per gene for multiple transcript genes (Fig. 5f). The Ensembl (release 83) chicken annotation contains 969 antisense genes as compared to 6446 antisense genes for the PacBio transcriptome. These are genes that overlap at least one gene on the opposite strand.

Out of the 64,277 PacBio transcripts, 21,887 had no overlap with Ensembl transcript models and are thus considered to be novel. Of these, 7414 transcripts had no sense exonic overlap with any Ensembl predicted transcript but were either antisense (exonic or intronic) or had a sense intronic overlap. These transcripts could be further classified based on their coding potential so that 5049 were noncoding and 2365 were protein coding (Table 3). The remaining transcripts were located in intergenic regions. Of these, 11,880 were predicted to be noncoding while there were 2593 intergenic coding transcripts (Table 3).

Comparative genomics and phylogenomic profiles of chicken PacBio transcripts

To understand the conservation of these sequences across birds and other vertebrate species, we mapped the transcript sequences (using GMAP with default parameters [16]) onto the genomes of several avian species

Table 3 Classification of coding and noncoding transcripts by gene overlap

# of Transcripts	Coding class	Exonic antisense	Intronic antisense	Intronic sense
1634	ncRNA	Yes	No	No
1262	ncRNA	No	Yes	No
2047	ncRNA	No	No	Yes
32	ncRNA	Yes	No	Yes
74	ncRNA	No	Yes	Yes
11,880	ncRNA	No	No	No
1478	coding RNA	Yes	No	No
200	coding RNA	No	Yes	No
575	coding RNA	No	No	Yes
55	coding RNA	Yes	No	Yes
57	coding RNA	No	Yes	Yes
2593	coding RNA	No	No	No

as well as representatives from other vertebrate classes (Table 4). The avian species with the best genome assemblies for each phylogenetic grouping were selected for this analysis. We used this selection criteria so as to avoid mapping biases from low quality genome assemblies. These mappings do not directly relate to orthologs but rather provide a general indication of transcript sequence conservation between species.

Only 0.2% of the total chicken PacBio transcripts did not map to any non-chicken species. While 8.8% of the chicken PacBio transcripts mapped to all species spanning 300Mys. Of these, 98.8% were predicted to be protein coding while 1.2% were predicted to be lncRNA. Of the lncRNA that mapped to all species, 52.3% were predicted to be intergenic. When focusing only on avian species, we see that 61.8% of the chicken PacBio transcripts map to all avian species. From these transcripts which map to all avian species, we see 82.8% predicted as protein coding and 17.2% predicted to be lncRNA (percentages given with respect to the total number of transcripts which map to all avian species included in this analysis). Out of the lncRNA transcripts that mapped to all avian species, 47.1% are classified as lincRNA. We produced heat maps to display this analyses with a colour scale indicating the quality percent of mapping for each transcript (Fig. 6 a-c). The quality percent is defined by the number of matching nucleotides divided by the total length of the transcript when aligning the chicken PacBio transcripts with their projected sequence when mapped to other species.

Discussion

Noncoding transcripts

Long noncoding RNA

In the Ensembl (release 83) annotation, there are 24,149 lncRNA transcripts predicted in human and 8391 predicted in mouse. Thus our 20,516 predicted lncRNA transcripts are similar in number to that found in the human annotation, which has the highest number of annotated lncRNAs of any Ensembl annotated vertebrate genome. While the mouse annotation usually benefits from homology based predictions from humans, the lack of conservation for lncRNA sequences has made homology methods mostly ineffective. Previous studies have shown that a large proportion of the human lncRNA are primate specific [32], which would explain the comparatively low number of identified lncRNA in mouse. The similar numbers of identified lncRNA in the Ensembl human annotation and the PacBio chicken annotation suggests that lncRNA are extremely underrepresented in the annotations of mouse and other species.

The Ensembl annotated lncRNAs are classified into three main categories: lincRNA, sense overlapping lncRNA, and antisense lncRNA. However, it is important

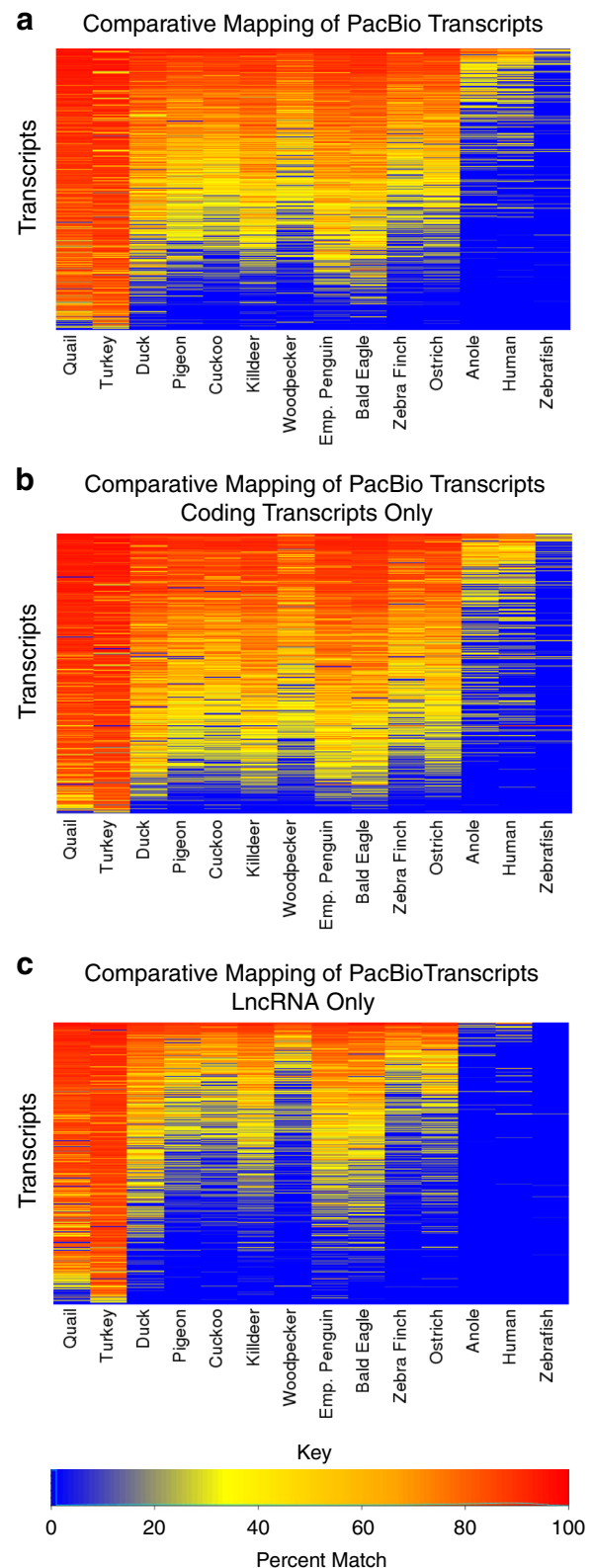


Fig. 6 Comparative mapping of PacBio transcripts. **a** Heatmap of PacBio transcripts mapped onto other species' genome assemblies. **b** Only the coding transcripts. **c** Only the lncRNA transcripts

to note that there are no biotype designations for sense exonic overlapping lncRNA in the Ensembl annotation. The sense overlapping class is comprised of two Gen-code defined biotypes termed *sense_overlapping* and *sense_intronic*. *Sense_overlapping* refers to lncRNA transcripts that have a protein coding gene within their introns. *Sense_intronic* refers to lncRNA transcripts that occur within the intron of a protein coding gene. Neither of these correspond to any exonic overlap, thus they are both sense intronic overlapping lncRNA. There is, however, a biotype classification labelled “*processed_transcript*” which is defined as a transcript with no open reading frame. There are transcript models within this group which meet the criteria for sense exonic lncRNA, however, due to the lack of evidence to support these models it is unclear how many represent true sense exonic lncRNA. Thus there are three sub-classes for lncRNA within the Ensembl annotation with a loosely defined 4th class which contains sense exonic lncRNA but not at an annotation level that can be used with high confidence. This means that the proportion of sense exonic lncRNA in human and mouse is unknown.

For both the human and mouse annotation, lincRNA make up roughly half of the total, while sense intronic lncRNA represent less than 10% of the total (Fig. 7a). Thus proportions of these classes seem to be well conserved within mammals. However, the relative proportions of the lncRNA sub-classes in the PacBio chicken annotation are very different. This difference seems to be due in large part to the inclusion of sense exonic overlapping lncRNA which make up 17% of PacBio chicken lncRNA transcripts (Fig. 7a). This difference could represent real biological differences between mammalian and avian genomes such that antisense lncRNA are more common in mammals while sense overlapping lncRNA are more common in birds. However, when we used our sense exonic overlap prediction tool on the Ensembl human and mouse *processed_transcript* models, we found 24,385 and 11,901 sense exonic lncRNA transcripts respectively. If these numbers are included in the proportion of lncRNA types then they would equate to 48 and 57% respectively. This would indicate that sense exonic lncRNA are actually the most abundant type of lncRNA. However, due to the dearth of evidence for these models, it is difficult to say whether this reflects reality. The proportions of lncRNA sub-classes within the PacBio chicken annotation may provide an estimate for the rate of occurrence of sense exonic lncRNA in human and mouse as well as other vertebrate species.

Another startling contrast between the human/mouse Ensembl (release 83) annotations and the PacBio chicken annotation is the proportion of the number of exons for lncRNA transcripts. In the human/mouse Ensembl

annotations, two exon lncRNA transcripts are the most commonly occurring (Fig. 7b). However, the PacBio chicken transcriptome show that single exon lncRNA transcripts are by far the most abundant. While this difference could be due to real biological differences between birds and mammals, no conclusions can be made because many of the lncRNA prediction methods for the human and mouse annotations removed single exon lncRNA models [33]. The practice of removing single exon lncRNA models is useful when dealing with models that are assembled from short read data since it is difficult to ascertain whether these models are truly single exon transcripts or the result of transcriptional noise. However, this puts a strong bias against the prediction of single exon lncRNA transcripts which has likely resulted in the underrepresentation of these transcripts. Thus the proportion of single exon lncRNA transcripts in the PacBio chicken annotation may indicate that these are also the largest group of lncRNA in other vertebrate species. If this is true, then a large portion of lncRNA have not been identified due to the practice of filtering out single exon lncRNA models.

Non-sense mediated decay transcripts

In comparison to the proportion of NMD products in human and mouse, our NMD predictions for chicken appear to be similar. Our predictions for NMD in chicken also show a more uniform ratio of NMD to coding transcripts across the chromosomes. Since 2517 Ensembl genes have NMD overlap, it appears that NMD may play a large role in protein expression regulation within the chicken. Considering the important biological implications of NMD products [34], the lack of annotated NMD transcripts in the public chicken annotation could have concealed important gene expression information in previous studies.

Antisense genes

The most common pairing for both exonic and intronic antisense genes is that of a protein coding gene with a noncoding gene which is supported by reports in mammals [35, 36]. The predominance of the coding to noncoding pairs suggests that there may be some regulatory relationship between the coding and noncoding genes in each pairing. While the mechanism of regulation is still mostly unknown, it has been proposed that one way in which an antisense gene can regulate a sense gene is by inhibiting transcription of the sense gene through transcriptional collision [37]. Thus the protein coding genes within these antisense pairings may be down regulated by the transcription of their noncoding antisense partners. In these situations, it is the action of transcription that is functional as opposed to the transcriptional product. Thus the sequence of the antisense partner is

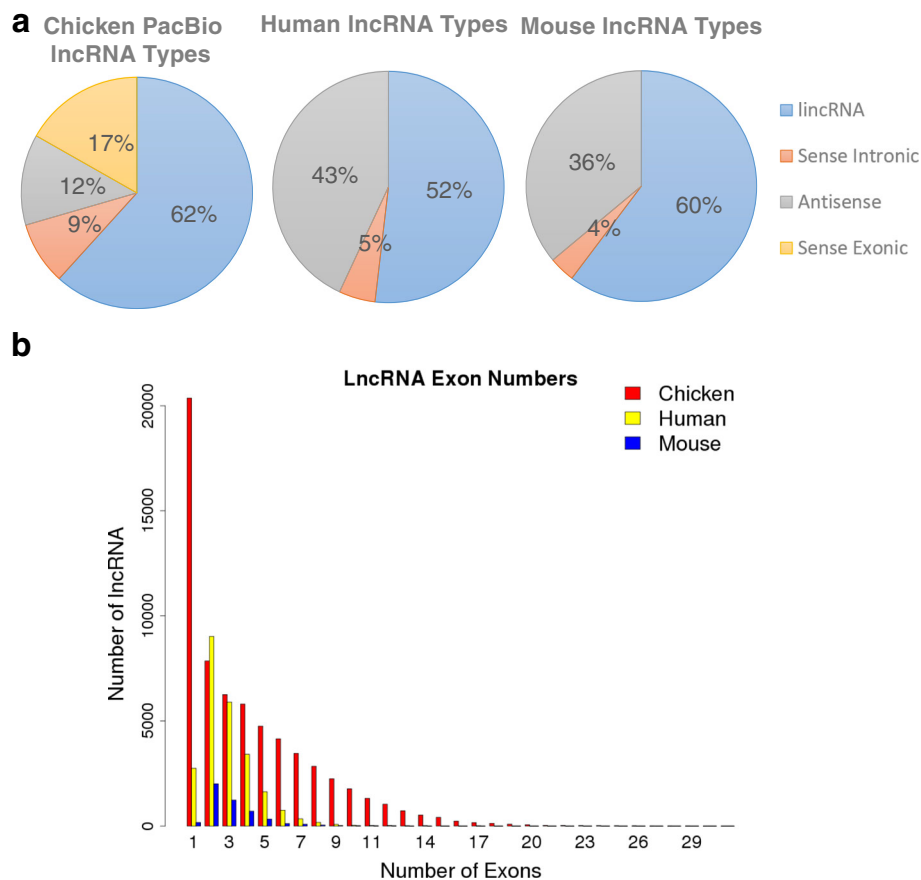


Fig. 7 Characterization of lncRNA. **a** Proportions of each class of lncRNA for chicken PacBio, Ensembl human, and Ensembl mouse annotations. **b** Proportion of exon numbers for lncRNA transcripts for chicken PacBio, Ensembl human, and Ensembl mouse annotations

essentially meaningless and almost completely free of selection pressure (aside from exonic overlapping regions). This would explain why lncRNA sequence conservation is so low as compared to protein coding genes.

The abundance ranking of antisense pairs from coding to noncoding, coding to coding, and then noncoding to noncoding has also been found in mammals [35]. The consistency of this ranking order within this study and within mammalian studies stands out as a peculiar coincidence. It suggests that coding to noncoding antisense regulation is a widely adopted and significant form of regulation within vertebrates. The coding to coding pairs may be a relic of ancient genomes where genomic compactness offered some selective advantage. However, it is perplexing as to why noncoding to noncoding pairs would be the least abundant. Due to the lack of sequence conservation for lncRNA, many believe that the majority of lncRNA lack function and refer to them as transcriptional noise. If they are truly non-functional, then their transcription near functional genes would likely have negative effects for several reasons. For instance, competition for access to the region by transcription factors. If lncRNA

are predominantly non-functional, it is more likely for them to occur near each other and not near useful genes. Thus lncRNA genes should make up the most abundant antisense pairs. The growing evidence that noncoding to noncoding pairs are the least abundant suggests that the majority of lncRNA are functional and their sequences are functionally important as well.

The prevalence of exonic pairs over intronic pairs offers another unintuitive result. Since intronic pairs have less sequence dependency between the two genes, it seems more probable for these pairings to arise. Yet there are less than a sixth of the amount of intronic gene pairs as compared to exonic. This large discrepancy suggests that there is some functional reason for why antisense exonic pairs dominate. It may be that the exonic sequence overlap allows RNA binding between the antisense products which could be used for up or down regulation. If this were the case, then perhaps the majority of antisense pairings represent a regulation relationship between the antisense genes. This would make sense from transcriptional collision alone but the RNA binding theory may add another level of regulation tuning.

Alternative transcription events

Multiple transcript genes within the PacBio chicken transcriptome display a high level of complexity with regard to transcription initiation and termination which is similar to that found in mammalian annotations [38]. When comparing the different alternative transcription events (Fig. 5b), the most dominant events are alternative TSS and TTS. This result matches the human and mouse annotations (Fig. 5b). It is possible that the TSS is a major factor in determining the splicing of the transcript such that each alternative transcript is somewhat defined by the TSS. This would mean that the polymerase binding site defines the alternative transcript which would be a mechanism for regulating alternative transcription.

Skipped exon events are the most dominant alternative splicing event with 49.0% of the PacBio genes having at least one occurrence and a rate of 51.0% for PacBio transcripts. While skipped exon events are the most dominant in both the human and mouse annotations, they occur at higher rates, 84.1 and 72.8% of genes, respectively. This may indicate that they play a lesser role comparatively in the chicken genome. The lower relative rate of occurrence for skipped exons in the chicken transcriptome may also contribute to the density of the genome, since skipped exons are related to greater overall intronic regions as they are effectively introns for other transcripts.

For retained introns, alternate exon starts, and alternate exons ends, there is a significant difference between the rate of occurrence when calculating per gene versus per transcript, with each event having a higher per gene rate than per transcript rate. This means that they tend to be spread out among genes but with fewer occurring within each gene. This contrasts the rates for skipped exons where there is actually a higher rate per transcript than per genes in the PacBio transcriptome. This suggests that these events may be related to a type of RNA product which does not benefit from a variety of these events, such as NMD products, where introducing an early stop codon is all that is needed.

Long read versus short read RNA sequencing data

If we consider the TGEA as a representation of an upper limit for transcribed loci discovery, then the difference in ratios between TGEA and PacBio for STSE to multiple transcript genes suggests that the majority of unannotated transcribed regions are STSE genes. However, another explanation is that many of these novel STSE genes in the TGEA transcriptome are a result of RNAseq noise. Since STSE genes have no splice junctions, there is no other supporting evidence for the existence of these genes except for read coverage. Since read coverage for a specific locus may be influenced by sequence similarity to another locus or errors in the genome assembly, it is possible that the supporting reads belong to another locus. It is also

possible that a STSE gene is actually an exon from another gene, but due to issues with low read coverage not linking the exon to the rest of the gene, the model was predicted incorrectly. Thus it is difficult to say how many of the TGEA transcripts/genes are accurate.

On the other end, the 9368 PacBio transcripts with no overlap from the TGEA indicates that there may be many transcripts which go undetected with short read sequencing. This under-prediction can be the result of genes with low expression levels or genes with sequence similarity to other loci (such as paralogs).

The relatively large number of gene merging events (4254) in the TGEA transcriptome indicate a clear issue with transcriptomes assembled from short read RNAseq data. Each gene merge event represents an incorrect transcript model that would be misidentified using standard annotation pipelines that rely on open reading frames and transcript length. While investigating gene merge events, we noticed that the transcript assembly errors seemed to be a result of short read noise. This noise is manifested as a low coverage of reads over intronic and intergenic areas. These noisy reads can bridge between transcripts thus resulting in merged gene models. This noise also makes it difficult to detect retained introns. Due to issues with noise, most assemblers use some method of thresholding to decipher when intronic reads are noise or real. However, filtering out noise reads from real reads is non-trivial and relies on low variance of read coverage over the transcript, which is rarely the case. Therefore, while the TGEA dataset can provide a rough estimate of transcribed loci, it is not recommended for identifying full length transcript sequences.

Comparing the PacBio transcriptome to the Ensembl annotation

The large difference in the number of anti-sense genes between Ensembl and PacBio is partly explained by the greater number of genes in the PacBio chicken transcriptome. However, it is also indicative of the limitations of the short read RNAseq data that was used for the Ensembl chicken annotation [5, 39]. Much of these data were generated using unstranded library preparation protocols which made it impossible to resolve anti-sense transcripts. Without stranded RNAseq data, anti-sense transcripts can look like extensions of the sense transcripts or can be filtered due to their non-conformance with the dominant transcript model. As a result, these models may have been omitted or represented incorrectly.

Due to the lack of lncRNA models in the Ensembl annotation, the large number of novel ncRNA predicted by PacBio sequencing is somewhat expected. The number of novel intergenic protein coding transcripts, however, was higher than we expected so we investigated the possible reasons for their absence in the Ensembl annotation. There

were 634 transcripts which had no hits against the Uniref 90 database. This subset represents transcripts with no or low sequence similarity to known proteins. Ensembl may have discarded these in their pipelines since they would be difficult to confirm as protein coding. These also represent possible avian specific proteins. There were 891 transcripts which were antisense (exonic or intronic) to a transcript either in the PacBio annotation or the TGEA. Similar to the transcripts that were antisense to Ensembl transcripts, these transcripts represent complex transcribed loci where short read data may not provide enough information to resolve the overlapping transcripts. There were 719 transcripts which did not have matching transcript models in the TGEA. Thus these transcripts could not even be assembled with short read data. This is mostly likely due to low and/or variable coverage. There were 967 transcripts left after removing the no hit, antisense, and no TGEA sense overlap transcripts. Therefore the majority of these previously unannotated transcripts can be explained by the limitations of short read RNA sequencing.

Comparative genomics provides functional support for PacBio transcript predictions

The mapping of over 99.8% of PacBio chicken transcripts to other genome assemblies provides support for the PacBio models and sequence conservation also predicts functional constraints on these transcripts. While there is some variability of genome assembly quality among the avian species, there is a clear trend for species that are evolutionarily closer to chicken to have more matching transcript sequences. The galliformes show a high proportion of similar transcripts (Fig. 6a), while there is a dropping off of the number of mapped transcripts for species at a further phylogenetic distance away from chicken. For the non-avian species, relatively few transcript mappings show similarity.

Comparing the coding and lncRNA transcripts, we see that there is a very different trend for sequence conservation with coding transcripts being relatively well conserved across the avian species (Fig. 6b) while a more noticeable drop off occurs with the lncRNA transcripts (Fig. 6c). This complements previous observations that lncRNA have low conservation as compared to protein coding genes [40] and homology approaches have limited effectiveness beyond closely related species.

Conclusions

We identified a large number of events where transcriptional complexity make it difficult or impossible to attain the true transcript sequences from short read data. As the current public annotation of the chicken genome by Ensembl has relied upon incomplete cDNA sequences (ESTs) and short read RNA-seq data, the complexity of the chicken transcriptome is currently underrepresented.

The current underrepresentation of transcriptional complexity with respect to the number of alternative transcripts can have consequences for analyses that rely on these models. Important transcriptional events can be missed or misrepresented thus obscuring underlying biological processes. Using PacBio sequencing to create a high quality transcriptome annotation can correct these issues that are common in many of the public annotations. More advanced analytical tools can be developed to take advantage of the long read transcriptome by using information which could identify problematic areas in short read data during transcript quantification experiments. These areas include multi-mapping loci, repeat regions and ambiguous splice junctions.

Long read transcript models also improve functional annotation since many annotation pipelines must assume that the supplied transcript sequences represent real splicing and correct reading frames. The ability to disambiguate overlapping transcripts or genes sheds light on transcriptome complexity that was previously unannotated in chicken. The PacBio chicken transcriptome suggests a level of transcriptional complexity that is more consistent with expectations based on the well-characterised human genome.

Methods

Pacific Biosciences Iso-Seq long read sequencing

For the brain library, brain tissue was collected from an adult J-Line chicken (brown leghorn) bred at the Edinburgh Poultry Research Centre and the extracted RNA sample was sent to GATC Biotech (Konstanz, Germany) for library preparation and sequencing. The total RNA sample was examined using capillary electrophoresis with a Shimadzu MultiNA microchip electrophoresis system (Shimadzu Corporation, Kyoto, Japan). Poly(A) + RNA was selected using an oligo(dT)-linker primer and cDNA was produced using M-MLV H⁻ reverse transcriptase. The cDNA was amplified using PCR with 16 cycles. Normalization was performed by denaturing and reassociating the cDNA. Double stranded cDNA were removed using a hydroxylapatite column. The remaining cDNA were then amplified using PCR with 8 cycles. The cDNA were then size selected for 1 kb and 2 kb lengths using Ampure beads (Agencourt BioSciences Corporation, Beverly, Massachusetts). For the 1 kb cDNA, 11 SMRT cells were used. For the 2 kb cDNA, 14 SMRT cells were used.

For the embryo library, an embryo was collected at Hamburger-Hamilton stage 26 from an ISA Brown chicken bred at the Edinburgh Poultry Research Centre and RNA was extracted. The RNA sample was sent to GATC Biotech for library preparation and sequencing. The total RNA sample was examined using capillary electrophoresis with a Shimadzu MultiNA microchip electrophoresis system. Poly(A) + RNA was selected and treated

with Terminator exonuclease (+TEEx). The 5'CAP structures were removed using tobacco acid pyrophosphatase (TAP). Then an RNA adapter was ligated to the 5'-monophosphate of the RNA. The cDNA was synthesized using an oligo(dT)-adapter primer and M-MLV H- reverse transcriptase. The cDNA was amplified using PCR with 13 cycles. Normalization was performed by denaturing and reassociating the cDNA. Double stranded cDNA were removed using a hydroxylapatite column. The remaining cDNA were then amplified using PCR with 8 cycles. The resulting cDNA was purified with the Agencourt AMPure XP kit. The cDNA were then size selected for 0.8 kb and 2 kb lengths using Ampure beads. For the 0.8 kb cDNA, 16 SMRT cells were used. For the 2 kb cDNA, 17 SMRT cells were used.

Long read transcriptome processing

Raw data was processed into error corrected reads of insert (ROI's) using the PacBio SMRT Analysis Package with default parameters. The ROI's were then processed using the Iso-Seq Tofu pipeline [15]. We used the Classify module with default parameters to remove adapter sequences, poly-A tails, artificial concatemers, and 3' truncated transcript sequences which resulted in our set of FLNC transcripts. For an additional level of error correction we ran PacBio ICE software without the Quiver step on the FLNC transcripts [15]. The ICE software clusters transcripts by alignment using BLASR and then error corrects using the alignments. This results in a higher quality set of transcript sequences and the removal of redundant data. Due to the computation time restraints we ran ICE independently on each size selection from the brain.

The resulting sequences were then mapped to the Galgal 4 reference genome assembly using GMAP [16] using default parameters. The GMAP result bam files are then processed using the Iso-Seq Tofu Collapse module (in the Iso-Seq pipeline) which merges transcripts based on genomic coordinates. There are two methods of doing this which are explained in the Results section.

For each transcript we collected the quality scores of the ICE cluster sequences contributing to that transcript model. We estimated quality score by aligning the pre-mapped sequence to the post-mapped sequence using MUSCLE [41] and counting the number of mismatches and gaps. We then took the longest supporting cluster for each transcript and removed the transcript from our working list if the quality percentage of the longest supporting cluster was less than 90%.

ROI quality scores were calculated with a similar method. Adapter sequences were aligned to the ROI sequences using MUSCLE and the quality score was calculated by counting the number of mismatches.

The collapsed transcripts from the brain library and the embryo library were then merged using in-house python scripts to create a PacBio transcriptome annotation.

Illumina RNA sequencing

RNA samples from 20 tissue types were collected from 9 16/17 weeks old female J-Line chickens bred at the Edinburgh Poultry Research Centre. The samples from the 9 individuals were pooled for each tissue type and sequenced by Edinburgh Genomics. The Illumina Total RNA Stranded kit was used to generate stranded cDNA fragments. In this stranded RNAseq method, random primers are used for reverse transcription to create a complementary strand to the original RNA template. Deoxyuridine Triphosphate (dUTP) is then incorporated into the original template [42]. Adapters are attached to both ends of the double strand and then the original template is degraded. The adapters provide strand information based on their orientation in the read. The cDNA fragments were then sequenced to produce paired end reads with an average length of 101 base pairs. The average size of cDNA fragments was 190 bases.

Short read transcriptome assembly

Edinburgh Genomics generated 8 fastq file pairs for each tissue. Each tissue had an average of 120,563,969 reads between all 8 fastq files. We checked the quality of data using Fastqc. We then mapped the reads to the Galgal 4 genome assembly using Tophat2 version 2.0.14 with Bowtie2 version 2.2.5. For this we used the parameters to define the inner insert size for each library and the strand orientation (`-library-type fr-firststrand`). We then ran Cufflinks version 2.2.1 to assemble transcripts using default parameters on each library (8 libraries per tissue). No annotations were provided at this step for guiding. We then merged the transcript models from each library using Cuffmerge. This final merged annotation was designated as our J-line derived annotation.

We also ran Cufflinks using force guided on the Ensembl (release 83) annotation, the PacBio annotation, and the J-Line derived annotation. This was performed to acquire FPKM estimates for each transcript model in each annotation set so that we could generate our expression atlases.

Other bioinformatics analyses

Prediction of coding and noncoding transcripts

To classify the PacBio transcripts as either protein coding or noncoding we used the criteria that transcripts without evidence for protein coding potential were labelled as noncoding RNA and transcripts with evidence were labelled as putative protein coding. We used three methods to find evidence for protein coding potential. The first method consisted of using Blastx [22] to find hits between the PacBio transcripts and the Uniprot

Uniref 90 protein database [23]. The second method consisted of using the Coding Potential Calculator (CPC) software [24]. CPC uses six different metrics to determine the coding potential of a transcript. These include using open reading frames (ORFs) and searching for Uniprot protein hits. The third method consisted of using the Coding Potential Assessment Tool (CPAT) [25]. CPAT uses ORF's and hexamer usage bias to produce protein coding probabilities. We used the recommended cut off of 0.3 for CPAT, designating any scores below this as noncoding.

Prediction of nonsense mediated decay products

Transcripts which were first identified as protein coding using our previously defined methodology were used for our NMD prediction. We investigated potential NMD products by identifying coding sequence regions within the PacBio transcript models. Coding sequence regions were identified by first converting the RNA sequences to peptide sequences in all three frames (single stranded data). The longest three ORF's were matched to the Chicken protein sequences from Uniprot using Blastp [22]. The ORF's with the highest match to a chicken protein were used as the representative or if no matches were found, the longest ORF was used. If the representative ORF had a stop codon that was more than 50-bp upstream of the final splice junction, it was labelled as an NMD candidate [31].

Additional file

Additional file 1: PacBio sequencing read lengths (pdf format) (a) Read lengths for chicken brain 1-kb size selection. (b) Read lengths for chicken brain 2-kb size selection. (c) Read lengths for chicken embryo 0.8-kb size selection. (d) Read lengths for chicken embryo 2-kb size selection. (PDF 184 kb)

Abbreviations

AEE: Alternative exon end; AES: Alternative exon start; ASV: Alternative splice variant; CPAT: Coding Potential Assessment Tool; CPC: Coding Potential Calculator; ECC: Exon Cascade Collapse; ICE: Iterative clustering for error correction; lincRNA: Long intergenic noncoding RNA; lncRNA: Long noncoding RNA; ncRNA: Noncoding RNA; NMD: Nonsense mediated decay; PacBio: Pacific Biosciences; STSE: Single transcript single exon; TGEA: Tissue gene expression atlas; TSE: Transcription start exon; TSS: Transcription start site; TSSC: Transcription Start Site Collapse; TTE: Transcription termination exon; TTS: Transcription termination site

Acknowledgements

The authors are thankful to Dr. Jacqueline Smith and Dr. Almas Gheyas for providing writing assistance.

Funding

We acknowledge funding support from the UK's Biotechnology and Biological Sciences Research Council (Institute Strategic Programme grant BB/J004235/1; and 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.

Availability of data and materials

Raw RNA sequencing data was deposited in the European Nucleotide Archive under accession numbers: PRJEB13246 (J-line Chickens Brain Isoseq), PRJEB13248 (ISA Brown Chickens Embryo HH26 Isoseq), PRJEB12891 (J-Line Chicken 21 Tissue RNA Expression Illumina). All databases used (Ensembl and Uniprot) are openly accessible by the public.

Authors' contributions

DWB conceived the idea for the study and selected the samples for sequencing. IRP prepared the samples for sequencing. RIK analysed the long read data. LE and RIK analysed the short read data. ET developed analyses pipelines and provided guidance on usage. RIK, DWB and ALA wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval

All animal work was conducted according to UK Home Office guidelines and approved by the Roslin Institute Animal Welfare and Ethical Review Body. All animals were sourced from the Edinburgh Poultry Research Centre.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK. ²Pacific Biosciences, Menlo Park, CA, USA. ³The University of Queensland, St. Lucia, Canberra, QLD 4072, Australia.

Received: 10 November 2016 Accepted: 6 April 2017

Published online: 24 April 2017

References

- Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, et al. Ensembl 2005. *Nucleic Acids Res* [Internet]. 2005;33:D447–53. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=540092&tool=pmcentrez&rendertype=abstract>. [cited 2015 Oct 23].
- Wang J, Yao A, Hauer-jensen M, Louis M, James W, Markham W, et al. cDNA cloning and sequencing, gene expression, and immunolocalization of thrombomodulin in the Sprague-Dawley rat. *DNA Res*. 1999;6:57–62.
- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10:59–63.
- Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*. 2008;453:1239–43.
- Ensembl gene annotation project (el71) *Gallus gallus* (chicken). 2013;1–13. Available from: http://jul2016.archive.ensembl.org/info/genome/genebuild/2013_04_chicken_genebuild.pdf.
- Steijger T, Abril JF, Engström PG, Kokocinski F, Akerman M, Alioto T, et al. Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* [Internet]. 2013;10:1177–84. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3851240&tool=pmcentrez&rendertype=abstract>, <http://www.ncbi.nlm.nih.gov/pubmed/24185837>.
- Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* [Internet]. 2011;27:2325–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21697122>. [cited 2014 Jul 14].
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* [Internet]. 2010;28:511–5. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3146043&tool=pmcentrez&rendertype=abstract>. Nature Publishing Group; [cited 2014 Jul 9].

9. Martin JA, Wang Z. Next-generation transcriptome assembly. *Nat Rev Genet* [Internet]. 2011;12:671–82. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21897427>. Nature Publishing Group; [cited 2014 Jul 11].
10. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. *Genome Biol* [Internet]. 2013;14:405. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23822731>.
11. Thomas S, Underwood JG, Tseng E, Holloway AK. Long-read sequencing of chicken transcripts and identification of new transcript isoforms. *PLoS One* [Internet]. 2014;9:e94650. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24736250>. [cited 2014 Apr 18].
12. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* [Internet]. 2013;31:1009–14. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24108091>. [cited 2014 May 28].
13. 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/doi/10.1038/ncomms11706>. Nature Publishing Group.
14. 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/doi/10.1038/ncomms11708>. Nature Publishing Group.
15. 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;10:1–15. Available from: <http://dx.doi.org/10.1371/journal.pone.0132628>.
16. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* [Internet]. 2005;21:1859–75. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15728110>. [cited 2014 Apr 30].
17. Haberer V, Forrest AR, Hayashizaki Y, Carninci P, Lenhard B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res* [Internet]. 2015;1–11. Available from: <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv054>. [cited 2015 Feb 6].
18. Chang H, Lim J, Ha M, Kim VN. TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications. *Mol Cell* [Internet]. 2014;53:1044–52. Available from: <http://dx.doi.org/10.1016/j.molcel.2014.02.007>. Elsevier Inc.
19. Burset M, Seledtsov IA, Solov'yev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res* [Internet]. 2000;28:4364–75. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=113136&tool=pmcentrez&rendertype=abstract>.
20. Wang ET, Sandberg R, Luo S, Khrebukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* [Internet]. 2008;456:470–6. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2593745&tool=pmcentrez&rendertype=abstract>.
21. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. 2012.
22. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST plus: architecture and applications. *BMC Bioinformatics*. 2009;10:1.
23. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* [Internet]. 2014;43:D204–12. Available from: <http://nar.oxfordjournals.org/content/43/D1/D204>. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4384041&tool=pmcentrez&rendertype=abstract>.
24. Kong L, Zhang Y, Ye Z-Q, Liu X-Q, Zhao S-Q, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* [Internet]. 2007;35:W345–9. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1933232&tool=pmcentrez&rendertype=abstract>. [cited 2014 Jul 9].
25. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* [Internet]. 2013;41:e74. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3616698&tool=pmcentrez&rendertype=abstract>. [cited 2015 Feb 19].
26. Maquat LE. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol* [Internet]. 2004;5:89–99. Available from: <http://www.nature.com/doi/10.1038/nrm1310>.
27. Brogna S, Wen J. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat Struct Mol Biol* [Internet]. 2009;16:107–13. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19190664>. [cited 2014 Jul 17].
28. Isken O, Maquat LE. The multiple lives of NMD factors: balancing roles in gene and genome regulation. *Nat Rev Genet* [Internet]. 2008;9:699–712. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3711694&tool=pmcentrez&rendertype=abstract>. [cited 2015 Aug 18].
29. Zhang J, Sun X, Qian Y, Duca JPLA, Maquat LE. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. *Mol Cell Biol*. 1998;18:5272–83.
30. Thermann R, Neu-yilik G, Deters A, Frede U, Wehr K, Hagemeier C, et al. Binary specification of nonsense codons by splicing and cytoplasmic translation. *EMBO J*. 1998;17:3484–94.
31. Nagy E, Maquat LE. A rule for termination-codon position within intron-containing genes: When nonsense affects RNA abundance. *Trends Biochem Sci*. 1998;23:198–9.
32. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*. 2012;22:1775–89.
33. 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.
34. Gardner LB. Nonsense-mediated RNA decay regulation by cellular stress: implications for tumorigenesis. *Mol Cancer Res* [Internet]. 2010;8:295–308. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2841721&tool=pmcentrez&rendertype=abstract>. [cited 2015 Sep 2].
35. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, et al. Antisense transcription in the mammalian transcriptome. *Science*. 2005;309:1564–6.
36. Kiyosawa H, Yamanaka I, Osato N, Kondo S, Hayashizaki Y, Ger R. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res*. 2003;13:1324–34.
37. Faghihi MA, Wahlestedt C. Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol* [Internet]. 2009;10:637–43. Available from: <http://dx.doi.org/10.1038/nrm2738>. Nature Publishing Group.
38. Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Lassmann T, et al. A promoter-level mammalian expression atlas. *Nature* [Internet]. 2014;507:462–70. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24670764>. Nature Publishing Group; [cited 2014 Jul 9].
39. Curwen V, Eyres E, Andrews TD, Clarke L, Mongin E, Searle SMJ, et al. The Ensembl automatic gene annotation system. *Genome Res* [Internet]. 2004;14:942–50. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=479124&tool=pmcentrez&rendertype=abstract>. [cited 2014 Sep 29].
40. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, et al. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet*. 2012;8:e1002841. <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002841>.
41. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–7.
42. Borodina T, Adjaye J, Sultan M. A strand-specific library preparation protocol for RNA sequencing. *Methods Enzymol* [Internet]. 2011. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21943893>. 1st ed. Elsevier Inc.; [cited 2015 Aug 30].

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit



2.3 Discussion

The major results from this work are based on sequencing pipeline performance and discoveries within the chicken transcriptome. Normalization of cDNA libraries prior to sequencing seems to significantly improve the ability to pick up lower abundance RNA. However, given the lower throughput of the PacBio RSII system that was the long-read sequencing platform available at the time, many transcript models were represented by only a single read. This makes distinguishing real transcript sequences from noise more difficult. The 5' cap selection of cDNA libraries seems to reduce the relative number of reads coming from degraded RNA. This in turn increases overall sequencing efficiency by reducing the number of reads spent on degraded RNA.

The long read RNA sequencing performed in this study represents one of the earliest attempts at using the PacBio platform for transcriptome annotation. As such, the bioinformatic landscape was sparse with very little understanding at the start of the study as to what were the appropriate types of data processing to use. Given that very few people had even seen this type of data let alone worked with it, the start of the project was very much focused on getting a better understanding of what we were seeing in the data. The comparatively low read depth from this study meant that many tools and methods that are available today are not very applicable for processing this data.

I consider this piece of work to be an introduction of long read RNA sequencing to the field with respect to usage in whole transcriptome annotation. There have been many developments in this area since the publishing of this article. Oxford Nanopore Technologies cDNA sequencing has become available, offering a viable alternative to Iso-Seq sequencing[122]. Direct RNA sequencing has emerged as a method that allows for full length transcript sequencing with detection of RNA modifications[123]. Single cell long read RNA sequencing has gained in popularity[124]. Library construction methods that target specific genes have been

developed[125]. Long read RNA sequencing has also become an important tool for understanding cancer[126].

After this paper was published, other studies used long read RNA sequencing for chicken transcriptome discovery. This includes a study using ONT cDNA sequencing for 19 chicken tissue types[127]. Despite the wider variety of tissue types and higher number of reads generated, the study identified a lower number of unique transcripts. This may be due to differences in sequencing technologies or bioinformatic processing, however it is likely that the library preparation played a large role. More specifically, they did not use normalization or 5' cap selection methods which may have reduced their sequencing efficiency for the discovery of novel transcripts.

Another study used PacBio Iso-seq sequencing on chicken embryo to discover novel long non-coding RNA[128]. They reported a larger number of lncRNA, however their predictions were based on both their long read data as well as their short read data. Therefore it is very likely that their high predictions were due to predictions from short reads which are less reliable than long read predictions for lncRNA.

Even with all these improvements in the field and newer studies using newer technologies or greater read depth, the sample processing techniques from this paper remain advanced and under utilized. The bioinformatic ideas have by and large remained relevant. However, as long read RNA sequencing improved in throughput, read length, and accuracy, many of the challenges encountered in this study have become lesser issues.

Using these novel sequencing pipelines, I was able to identify thousands of potential novel genes. These include thousands of potential lncRNA. I confirmed that the lncRNA models seemed to show a significant reduction in sequence conservation across other avian species as compared to protein coding genes. There also seems to be a large number of mono-exonic and

sense-exonic lncRNA. However, these transcript models are the most difficult to differentiate from noise so further work needs to be done to validate these.

Chapter 3: Paper - Illuminating the dark side of the human transcriptome with long read transcript sequencing

3.1 Introduction and aims

In this chapter I explore the effects that different long read RNA sequencing analysis pipelines can have on annotation results. I explain the TAMA software package that I developed and demonstrate its performance on an Iso-Seq dataset that was produced by PacBio.

In this study, all work was performed by myself except as further specified: Samples were prepared and sequenced by PacBio.

3.2 Research Paper


This research was published as “Illuminating the dark side of the human transcriptome with long read transcript sequencing” in BMC Genomics [129]. BMC Genomics does not require authors to obtain permission to include these papers in their thesis, provided the original work is properly cited.

RESEARCH ARTICLE

Open Access



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

Richard I. Kuo^{1*} , Yuanyuan Cheng^{2,3}, Runxuan Zhang⁴, John W. S. Brown^{5,6}, Jacqueline Smith¹, Alan L. Archibald¹ and David W. Burt^{1,2}

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 additional evidence for rare transcripts and genes such as mono-exonic and non-coding genes that were previously either undetectable or impossible to differentiate from sequencing noise.

Results: We developed the Transcriptome Annotation by Modular Algorithms (TAMA) software to leverage the power of long read transcript sequencing and address the issues with current data processing pipelines. TAMA achieved high sensitivity and precision for gene and transcript model predictions in both reference guided and unguided approaches in our benchmark tests using simulated Pacific Biosciences (PacBio) and Nanopore sequencing data and real PacBio datasets. By analyzing PacBio Sequel II Iso-Seq sequencing data of the Universal Human Reference RNA (UHRR) using TAMA and other commonly used tools, we found that the convention of using alignment identity to measure error correction performance does not reflect actual gain in accuracy of predicted transcript models. In addition, inter-read error correction can cause major changes to read mapping, resulting in potentially over 6 K erroneous gene model predictions in the Iso-Seq based human genome annotation. Using TAMA's genome assembly based error correction and gene feature evidence, we predicted 2566 putative novel non-coding genes and 1557 putative novel protein coding gene models.

Conclusions: Long read transcript sequencing data has the power to identify novel genes within the highly annotated human genome. The use of parameter tuning and extensive output information of the TAMA software package allows for in depth exploration of eukaryotic transcriptomes. We have found long read data based evidence for thousands of unannotated genes within the human genome. More development in sequencing library preparation and data processing are required for differentiating sequencing noise from real genes in long read RNA sequencing data.

Keywords: Human, Transcriptome, Long read RNA sequencing, Iso-Seq, TAMA, Annotation, Pacbio, Nanopore, Gene models, Bioinformatics

* Correspondence: richard.kuo@roslin.ed.ac.uk

¹The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian EH25 9RG, UK

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

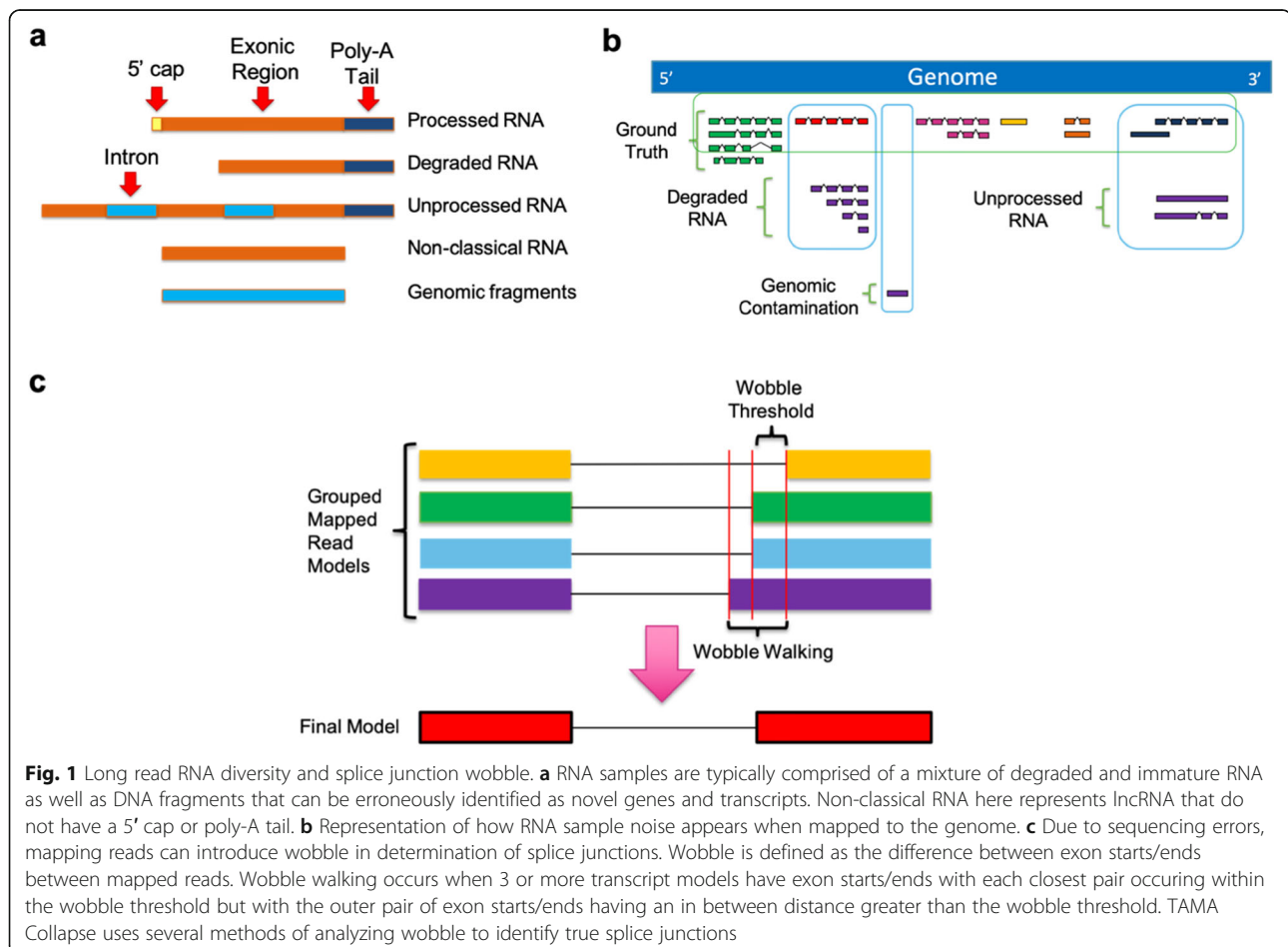
Background

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 fails to provide coverage for rare/unstable transcripts, while short read RNA sequencing (RNA-seq) present computational challenges in accurate transcript model reconstruction [4–6]. 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 genes from sequencing noise [5]. While there have been many studies using long read transcript sequencing for transcriptome discovery [7–11], their sensitivity may have been compromised by the use of orthogonal verification/filtering. 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 [6].

The use of inter-read error correction in previous studies by either hybrid approaches (aligning short reads to long reads) or long read methods (aligning long reads into clusters) could also cause issues with both reducing gene detection sensitivity and producing erroneous gene models. Long read inter-read error correction methods such as PacBio's Cluster/Polish method [7] filter out any reads that do not cluster with at least one other read. Due to the lower read depth of long read sequencing methods (relative to short read sequencing), this results in the removal of many low expressed genes and transcripts. Inter-read error correction methods can also produce erroneous hybrid sequences since long reads



with higher error rates have an increased probability of grouping with other high error rate reads from different transcripts either from the same or paralogous genes. This type of error occurs when the alignment of reads is compromised by regions of high error density. While this effect could be reduced by requiring high alignment scores for clustering reads, this would also decrease the intended effect of rescuing low quality reads.

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 uses long read transcript data and high-quality reference genome assemblies to produce accurate and informative transcript models. TAMA is designed to improve transcript model prediction accuracy and increase transcriptome discovery with transparent and traceable steps. Evidence, including raw reads, read counts, local sequencing characteristics (e.g. mismatches, internal poly-A sequencing) supporting or compromising each transcript model is recorded and presented. This makes TAMA useful for situations where additional types of data, such as public annotations or short read RNA-seq, are not available [8]. In addition, by not relying on orthogonal information and having transparent and traceable steps, TAMA also provides a more agnostic approach to transcriptome annotation which can reveal problems with prior assumptions from previous annotation efforts.

We report the use of TAMA to analyze the Universal Human Reference RNA (UHRR) Sequel II Iso-Seq data released to the public by Pacific Biosciences (PacBio). This dataset represents the combination of the highest read depth for long read sequencing on a single human RNA sample with the highest long read accuracy. As such, the challenges of analyzing this dataset are applicable to all long read transcriptome datasets. We compared different long read based transcriptome assembly methods to identify corresponding benefits and issues. Our analyses indicate that long read transcript sequence data together with appropriate analysis tools has the potential to reveal yet further complexity in eukaryote transcriptomes.

Results

TAMA – Transcriptome annotation by modular algorithms

TAMA is comprised of modular tools with transparent algorithms, precise parameter control, and traceable outputs to allow users to analyze, interpret, and diagnose the resulting transcript models. The main analysis functions consist of two modules: TAMA Collapse and TAMA Merge.

TAMA Collapse uses mapped reads and a reference genome assembly to create a transcriptome annotation. TAMA Collapse uses four main methods for identifying

true splice junctions: alignment quality filtration, local density error filtration (LDE), splice junction ranking, and splice junction coverage. All of these methods can be tuned by the user. First, alignment quality filtration is applied by assessing the alignment length coverage and alignment identity of each mapped read with respect to the reference genome. Reads below the user defined thresholds are discarded. The reads passing this first step are then examined via the LDE algorithm for the number of mismatches flanking each predicted splice junction. Errors around splice junctions exacerbate mis-mapping and cause the prediction of false splice junctions. This assessment removes reads with high error density within a specified base pair distance from each splice junction. The remaining reads are then grouped based on exon-intron structure allowing for user defined differences (called wobble in the TAMA nomenclature) in exon starts and ends measured in base pairs (Fig. 1c). The predicted splice junctions for the grouped reads are then ranked based on the flanking mismatch profiles and coverage. The highest ranked splice junctions are then used in the final transcript model. A large wobble threshold can help remove false positive predictions for splice junctions but may remove real splice junctions within the wobble length. Thus the LDE algorithm and splice junction ranking allows for smaller wobble lengths while also reducing false splice junction predictions.

In addition to rigorously identifying splice junctions, TAMA Collapse also allows the incorporation of the confidence of transcript starting sites by running the program in a capped or non-capped mode. For example, for 5' captured RNAs, the capped mode will allow the transcripts with alternative transcript starting sites to be retained; while for non 5' captured RNAs, the non-capped mode removes transcript models which appear to be 5' degraded. The capped mode, requires grouped mapped reads to have the same number of exons and the same exon-intron structure. The non-capped mode is similar to the capped mode but allows for grouped reads to have differences in the number of exons on the 5' end reflecting reads derived from RNAs with degradation from the 5' end. Thus, all predicted splice junctions for the shorter mapped read model and the 3' end would have to match those of the longer model. These two methods of grouping are described in a previous study where they were referred to as Transcription Start Site Collapse (equivalent to capped mode) and Exon Cascade Collapse (equivalent to non-capped mode) [4].

In addition to the transcriptome assembly, TAMA Collapse also outputs detailed information showing read mapping quality, collapsed read groups, predicted sequence variation, and transcript models with 3' genomic poly-A (genomic contamination or truncated transcript). These outputs are intended to provide users with a full

understanding of the behavior of TAMA Collapse and thus allow users to trace, diagnose, and improve their transcriptome assemblies.

TAMA Merge combines transcript models by examining exon-intron structures of transcript models to create a non-redundant set of genes and transcripts. TAMA Merge can be used on a single input transcriptome annotation to remove redundancy or can be used on multiple transcriptome annotations to create a unified annotation. TAMA Merge also produces output files that can be used to understand the differences between the input annotations. TAMA Merge uses the same collapsing mode algorithms from TAMA Collapse. One unique feature of TAMA Merge is the ability to merge transcript assemblies by assigning different collapsing modes and transcript model feature priorities between different annotations. For example, when using TAMA Merge to combine a long read sequencing derived annotation to a reference annotation, the reference annotation can be given priority for transcription start/end sites and splice junctions. The user created annotation can also be set to the non-capped mode to allow user produced models to collapse with 5' longer reference models. The output files from TAMA Merge include detailed reports on how merging was done. These report files show which input annotations supported each of the final transcript and gene models as well as the amount of wobble that occurred at each exon start and end between merged models.

Along with TAMA Collapse and TAMA Merge, the TAMA toolkit contains many other tools that either apply additional filters or add information. Other TAMA tools used in this study are explained in further detail in the [Methods](#) section. A more detailed description of how TAMA works can be found here: github.com/GenomeRIK/tama/wiki/.

Benchmarking TAMA and related software

We benchmarked the long read based transcriptome assembly of TAMA, Stringtie2 [9], TALON [10], and Cupcake [7] using three different datasets: simulated PacBio data, simulated Nanopore data, and PacBio Sequel II Iso-Seq data from Lexogen's Spike-in RNA Variant (SIRV) control mix. The simulated PacBio and Nanopore reads were produced in a previous study [11] using PBSIM [12] and were also used for benchmarking in the Stringtie2 study [9]. The simulated datasets were based on the annotations of chromosome 19 of the human reference annotation. Details of the simulated and human datasets can be found in the supplementary files (Table S1). Using these simulated datasets, the Stringtie2 study showed that Stringtie2 outperformed both FLAIR [13] and Traphlor [14]. We used the same method of assessment as was used in the Stringtie2 study. While these

simulated datasets are useful due to having a ground truth, they are not entirely accurate in their representation of long read sequencing data. In particular, the simulated reads were created by fragmenting transcript models at random which is not realistic since the fragmentation of transcripts is non-random and influenced by sequence characteristics and sample processing methods. The simulated PacBio dataset represents reads equivalent to PacBio Full Length Non-Chimeric (FLNC) reads. This means that they assume Circular Consensus Sequence (CCS) intra-read correction was performed and that adapters and poly-A tails were removed. The simulated Nanopore dataset is equivalent to Nanopore reads after removing poly-A tail and adapter sequences. Since PacBio's Iso-Seq software (Cupcake) requires specific PacBio generated metadata that these simulated datasets do not contain, we could not benchmark PacBio's Cupcake software on these datasets. This means that we could not use PacBio's Cluster/Polish inter-read error correction on these datasets. Thus, these simulated datasets can only be used to assess the effect of random errors in long reads on the performance of mapping tools and transcriptome assemblies tools.

To address the issues with simulated datasets, we also used reads from the Lexogen SIRV spike-in from the PacBio UHRR Sequel II Iso-Seq dataset. The Lexogen SIRV control mix contains synthesized RNA molecules representing 7 expressed loci (18 genes when strand is accounted for) with 69 unique transcripts. The ground truth in this dataset is provided by Lexogen in the form of expected gene models based on their synthetic genome. However, it is possible that not all RNA from the SIRV dataset were sequenced and/or there are other RNA in the SIRV sample which are not represented in the annotation file provided on the Lexogen website. This may explain the lower precision of all unguided pipelines for the SIRV dataset (<68% precision for all unguided approaches).

We used GffCompare [15] to calculate the sensitivity and precision for each pipeline. Sensitivity is defined as the number of correct transcript models in the predicted annotation divided by all the transcript models used for simulation. Precision is defined as the number of correct transcript models in the predicted annotation divided by the number of all predicted transcript models. These scores can be calculated at either the transcript or gene loci level. These definitions are from the GffCompare software. This method of calculation is identical to the method used in the Stringtie2 study [9]. Since TAMA, Stringtie2, and TALON can be run either with an unguided approach or a reference annotation guided approach, we tested both methods for each of these tools. Since TAMA is designed for parameter tuning, we applied two parameter sets for the unguided TAMA

pipelines which we refer to as TAMA Low and TAMA High. TAMA Low uses parameters to maximize genic loci sensitivity at the cost of transcript model precision while TAMA High uses more stringent parameters to remove erroneous transcript models. The parameter selection for TAMA High and TAMA Low differs between the synthetic datasets and the PacBio Sequel II Iso-Seq data (SIRV and UHRR) since the synthetic datasets have higher error rates. TAMA High and TAMA Low parameter selection is described in more detail in the [Methods](#) section. Briefly, the TAMA High pipeline uses a more stringent LDE setting (fewer mismatches surrounding splice junctions), and requires read support from both SMRT Cells (in the PacBio Sequel II Iso-Seq data) while TAMA Low has lower stringency settings for LDE and requires support from only a single read. The TAMA High requirement of read support from both SMRT cells can be viewed as a modified form of the method that the Cluster/Polish step uses to filter out erroneous transcript models (removing all reads that do not cluster). However, the TAMA High approach can provide more sensitivity since it allows for greater variance on the 5' end of the transcript models to account for low expressed genes which may only be represented by a 5' truncated model in one of the SMRT cells (where the predicted 5' complete model was picked up in the other SMRT cell). This method of filtration can also provide greater precision since requiring read support across sequencing runs can help reduce artifacts caused by technical batch effects. This algorithm can be adjusted where only a single SMRT cell or sequencing run was performed by only requiring multiple read support for each transcript model. This would still provide greater sensitivity than the Cluster/Polish method due to the greater allowance in 5' variability. The TAMA Guided pipeline matches the transcript models from the long read data to the input reference annotation and adopts the splice junction predictions from the reference annotation. It discards any models not matching the reference annotation using the TAMA Merge algorithm. See [Methods](#) section for description of TAMA Merge and pipeline parameter selection.

For both the PacBio and Nanopore simulated datasets, guided approaches achieved better sensitivity and precision as compared to unguided approaches (Fig. 2). The TAMA Guided approach had the highest precision across all datasets with slightly less sensitivity as compared to the Stringtie2 Guided approach for the simulated datasets. In the SIRV dataset, the TALON Guided method achieved a slightly higher sensitivity score as compared to TAMA Guided. The higher sensitivity score for TALON Guided was due to the inclusion of one more transcript model as compared to TAMA Guided. When we inspected this transcript model found only in the TALON Guided assembly, we found that it

did not match the supporting reads (Fig. 2f). The reads used to support the TALON Guided prediction of that particular transcript model have a long 3' extension as compared to the predicted transcript model. This extension is present in other transcript models in the SIRV annotation and it appears that these reads likely originated from 5' truncated/degraded RNA from those transcripts. This raises the question of why these reads were assigned to the transcript model and how this might affect unguided TALON.

The overall better performance of guided approaches is to be expected because guided approaches essentially fit the transcript models to an annotation which has high similarity to the assessment annotation. However, guided approaches are not as useful for transcriptome discovery since they only confirm already known gene/transcript models. Among all the unguided methods, TAMA Low achieves the best sensitivity for the gene loci level while TAMA High achieves the highest precision and sensitivity at the transcript level compared to the non-TAMA approaches. The SIRV gene loci comparison was not included since the SIRV transcriptome is comprised of only 18 gene loci across 7 scaffolds. All methods had perfect sensitivity and precision at the gene loci level for the SIRV dataset.

Effect of inter-read error correction on gene model discovery

We processed the UHRR Iso-Seq data using four different pipelines to understand the effect of pre-mapping inter-read error correction on gene discovery and model prediction accuracy (Fig. 3a). The UHRR Iso-Seq dataset was comprised of two separate Sequel II runs using the 8M SMRT Cells. There were 4,461,529 and 4,473,633 CCS reads generated by the two SMRT Cells which resulted in 3,504,905 and 3,447,471 FLNC reads, respectively. A plot of FLNC read lengths can be found in the supplementary files (Figure S1). All four pipelines use TAMA tools since the TAMA High pipeline has the highest combination of sensitivity and precision compared to all other non-guided methods in the benchmarking tests and the TAMA Low pipeline has the highest sensitivity. We compared two pipelines without inter-read error correction (TAMA Low and TAMA High pipelines), one pipeline using long read inter-read error correction (Polish Pipeline), and one pipeline using hybrid inter-read error correction (Lordec Pipeline). The Polish pipeline, uses inter-read error correction (in the form of clustering long reads and using the alignment to polish the sequences prior to mapping) along with TAMA Collapse using the same parameters as the TAMA Low pipeline. The Lordec pipeline, uses LoRDEC [16] inter-read error correction (aligning short read RNA-seq data to long reads prior to mapping) with

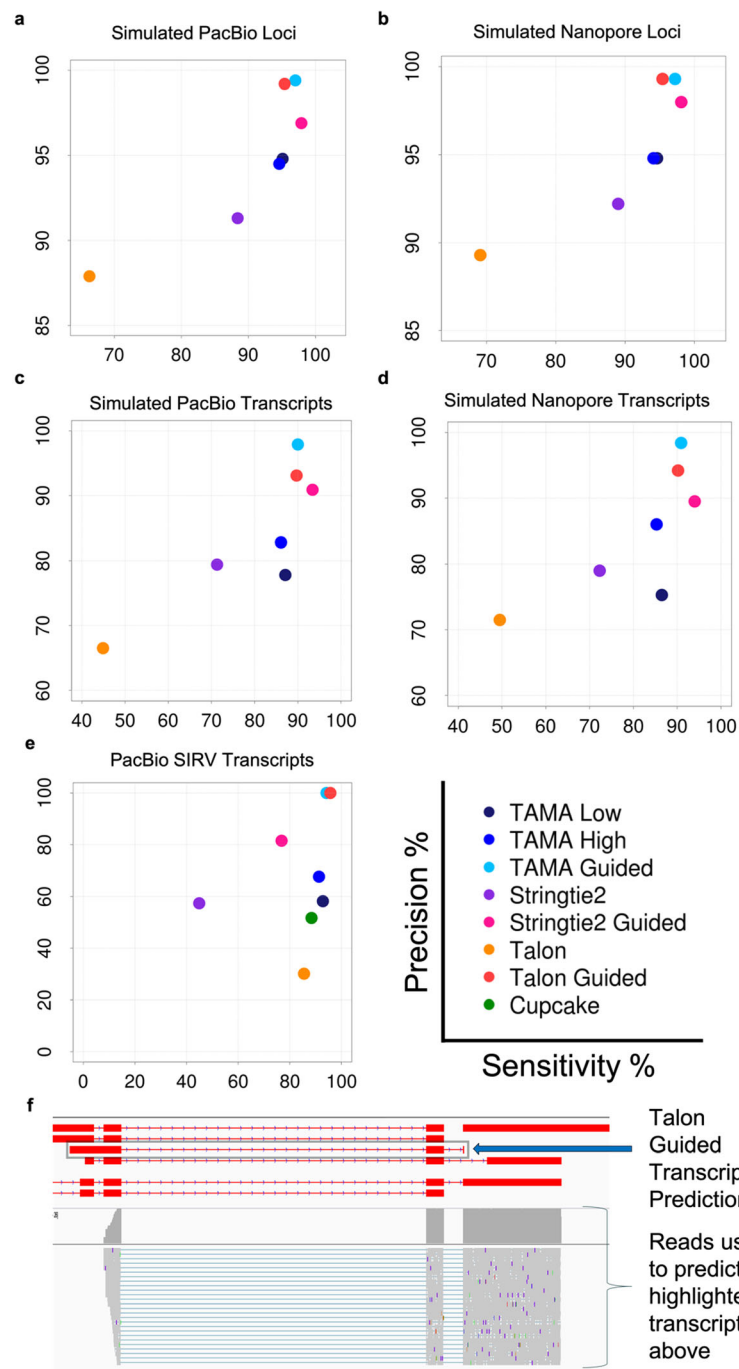


Fig. 2 Long transcript assembly benchmarking. Sensitivity and precision of guided and unguided long read transcriptome assembly methods. **a** Gene loci level for simulated PacBio reads. **b** Gene loci level for simulated Nanopore reads. **c** Transcript level for simulated PacBio reads. **d** Transcript level for simulated Nanopore reads. **e** Transcript level for PacBio Sequel II Iso-Seq SIRV reads. **f** Example of erroneous transcript prediction by guided Talon where supporting reads are from another transcript model. These supporting reads are from 5' degraded RNA resulting in the confusion

TAMA Collapse (same settings as TAMA Low). For the Lordec pipeline we used short read RNA-seq data from the UHRR but from another study [17].

The TAMA Low and Lordec pipelines produced the most predicted gene and transcript models with more than 160 K genes and 750 K transcripts (Table 1). These

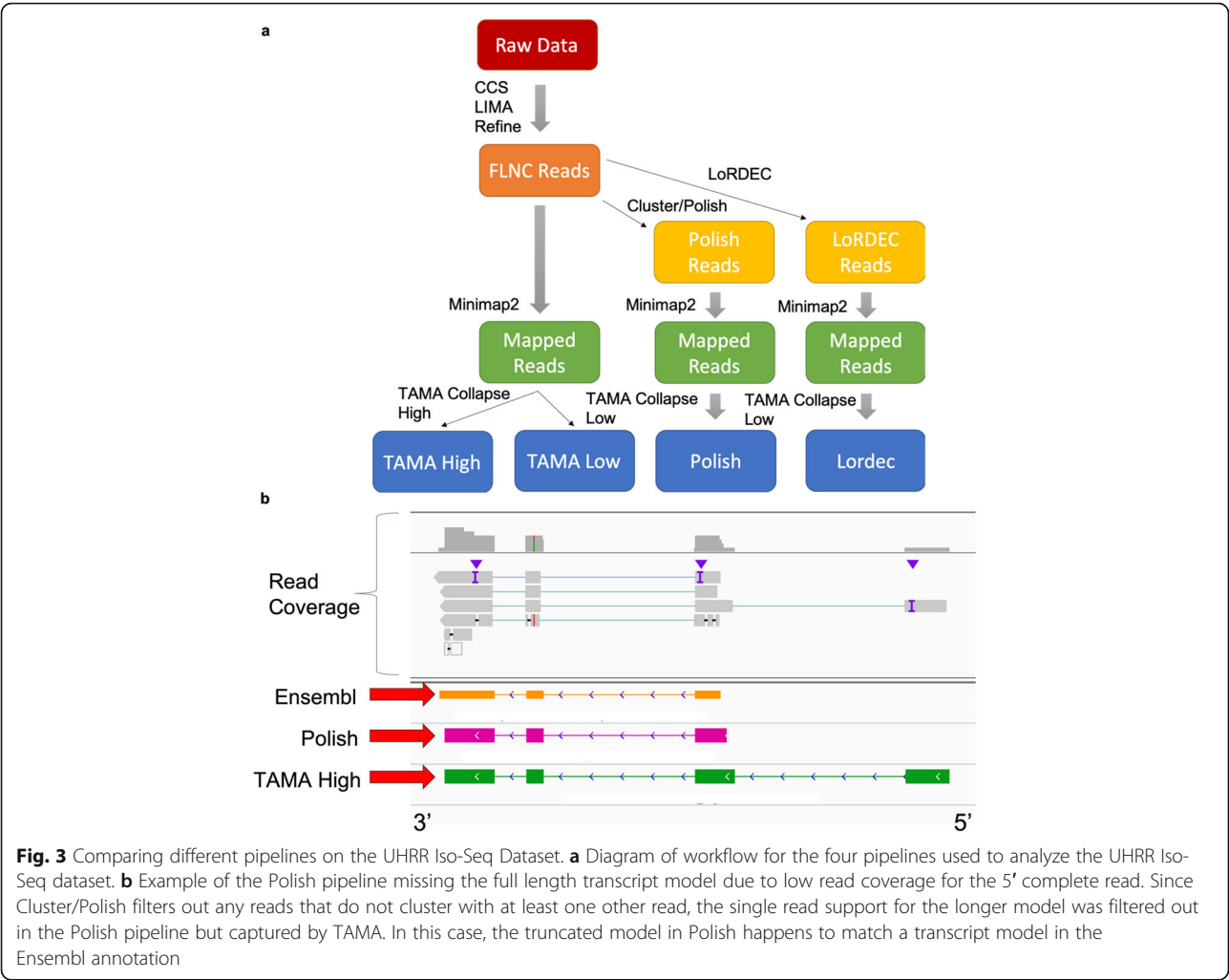


Table 1 Pipeline comparison

Match type	Polish	Lordec	TAMA low	TAMA high
Total Genes	25,731	166,766	168,328	38,743
Total Transcripts	126,288	753,756	752,996	135,218
Ensembl Loci Overlap	19,348	30,835	30,947	21,284
Ensembl Transcript Matches	17,948	24,660	24,691	15,854
Predicted Novel Gene Loci	8519	139,769	141,097	23,302
Predicted Novel Transcripts	106,243	724,316	723,759	118,148

Comparison of gene and transcript numbers across pipelines broken down into different categories. Ensembl loci overlap refers to the number of Ensembl v94 annotation gene models that are overlapped on the same strand by gene models from each Iso-Seq annotation. Transcript matches refer to Ensembl v94 transcript models with identical exon-intron structures as transcript models in each Iso-Seq annotation. The Ensembl v94 human annotation consists of 58,735 gene loci and 206,601 unique transcript models. In some cases, multiple Ensembl gene loci are overlapped by a single Iso-Seq gene locus leading to the differences between matching loci and predicted novel loci

extremely high numbers are likely due to issues with the use of reads with high error rates and reads originating from transcriptional noise. The Polish pipeline produced the fewest number of genes and transcript models (Table 1) while the TAMA High pipeline had over 1.5 times the number predicted genes but with a similar number of predicted transcripts.

Estimating gene model detection accuracy

While there is no ground truth for the human transcriptome, we used the Ensembl v94 (Release 94, October 2018) human genome reference annotation [18] as a reference to understand how our results compare to current annotations. We identified the number of gene loci and transcript models from the Ensembl annotation with representation from each pipeline. The TAMA Low and Lordec pipelines had the highest number of matches for both gene loci and transcript models indicating high sensitivity. However, given the high total numbers of genes and transcripts, the annotations from these pipelines likely contain many erroneous gene and transcript models. The TAMA High pipeline had more gene loci matches but slightly fewer transcript matches compared to the Polish pipeline. This means that there were more transcripts per gene in the Polish pipeline annotation (4.9:1) versus the TAMA High annotation (3.5:1). The higher ratio of transcripts to genes in the Polish pipeline, as compared to the TAMA High pipeline, suggests that either TAMA High is filtering out many real alternative transcripts or that Cluster/Polish is somehow predicting more erroneous alternative transcript models.

When we investigated the reason for the higher number of transcript model matches in the Polish annotation, we discovered that in some cases the Polish transcript models matched the models in the Ensembl annotation due the removal of reads (by the Cluster/Polish step) which supported 5' longer transcript models (Fig. 3b). In these cases, the mapped reads showed 5' extended transcript models with additional 5' exons along with 5' shorter models that may have originated from 5' degraded RNA molecules. However, since the longer models had lower read coverage, the Polish pipeline removed them from the transcriptome assembly leaving only the shorter models that sometimes matched models in the Ensembl annotation. This tendency toward producing truncated transcript models could explain the expansion of alternative transcript predictions in the Polish

pipeline. While it could be argued that these shorter models are real since they are represented in the Ensembl annotation, it is also possible that these RNA are typically rapidly degraded and thus full length representations have not been identified in the Ensembl annotation due to a lack of coverage from the supporting data used by the Ensembl pipelines.

Assessing RNA degradation from Iso-Seq data

To gain a better understanding of the effect that RNA degradation may have on long read based annotations, we analyzed the transcript models which had matching 3' exon-intron structure between the TAMA High (135,218 transcripts), Polish (126,288 transcripts), and Ensembl v94 (206,601 transcripts) annotations to see which annotation had longer 5' representation (Table 2). When comparing the TAMA High annotation to the Polish annotation, there were 67,480 transcript models with matching 3' exon-intron structure. Out of those 3' matching transcript models, 56,198 (83.2%) showed the TAMA High models as having the longer 5' representation with 3357 models (5%) having additional 5' exons. This indicates that the Polish pipeline may be producing a large number of 5' incomplete transcript models. While the TAMA High and Polish annotations had similar numbers of transcript models, roughly half of those models in each annotation did not have matches between the annotations. This may be due to differences in splice junction calls between the two pipelines which is referred to in this text as splice junction wobble.

When we compared the TAMA High annotation to the Ensembl annotation using the same method, we found 23,542 3' exon-intron structure matching transcript models. Out of those matching models, 15,230 (64.7%) showed the TAMA High models as having the longer 5' representation with 3521 models (15%) having additional 5' exons. Comparing the Polish pipeline annotation to the Ensembl annotation using the same method, we found 26,186 3' exon-intron structure matching transcript models. Out of those matching models, 15,496 (59.2%) showed the Polish models as having the longer 5' representation. This could indicate that over three thousand Ensembl transcript models have incomplete 5' ends with missing 5' exons or that at least these represent novel alternative transcripts for these genes. Even though roughly half of the transcript models (67,480) from the TAMA High and Polish

Table 2 Comparing 5' completeness of transcript models between annotations

Match comparison	TAMA high longer	Polish longer	Ensembl longer	Total matches
TAMA High - Polish	56,198	11,282	–	67,480
TAMA High - Ensembl	15,230	–	8312	23,542
Polish - Ensembl	–	15,496	10,690	26,186

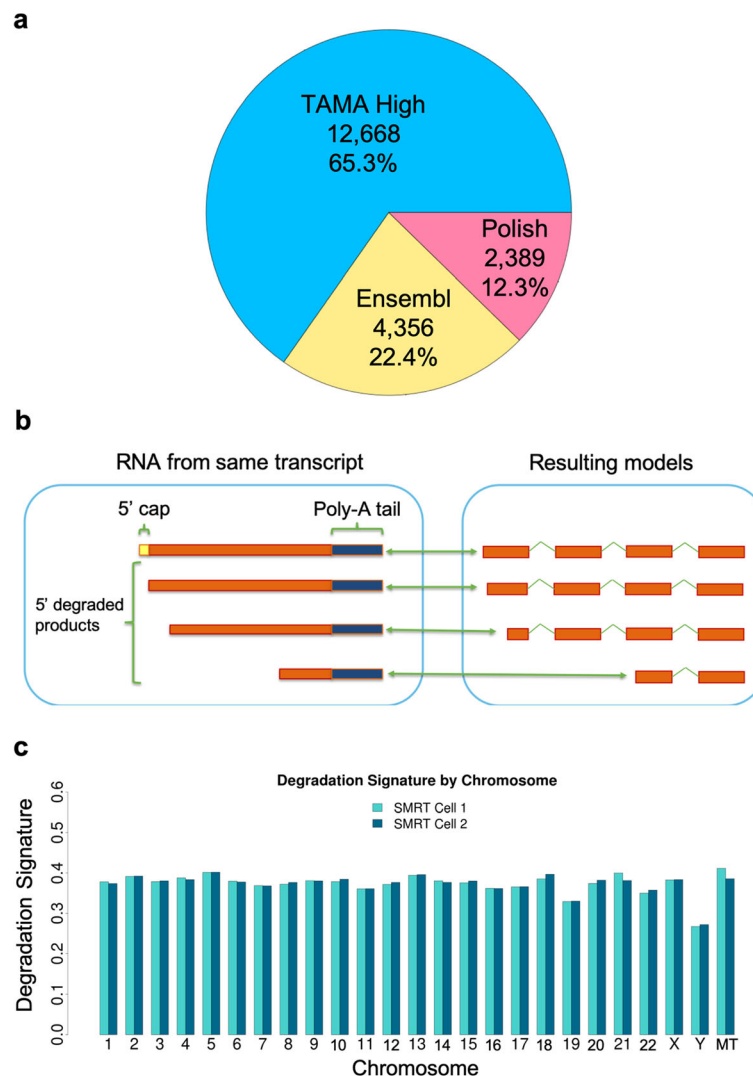


Fig. 4 Degradation signature analysis. **a** Pie chart of the 3' transcript level intersection between the TAMA High, Ensembl, and Polish annotations with the sections representing the number of 5' extended transcript models from each annotation. **b** Diagram of degraded RNA representation with respect to a genome assembly. The reduced 5' coverage results in 5' variability in mapped reads. **c** Degradation signature by chromosome per SMRT Cell run

pipelines had matches between the two pipelines, less than half (23,542 for TAMA High and 26,186 for Polish) of those transcript models also matched the Ensembl annotation. This suggests that the models matching between the TAMA High and Polish pipelines but not found in the Ensembl annotation may represent novel alternative transcript models. Alternatively, they may indicate a type of systemic error in the transcript model prediction pipelines.

We then compared the intersection between all three annotations and identified 19,413 transcripts with common 3' regions. Of these transcripts, TAMA High had the longest transcripts in 65.3% of the matches, Ensembl in 22.4%, and Polish in 12.3% (Fig. 4a). Although the Polish pipeline annotation had more 3' matching

transcript models with the Ensembl annotation in the two way comparison, the number of 5' longer transcripts were similar to the TAMA High annotation suggesting that the increase in matches came from Polish pipeline models which were shorter on the 5' end as compared to the matching Ensembl transcript models. While the 5' shorter transcript models from the Polish pipeline may be accurate, these results demonstrate that the use of transcript model matching for assessing pipeline performance (as is used in GffCompare) can be affected by false positives from 5' incomplete models where these models happen to match the reference annotation. Thus we suggest in depth evaluation of transcript models for a more accurate understanding of pipeline performance.

A method for estimating RNA degradation from Iso-Seq data

To measure the relative amount of reads originating from 5' degraded RNA, we developed a metric called the "Degradation Signature" (DegSig) which evaluates the amount of 5' exon variability in transcript models (Fig. 4b). The DegSig metric is calculated using the outputs from TAMA Collapse runs and inputting them into the TAMA Degradation Signature tool. The value of DegSig is given as a percentage which represents the proportion of reads derived from 5' degraded RNA (see [Methods](#) for formula). It is important to note that DegSig only provides an estimate of 5' degradation with the caveat that bona fide alternative transcription start sites and incomplete first strand synthesis in the preparation of the cDNA library can also produce 5' exon variability which can mimic 5' degradation. To test our DegSig metric we applied it to two Iso-Seq datasets from Chicken brain RNA. One dataset was produced from TeloPrime [19] 5' cap selected RNA and the other was produced without 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. However, there is no ground truth in any species for the actual amount of 5' shorter models with the same 3' exon-intron structure as longer models, thus DegSig is only a rough gauge of the proportion of models which may be from degraded RNA.

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. 4c). 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 reads from degraded RNA and thus reduced representation of full-length transcripts.

Comparing splice junction identification accuracy

To understand the accuracy of each pipeline for predicting splice junctions, we looked at both mapping mismatch 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. 1c) (See [Methods](#) for more detailed explanation of wobble). While the mismatch percentage of mapped reads are often used to assess the improvement of long read data from different error correction pipelines [20], 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 that 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, another 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 mapping mismatch profiles for each mapped read for the inter-read error correction pipelines (Polish and Lordec) and the pipelines using the mapped FLNC reads (TAMA High and TAMA Low). Note that the mapped FLNC reads are the same for the TAMA High and TAMA Low pipelines.

Using the output from TAMA Collapse we looked at length of mapped read 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 since difference between the reads and the reference genome assembly may be caused by real polymorphism. We calculated the average mismatch rates 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. Mismatches evaluated include soft clipping, insertion, deletion, and substitution mismatches but do not include hard clipping.

The mapped FLNC reads (used in TAMA High/Low pipelines) had the highest average predicted mismatch rate (2.83%) and the highest amount of each type of mismatch while the Cluster/Polish reads had the lowest mismatch rates (0.52%) with the lowest amount of each type of mismatch. The LoRDEC error corrected reads (average 1.38% mismatch rate) had a similar amount of clipping mismatches as compared to the mapped FLNC reads (Fig. 5a). This indicates that LoRDEC correction may have some issues correcting the ends of reads that may be due to lower short read coverage at the ends of transcripts.

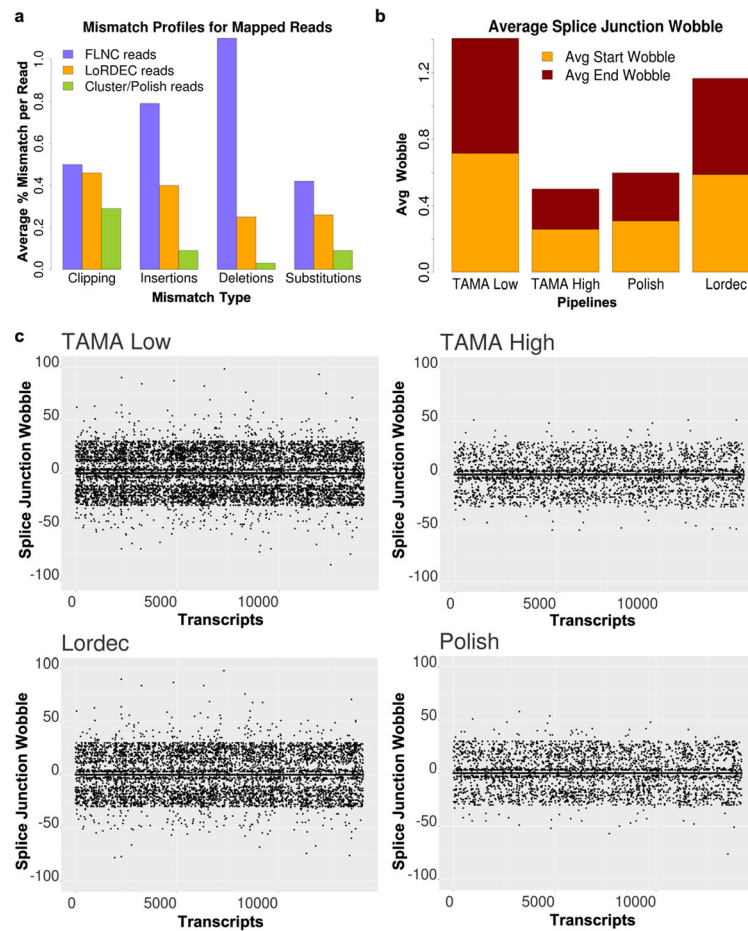


Fig. 5 Error rate estimation and wobble across pipelines. **a** The average percent of alignment mismatch by mismatch type across pipelines. **b** Average splice junction wobble across all transcript models which matched the Ensembl annotation in all four pipelines. A splice junction wobble threshold of 30 bp on each side of the splice junction was allowed for matching for these plots. Note that wobble greater than 30 bp is possible due to wobble walking. **c** Scatter plots to illustrate the amount of wobble across all pipelines assessed on the transcript models used in Average Splice Junction Wobble plot

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 four different pipelines (Fig. 5b-c). Wobble typically occurs due a large number of read errors immediately flanking the splice junctions leading to small shifts in mapping the ends of each exon [21]. The total wobble for a splice junction within grouped reads can be larger than the specified wobble threshold due to a phenomenon we call wobble walking. Wobble walking occurs when the predicted exon starts/ends are represented in staggered formation so that the difference between each closest pair is still within the wobble threshold but the difference between the most distant pair is greater than the threshold (Fig. 1c). 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. For instance, the expectation is that if a transcript model from a long read based annotation contains identical splice junctions (a splice junction wobble of zero) as compared to a reference annotation, then the long read based transcript model has the correct predicted splice junctions. We ignored wobble at the transcript start and end sites due to the high variance of these features in natural RNA [22, 23]. 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 High pipeline with stringent LDE filtration had the lowest average wobble values per splice junction while the TAMA Low pipeline produced the highest average wobble (Fig. 5b-c). Thus, despite the lower overall error rates in the mapped reads from the Polish pipeline, the TAMA High pipeline had more splice junctions matching the Ensembl annotation. This

suggests that the LDE filtration in the TAMA High pipeline resulted in more accurate identification of splice junctions.

Inter-read error correction mis-clustering may produce erroneous gene models

One of the major concerns when using inter-read error correction methods such as Cluster/Polish and LorDEC is the possibility of combining read sequences from different transcripts that would result in erroneous transcript models. The different transcripts could 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 [23]. 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 mapped FLNC reads (TAMA Low) to the inter-read error corrected reads (Polish and Lordec) to find reads that mapped to different genes and transcripts in each comparison. While it is possible that the FLNC read mappings are erroneous, they represent the read sequences without any over-correction. Also reads that map to different loci after inter-read error correction indicate that there is enough sequence ambiguity to call into question the effect of the inter-read error correction.

Comparing the mapped FLNC reads to the Cluster/Polish mapped reads, we found 34,637 reads (0.6% of mapped reads) that switched from one gene locus to another after Cluster/Polish correction (Fig. 6a). This gene loci switching involved 6774 genes, 3230 of which were only found with the TAMA 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 may reduce gene discovery by combining reads from low expression genes with high expression genes.

To assess the effect of hybrid inter-read error correction on gene level read jumbling, we compared the mapped FLNC reads to the mapped LorDEC corrected reads. There were 19,064 reads (0.3% of mapped reads) which switched from one gene locus to another (Fig. 6b), involving a total of 3476 genes, 775 of which were only found with the TAMA Low pipeline while 675 genes were only found with the Lordec pipeline.

To gain a more detailed understanding of what happens during a read jumble event, we examined the PReferentially expressed Antigen of MELanoma (PRAME) gene family. The PRAME gene family is highly associated with cancer development [24] and is used as a biomarker for identifying various forms of cancer. Within the PRAME gene family there are 24 annotated paralogues [25]. In this example, the Polish pipeline fails to detect one of the PRAME paralogues (PRAMEF8) while erroneously predicting the expression of another paralogue (PRAMEF15) which has no FLNC mapped read support. The TAMA Low pipeline (using FLNC mapped reads) finds 9 reads mapping to PRAMEF8 (Fig. 6c) while the Polish pipeline (using Cluster/Polish mapped reads) shows no reads mapping to PRAMEF8. Of the 9 PRAMEF8 reads from the TAMA Low pipeline, 5 of these reads were clustered and combined with other reads (3 from PRAMEF11, 4 from PRAMEF4, 2 from PRAMEF7, and 3 from PRAMEF27 according to FLNC mapping) into 1 cluster read by Cluster/Polish resulting in a jumbled cluster read mapping to the PRAMEF15 gene (Polish pipeline). We analyzed the sequence similarity between the two paralogues by aligning the PRAMEF8 and PRAMEF15 transcript sequences with Muscle [26] and found that they had 76% identity. While the two genes have similar exonic sequences, the genome mapping identity for the reads were higher than the sequence similarity between the two paralogues. The PRAMEF8 FLNC read with the lowest genome mapping identity score had a mapping identity of 89% and 6 PRAMEF8 FLNC 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.

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 1372 transcript were found only in the Polish annotation. 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 7117 transcripts found only in the TAMA Low annotation and 11,732 transcript found only in the Lordec annotation. It is important to note that this transcript level

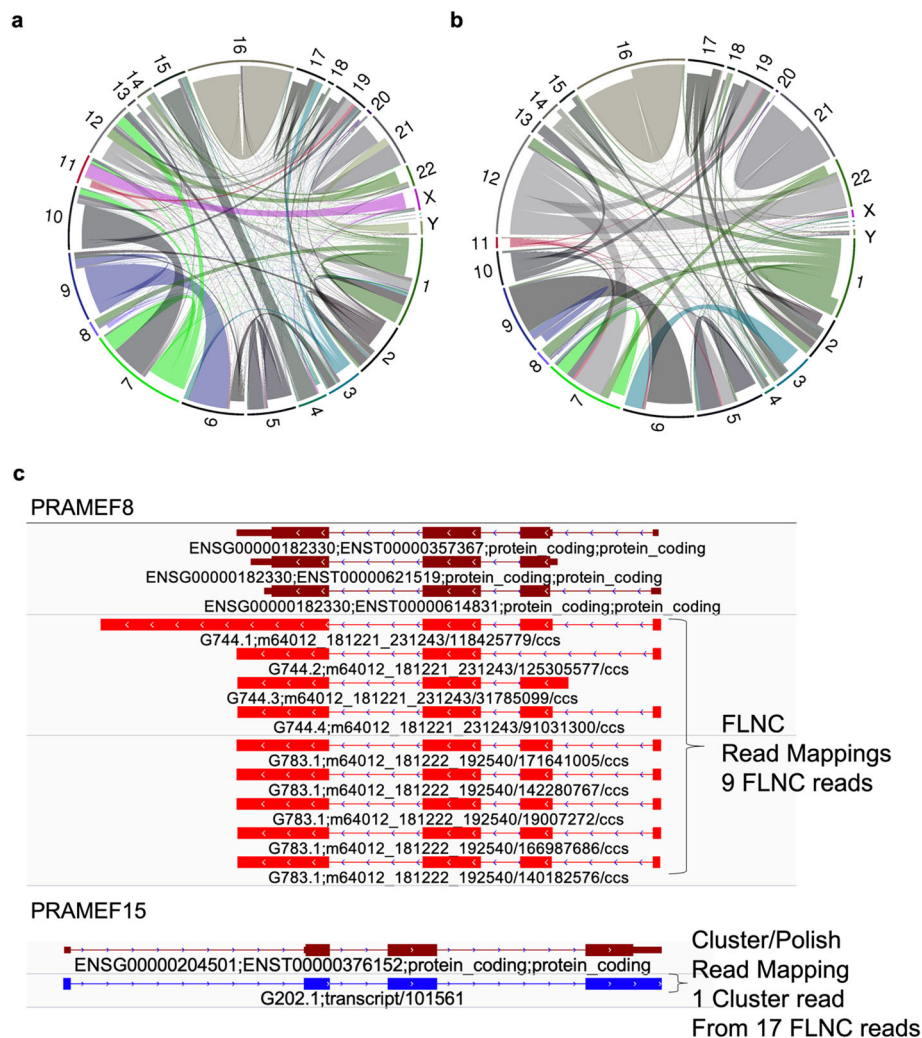


Fig. 6 Gene and transcript read swapping from error correction. **a** Circos plot showing reads mapping to different loci after using Cluster/Polish for long inter-read error correction. Each line represents one read and the width of each chromosome bin represents the number of reads (combined thickness of each line). The indented line ends shows FLNC read location and non-indented ends shows read allocation after inter-read error correction. This plot shows 34,637 reads from 4799 genes moving to 2793 genes after Cluster/Polish. The reads are organized by chromosome however swapping occurs within chromosome and between chromosomes. **b** Circos plot as above but after hybrid inter-read correction with LoRDEC. Each line represents a single read moving from one gene to another with 19,064 reads from 2292 genes moving to 2319 genes after LoRDEC error correction. **c** The PRAMEF8 gene has coverage from 9 FLNC mapped reads (TAMA Low). Five of these reads were clustered and combined with other reads into one cluster read by Cluster/Polish resulting in a jumbled cluster read mapping to the PRAMEF15 gene (Polish pipeline). This suggests a false negative for PRAMEF8 and false positive for PRAMEF15 in the Polish pipeline due to the use of Cluster/Polish

jumbling assessment is only a rough indication since without a ground truth for real transcripts it is impossible to know which transcript model is accurate.

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, to avoid read jumbling issues we suggest foregoing 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.

Analysis of predicted expressed loci not found in the Ensembl human annotation

Given that the TAMA High pipeline had the highest sensitivity and precision scores for non-guided annotation in the benchmarking datasets, we used the gene loci predicted by the TAMA High pipeline to investigate potentially novel genes within the UHRR dataset. To gain insight into the 23,302 TAMA High predicted gene

models not found in Ensembl (TAMA High specific gene models), we looked at several features which provide support for or against real gene models: coding potential, number of exons, intronic overlap with other genes, overlap with regulatory features, and the presence of immediately downstream genomic poly-A stretches. The combination of coding potential and splice junctions is often used as 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 [27], 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 [28] database. We combined these three methods to account for the various errors that can cause false negatives in protein coding gene prediction.

Only a small number of the TAMA High predicted gene models which were not found in the Ensembl v94 annotation (18 out of 23,302) were supported by all features which are considered evidence for functionality (multi-exonic, coding, intergenic, and processed poly-A) (Fig. 7). 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 identified within the Ensembl annotation.

There were 1059 TAMA High specific gene models which were intergenic, single exonic, and had genomic poly-A. These features are commonly ascribed to genomic DNA contamination. However, the precise mechanism for how these sequences make it through to the final sequencing library is not well characterized.

The two most common sets of features for the TAMA High specific gene models are “single exonic, non-coding, intronic gene overlap, and genomic poly-A” at 24% (5679) and “single exonic, coding, intronic gene overlap, and genomic poly-A” at 19% (4440). These feature sets are typically used as indicators for non-real models since they could be derived from internal priming of unprocessed RNA. However, this would require further truncation of the template so that the resulting model does not overlap with transcripts from the gene of origin. In theory a subset of loci with the first feature set could be comprised of lncRNA while a subset of loci

with the second feature set could be comprised of processed pseudogenes. Together, these account for over 43% of the TAMA High specific gene models.

There were 2566 (11% of TAMA High specific gene models) gene models that were predicted to be non-coding with processed poly-A tails. Of these, 461 were multi-exonic while 2105 were single exon genes (Fig. 7). Given that these models did not overlap any exonic regions of gene models in the Ensembl annotation, this would represent a large increase in the number of predicted lncRNA for the human genome.

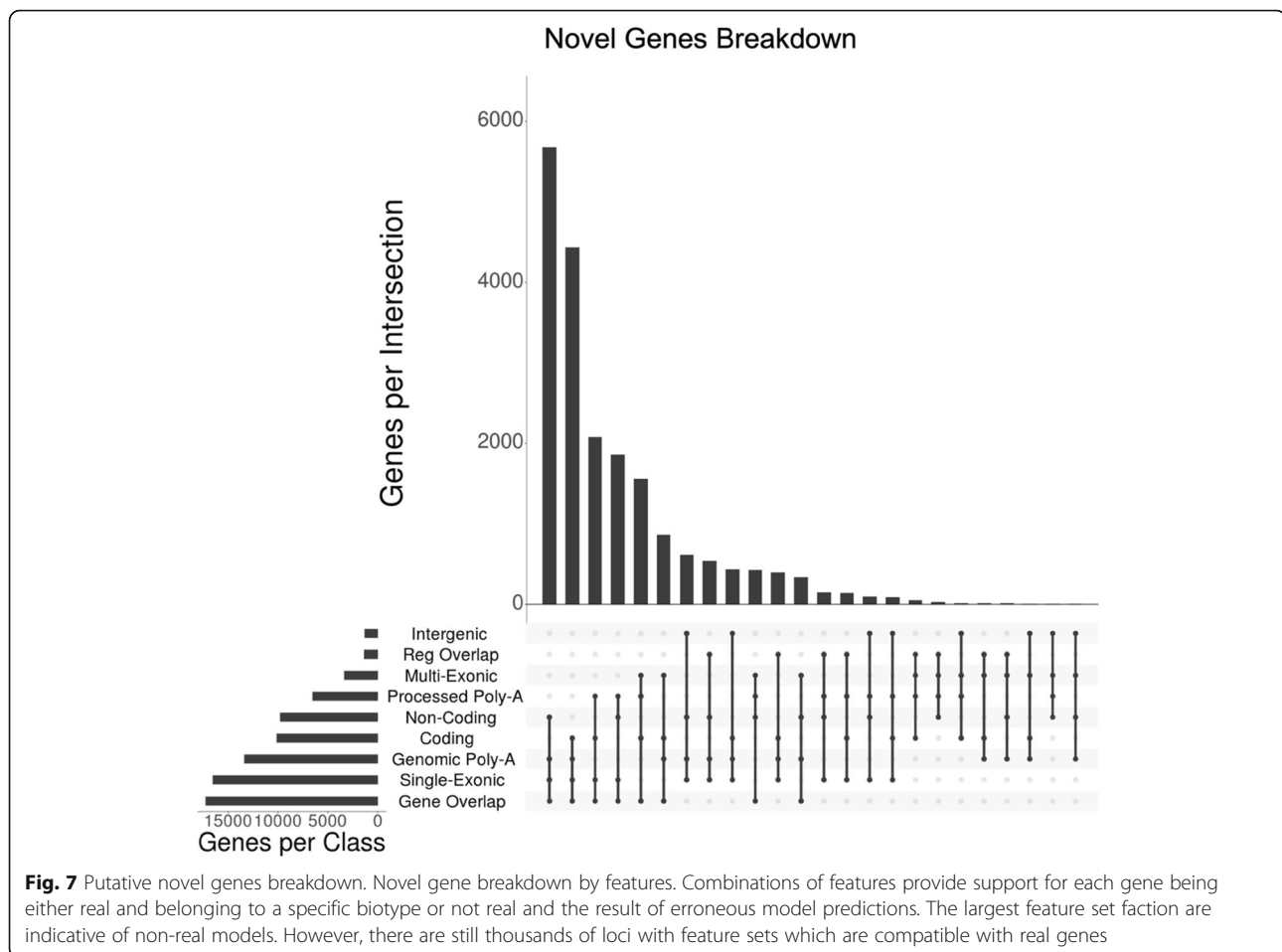
There were 1557 (7%) TAMA High specific gene models with features (multi-exonic, coding, intron overlapping, and processed poly-A) that are indicative of real protein coding genes that exist within the introns of larger genes. However, it is possible that these are alternative transcripts from the surrounding genes but due to lack of 5' completeness, the overlapping 5' exons were not represented in the transcript models. If these gene models are derived from alternative transcripts of their surrounding genes, these models would represent novel transcripts.

These analyses were based on the Ensembl v94 human annotation, the Ensembl v100 annotation has since been released. This new Ensembl version has more than a thousand new lncRNA gene models as compared to v94. We compared the TAMA High annotation to v100 and found 144 matching lncRNA genes that were not present in v94. 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.

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 and/or sequencing analysis to confidently identify novel genes.

Discussion

The UHRR PacBio Sequel II Iso-Seq dataset is the result of one of the most accurate high-throughput long read transcript sequencing technologies [29] 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. To date, 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 TAMA, we have demonstrated some key issues



with current long read RNA sequencing data pipelines that could have a major influence on current transcriptomic studies. Firstly, mis-mapping of reads with sequence errors around splice junctions (error generated wobble) can produce transcript models with false splice junction predictions. Secondly, RNA degradation can result in 5' incomplete transcript models that can have large downstream effects for data processing and interpretation. Thirdly, inter-read error correction can also cause false positives and negatives for gene and transcript model predictions. Finally, the combination of these problems also brings up challenges for using long read data in expression quantification experiments. If a significant number of reads can change transcript assignment due to either lack of 5' completeness or changes in mapping loci after inter-read error correction, quantification estimates may not reflect the true biological state. While sequence error correction is currently the main focus of many long-read bioinformatic tools, it should not be applied at the cost of biological accuracy as could be the case for the gene and transcript read jumbling events from long read and short read inter-read error correction.

The resulting transcriptome annotation with TAMA portrays a very different composition of gene models compared to public transcriptome annotations. These differences suggest the existence of possibly thousands of potential novel genes (many of which are classified within under-represented biotypes) and/or artifacts arising during the sequencing pipeline.

The underlying issue in all methodologies is the balance between retaining useful information and discarding misleading information. However, the treatment of long read RNA data requires customization to accommodate both the sequencing technology as well as the biological assumptions. The TAMA tool kit is designed to allow the user to tune its behavior. This means that TAMA Collapse and TAMA Merge can be used with less stringent settings for maximum discovery potential and/or high stringent parameters for curating reference annotations. The resulting gene models can be assessed with the TAMA ORF/NMD pipeline for identifying coding similarity to know protein coding genes. However, more development is needed for discerning between long non-coding RNA and RNA sample noise. This may require wet lab methods such as improved 5' cap selection for biasing against RNA sample noise.

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

Conclusions

Long-read transcript sequencing presents new challenges for annotating transcriptomes. Analysis of the UHRR PacBio Sequel II Iso-Seq data suggests that there may be thousands of unannotated non-coding genes within the human genome. However, the methods for sequencing library preparation and data processing require more development to distinguish expressed genes from sequencing noise. Wobble analysis should complement read mapping identity for assessing transcript assembly pipelines. Degraded RNA within samples can lead to 5' incomplete transcript model predictions. Inter-read error correction (prior to mapping) can cause read jumbling and gene model ambiguity. Read jumbling represents one challenge for using long read data for quantification. Long-read sequencing analysis benefits from tools (such as TAMA) which allow for algorithmic tuning to accommodate sequencing error rates and biological assumptions.

Methods

TAMA Version Commit 39c1270c6e1ef2cf5d39f7f047-fa15e0f1a6c790 was used for this study.

More detailed information on how TAMA works can be found here:

<https://github.com/GenomeRIK/tama/wiki>

Wobble

Wobble is defined in this text as the distance measured in bases between the mapped starts and ends for exons. This term is used to describe small differences (< 50 bp) in predicted starts/ends based on mapped reads. These differences can occur due to real differences in starts/ends or due to errors in the reads flanking the starts/ends. For example, if a read has a number of missing bases immediately flanking a splice junction (SJ are comprised of one exon start and one exon end), the predicted splice junction from mapping may be off by the same number of missing bases. TAMA Collapse and TAMA Merge both use wobble to allow for the grouping of reads to be collapsed into a single transcript model. This is assessed by comparing every pair of transcript models within the same genomic loci (at least 1 bp same strand overlap connecting all loci grouped reads). In each pair assessment, each exon start and end from each predicted transcript model is compared to see if they occur within the user defined wobble threshold.

Due to this allowance of wobble between predicted starts and ends of exons, a phenomenon termed in this

text as wobble walking can occur (Fig. 1c). Wobble walking is defined as a situation where 3 or more transcript models have exon starts or ends where the most upstream exon start/end prediction and the most downstream exon start/end prediction occur at a distance greater than the wobble threshold. However, the other exon start/end positions occur in such a way that when ordered based on genomic position there are no consecutive pairs of exon starts/ends which are farther apart than the wobble threshold. Thus by using the pairwise non-stochastic method of matching transcript models, all transcript models in this situation would match due to the linking effect across all represented exons starts/ends. When this situation occurs, the distance between the exon starts/ends between the grouped transcripts used for collapsing can be greater than the user defined wobble threshold.

TAMA collapse

TAMA Collapse performs multiple functions: transcriptome assembly, variant calling, genomic downstream poly-A detection, and transcript/gene level quantification. The primary function is to create a non-redundant error corrected genome reference based transcriptome annotation. TAMA Collapse takes as input a sorted SAM/BAM file representing long read RNA sequencing data mapped onto a reference genome assembly as well as a fasta file representing the reference genome assembly used for mapping. TAMA Collapse is designed to be highly tunable and relies on 4 main parameters to define its behaviour: wobble thresholds, collapse mode, splice junction ranking, and the amount of mapping mismatch surrounding splice junctions (LDE).

The wobble thresholds and collapsing modes are used to define how mapped reads are grouped for collapsing. Wobble thresholds can be defined for the TSS, TES, and SJ. Wobble thresholds are given in integer values representing base pair distances. These thresholds define the limit between two features (such as TSS) to be considered a matching feature. There are two collapsing modes which are termed capped and non-capped modes. The capped mode requires that all grouped transcript models (mapped reads) have the same number of exons and all their exons have matching start and end sites as per the user defined wobble thresholds. Matches are performed pairwise in a non-stochastic algorithm. This pair-wise matching is what leads to wobble walking.

The splice junction ranking and local density error algorithm are designed to identify the most likely real splice junctions given a group of matching transcript models. Both the splice junction ranking and LDE rely on user defined threshold of distance from SJ to assess. The LDE feature can be turned on or off. When turned on, the user can specify the distance from the splice

junction to assess and the number of allowed mismatches within that distance. If the number of mismatches exceed the threshold, the read is discarded. This is intended to prevent erroneous splice junction predictions. The splice junction ranking can be turned on or off by the user. When turned off, the splice junctions are selected based on the the highest read coverage. When splice junction ranking is turned on, TAMA Collapse ranks the splice junction read support based on the amount of mismatches flanking the splice junctions. In this method, a splice junction with read support where there are no mismatches flanking the splice junction is given the highest rank and chosen as the final predicted splice junction.

While TAMA Collapse has multiple file outputs, the main output is a bed12 formatted annotation file containing all non-redundant transcript models.

TAMA merge

TAMA Merge is designed to remove transcript model redundancy either between multiple input annotations or within a single input annotation. TAMA Merge accepts as input 1 or more annotations in bed12 format. TAMA Merge has multiple output files, however the main output file is an annotation file in bed12 format. TAMA Merge also keeps track of the transcript models and their source annotation which were “merged”. This means that for each transcript model, TAMA Merge provides information on which input annotations had transcripts matching it. TAMA Merge uses the same wobble parameter/algorithm and collapsing modes as TAMA Collapse. However, individual input files can be assigned different collapsing modes. This is useful for merging long read data which is likely to contain 5’ truncated transcript models with a reference annotation. In addition to collapsing mode and wobble thresholds, TAMA Merge allows user to assign priority to different input annotation for features such as TSS, TES, and SJ. For instance, a short read derived annotation can be given priority for SJ, while a long read annotation can be given priority for TSS and TES.

TAMA read support levels

The `tama_read_support_levels.py` tool is designed to generate a file that relates each transcript and gene model with the ID’s of reads which were used to generate those models. This can also be thought of as producing read count information for transcripts and genes. The `tama_read_support_levels.py` tool works on all annotation output files from all TAMA modules as well as on PacBio annotation files. This tool was used to identify reads that were involved in read jumbling.

TAMA filter fragments

The `tama_remove_fragment_models.py` tool is used to remove transcript models that appear to be fragments of full length models. The criteria for fragment models is that they contain the same internal exon-intron structure as a transcript that is longer on both the 5’ and 3’ ends. The splice junction wobble can be adjusted by the user.

TAMA remove single read models

The `tama_remove_single_read_models_levels.py` tool is used to filter a transcriptome annotation based on the amount of read support for each transcript model. This can be run on either the results of TAMA Collapse or the results of TAMA Merge. When used with TAMA Merge with multiple input annotations, `tama_remove_single_read_models_levels.py` can filter out models based on the number of supporting sources for each transcript model. When TAMA Merge is used to merge a long read data based annotation with a reference annotation, `tama_remove_single_read_models_levels.py` can be used to filter out models in the long read annotation that do not match the reference annotation. This is how TAMA performs guided annotation.

TAMA find model changes

The `tama_find_model_changes.py` tool is designed to identify reads which have different transcript/gene model assignments between different pipelines. This is referred to as read jumble in this study. This tool takes as input a TAMA Merge annotation which was generated by merging annotations from the 2 pipelines to be compared. This tool also requires a read support file generated by `tama_read_support_levels.py`. Read jumbles are identified by using the read ID’s and comparing the transcript models they are assigned to within the TAMA Merge annotation file. Any read that supports more than 1 transcript model is considered to be involved in a read jumbling event.

TAMA ORF/NMD pipeline

The TAMA ORF/NMD pipeline is a method for identifying open reading frames (ORF) from transcript models and relating them to known protein coding genes. Non-sense mediated decay (NMD) product predictions are also made by identifying stop codons which occur 50 bp upstream of a splice junction. The first step of the pipeline is the conversion of the transcript nucleotide sequences into amino acid sequences. This is done by looking for all ORF’s which have a stop codon and selecting the longest ORF’s from each frame (3 forward strand frames). Start codons are not required for an ORF prediction, however, if a start codon is not found, the corresponding ORF is labeled as evidence that the

transcript is from a degraded RNA. BlastP is then used to relate the resulting amino acid sequences to a protein database. The ORF from each transcript with the best hit to the database is then selected as the predicted true ORF. Using the ORF information, the transcripts are then labeled with attributes based on the protein hit.

TAMA degradation signature

The TAMA Degradation Signature (DegSig) score is intended to provide a metric for the relative amount of sequencing reads originating from degraded RNA. The DegSig score is calculated by the following formula:

$$\text{DegSig} = (\text{CT} - \text{NT}) / \text{CT}$$

Where CT is the number of multi-exon transcript models from genes with more than 1 read support after using TAMA Collapse with the capped mode, and NT is the number of multi-exon transcript models from genes with more than one read support after using TAMA Collapse with the no_cap mode.

Simulated long read datasets and processing for benchmarking

The simulated PacBio and Nanopore datasets (https://figshare.com/articles/RNA_benchmark_datasets/5360998) were produced in another study [11] using PBSIM [12]. These datasets were also used and described in the Stringtie2 paper [9].

Both datasets were mapped to chromosome 19 of the human reference genome as provided in the simulated dataset. Minimap2 [30] (version 2.15-r915-dirty) with the parameters “--secondary=no -ax splice -uf” was used for mapping. Samtools [31] (version 1.9) was used for all SAM/BAM file handling.

For the TAMA Low processing, TAMA Collapse was used with the parameters “-d merge_dup -x \${capflag} -a 200 -z 200 -sj sj_priority -log log_off -b BAM”. For the TAMA High processing, TAMA Collapse was used with the parameters “-d merge_dup -x no_cap -a 300 -m 20 -z 300 -sj sj_priority -lde 3 -sjt 10 -log log_off -b BAM”. After TAMA Collapse, both TAMA Low and TAMA High shared the same processing with `tama_remove_fragment_models.py` used with default parameters to remove transcript models that appear to be fragments of longer models. This resulted in the final annotations for both pipelines.

For the TAMA Guided pipeline, the output from the TAMA Low TAMA Collapse run was merged with the reference annotation containing both expressed and non-expressed transcript models using TAMA Merge with “-a 300 -z 300 -m 20 -d merge_dup” parameters. The input `filelist.txt` file for TAMA Merge set both annotations to capped mode with full priority (1,1,1) given to the reference annotation. The `tama_remove_single_read_models_levels.py` tool was then used with “-l

transcript -k remove_multi -s 2” parameters resulting in the final annotation. The `tama_read_support_levels.py` tool was used at each step of processing to keep track of read support for each transcript model.

For the Stringtie2 pipeline, Stringtie2 (v2.1.3b) was used with the “-L” parameter after mapping.

For the Stringtie2 Guided pipeline, Stringtie2 (v2.1.3b) with “-L -G <reference annotation>” parameters was used. The reference annotation used was the same annotation as used in in TAMA Merge for the TAMA Guided pipeline.

For the TALON pipeline (unguided), a blank database was created using “`talon_initialize_database`” with default settings and an empty GFF file. Then “`talon_label_reads`” was used with “--t 1 --ar 20 --deleteTmp” parameters. Then default “`talon`” was used. This was followed by “`talon_filter_transcripts`” using “--maxFracA 0.5 --minCount 5 --minDatasets 1” parameters. The default “`talon_create_GTF`” was used to create a GTF file for the annotation.

For the TALON guided pipeline, a database was created using “`talon_initialize_database`” with default settings and the same GFF reference annotation file used for TAMA Guided and Stringtie2 Guided. Then default “`talon`” was used. This was followed by “`talon_filter_transcripts`” using “--maxFracA 0.5 --minCount 1 --minDatasets 2” parameters. The default “`talon_create_GTF`” was used to create a GTF file for the annotation.

All resulting annotations were compared to the annotation file containing all expressed transcript models using GffCompare (v0.11.2).

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](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.

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*).

Lexogen SIRV Iso-Seq dataset benchmarking

The UHRR Sequel II Iso-Seq data also contained a spike-in of Lexogen SIRV RNA. For the Cupcake pipeline we used the FLNC reads from each SMRT cell and used Cluster/Polish for long read inter-read error correction. We then mapped the resulting reads using Minimap2 (*--secondary=no -ax splice -uf -C5*) to the “SIRV_isoforms_multi-fasta_170612a.fasta” reference genome assembly provided by Lexogen. After mapping we ran Cupcake Collapse “collapse_isoforms_by_sam.py” with the Cupcake manual recommended settings “*--dun-merge-5-shorter*”. We then used Cupcake “chain_samples.py” to merge the assemblies from each SMRT Cell. This resulted in the final annotation for the Cupcake pipeline.

For all the other pipelines (TAMA Low, TAMA High, TAMA Guided, Stringtie2, Stringtie2 Guided, TALON, and TALON Guided), we mapped the FLNC reads to the same reference genome as above using the same parameters for Minimap2.

For the TAMA Low processing, TAMA Collapse was used with the parameters “*-d merge_dup -x no_cap -sj sj_priority -log log_off -b BAM -lde 5 -sjt 20 -a 100 -z 100*”. For the TAMA High processing, TAMA Collapse was used with the parameters “*-d merge_dup -x no_cap -sj sj_priority -log log_off -b BAM -lde 1 -sjt 20 -a 100 -z 100*”. After TAMA Collapse, both the TAMA Low and TAMA High pipelines used TAMA Merge (*-a 100 -z 100 -d merge_dup*) was used to merge the TAMA Collapse outputs from each SMRT Cell. The *tama_remove_single_read_models_levels.py* tool was then used with “*-l transcript -k remove_multi -s 2*” parameters resulting in the final annotation. The *tama_read_support_levels.py* tool was used at each step of processing to keep track of read support for each transcript model.

For the TAMA Guided pipeline, TAMA Collapse (*-d merge_dup -x capped -sj sj_priority -log log_off -b BAM -a 0 -m 0 -z 0*) was used on the Minimap2 output files for each SMRT Cell. TAMA Merge (*-d merge_dup -a 0 -m 0 -z 0*) was then used to combined the TAMA Collapse outputs from each SMRT cell. TAMA Merge (*-d merge_dup -a 0 -m 0 -z 0*) was then used again to match the output with the SIRV annotation file (SIRV_isoforms_multi-fasta-annotation_C_170612a.gtf).. The *tama_remove_single_read_models_levels.py* tool was then used with “*-l transcript -k remove_multi -s 2*” parameters resulting in the final annotation. The *tama_read_support_levels.py* tool was used at each step of processing to keep track of read support for each transcript model.

For the Stringtie2 pipeline, Stringtie2 (v2.1.3b) was used with the “*-L*” parameter after mapping.

For the Stringtie2 Guided pipeline, Stringtie2 (v2.1.3b) with “*-L -G <reference annotation>*” parameters was used. The reference annotation used was the same annotation as used in in TAMA Merge for the TAMA Guided pipeline.

For the TALON pipeline (unguided), a blank database was created using “*talon_initialize_database*” with default settings and an empty GFF file. Then “*talon_label_reads*” was used with “*--t 1 --ar 20 --deleteTmp*” parameters. Then default “*talon*” was used. This was followed by “*talon_filter_transcripts*” using “*--maxFracA 0.5 --minCount 10 --minDatasets 2*” parameters. The default “*talon_create_GTF*” was used to create a GTF file for the annotation.

For the TALON guided pipeline, a database was created using “*talon_initialize_database*” with default settings and the same GFF reference annotation file used for TAMA Guided and Stringtie2 Guided. Then default “*talon*” was used. This was followed by “*talon_filter_transcripts*” using “*--maxFracA 0.5 --minCount 5 --minDatasets 2*” parameters. The default “*talon_create_GTF*” was used to create a GTF file for the annotation.

All resulting annotations were compared to the Lexogen SIRV annotation file (https://www.lexogen.com/wp-content/uploads/2018/08/SIRV_Set2_Sequences_170612a-ZIP.zip) using GffCompare (v0.11.2).

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 [4].

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 Tel-oPrime kit. The resulting cDNA library was sent to Edinburgh Genomics for sequencing on the Sequel system using 2.0 chemistry.

TAMA low pipeline for UHRR

Full descriptions of the TAMA algorithms can be found in the wiki pages of the Github repository (<https://github.com/GenomeRIK/tama/wiki>). FLNC reads were mapped to GRCh38 (Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa) 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). The *tama_read_support_levels.py* tool was used at each step of processing to keep track of read support for each transcript model.

TAMA high pipeline for UHRR

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 TAMA Low pipeline. Transcript models supported only by reads from a single SMRT Cell were filtered out using *tama_remove_single_read_models_levels.py* (*-l transcript -k remove_multi -s 2*). The *tama_read_support_levels.py* tool was used at each step of processing to keep track of read support for each transcript model.

Polish pipeline for UHRR

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*). The *tama_read_support_levels.py* tool was used at each step of processing to keep track of read support for each transcript model.

Lordec pipeline for UHRR

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://rna-journal.cshlp.org/content/22/4/597.full.pdf>). The resulting error-corrected reads were processed in the same way as the TAMA Low starting from the mapping step. The *tama_read_support_levels.py* tool was used at each step of processing to keep track of read support for each transcript model.

Finding transcript matches and loci overlap between Iso-Seq annotations and the Ensembl annotation

We used TAMA Merge to compare the annotations from each Iso-Seq pipeline (TAMA Low, TAMA High, Polish, and Lordec) to the Ensembl v94 annotation. All input annotations were set to capped mode in the input fielist.txt files. The “-a 300 -z 300 -m 0 -d merge_dup” parameters were used to run TAMA Merge. Transcript matches were identified from the *trans_report.txt* file while gene loci overlap was identified from the *gene_report.txt* file.

Comparing 5' completeness between the TAMA high, polish, and Ensembl v94 annotations

We used TAMA Merge to compare the annotations for pairs of annotations (TAMA High-Polish, TAMA High-Ensembl, Polish-Ensembl). Both annotations in each merging were given *no_cap* parameters in the *fielist.txt* input file. We used the same TAMA Merge settings as were used for identifying matching transcript models between annotations. We used the TAMA Merge *trans_report.txt* output file to identify which source annotation had the longer 5' representation for each matching transcript model.

Degradation signature analysis

We split the SAM files from the mapping by chromosome. We then used these single chromosome SAM files as inputs to 2 TAMA Collapse runs. One TAMA Collapse run used the capped mode and the other run used the *no_cap* mode. Both runs used “-a 100 -z 100 -sj sj_priority -lde 5 -sjt 20 -log log_off -b BAM” parameter settings. We then used the *trans_read.bed* files from each pair of TAMA Collapse runs as inputs for the *tama_degradation_signature.py* tool which calculated the DegSig scores.

Mismatch and wobble analysis

The mismatch profiles for the mapped FLNC, Cluster/Polish corrected, and LoRDEC corrected reads were extracted from the TAMA Collapse *read.txt* output files generated in each pipeline.

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.

Read jumbling analysis

Read ID's were tracked through each processing step using the *tama_read_support_levels.py* tool. TAMA Merge was used to combine the annotations from the different pipelines (TAMA Low-Polish, TAMA Low-Lordec) using the same parameters as the was used in the wobble analysis. The TAMA Merge output and *tama_read_support_levels.py* outputs were used as input for the *tama_find_model_changes.py* tool that identified reads which had different transcript model assignment between each pair of pipelines.

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 [27]) was used to segregate between coding and non-coding transcripts.

We used the TAMA ORF/NMD pipeline for the third source of coding evidence. The transcript models were converted into fasta sequences using Bedtools [32]. ORFs were predicted for each transcript from the fasta file then translated into amino acid sequences. BlastP [33] (*-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.

Matching TAMA high annotation to Ensembl v100

For identifying gene models found in the Ensembl v100 human annotation matching gene models predicted in the TAMA High annotation which were not present in the Ensembl v94 human annotation, we used TAMA Merge with “-m 0 -a 300 -z 300” parameters in capped mode for all three annotations (TAMA High, Ensembl v94, and Ensembl v100). These parameters group transcript models between the annotations if they share the exact same splice junctions and exon chaining but with an allowance of up to 300 bp difference in TSS and TES. We then identified all gene models which were the product of merging a TAMA High annotation gene with an Ensembl v100 gene and with no Ensembl v94 gene represented for that loci.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12864-020-07123-7>.

Additional file 1: Figure S1. Histogram plot of FLNC read lengths.

Additional file 2: Table S1. Table of long read datasets information on number of mapped reads, reference annotation, and genome assembly scaffold numbers.

Abbreviations

CCS: Circular consensus sequence; DegSig: Degradation signature; ECC: Exon cascade collapse; FLNC: Full length non-chimeric reads; LDE: Local density error; lncRNA: Long non-coding RNA; NMD: Nonsense mediated decay; PacBio: Pacific bioscience; RNA-seq: RNA sequencing; SIRV: Spike-in RNA variant control; TAMA: Transcriptome annotation by modular algorithms; TSSC: Transcription start site collapse; UHRR: Universal human reference RNA

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.

Code availability

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

Authors' 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 contributed the conception and tested the TAMA Collapse LDE feature. JS, ALA, JWSB, and RZ reviewed and edited the manuscript. The author(s) read and approved the final manuscript.

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.

Availability of data and materials

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](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 Telo-Prime cap selected chicken brain Iso-Seq data is available from the European Nucleotide Archive under accession number PRJEB25416. The simulated long read dataset is available here: https://figshare.com/articles/RNA_benchmark_datasets/5360998

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian EH25 9RG, UK. ²The University of Queensland, St. Lucia, Brisbane, QLD 4072, Australia. ³School of Life and Environmental Sciences, University of Sydney, Sydney, New South Wales, Australia. ⁴Information and Computational Sciences, The James Hutton Institute, Invergowrie, Dundee, Scotland, UK. ⁵Plant Sciences Division, School of Life Sciences, University of Dundee, Invergowrie, Dundee, Scotland, UK. ⁶Cell and Molecular Sciences, The James Hutton Institute, Invergowrie, Dundee, Scotland, UK.

Received: 15 April 2020 Accepted: 6 October 2020

Published online: 30 October 2020

References

- Salzberg SL. Next-generation genome annotation: We still struggle to get it right. *Genome Biol.* 2019;20(1):19–21.
- 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.
- Minoche AE, Dohm JC, Schneider J, Holtgräwe D, Viehöver P, Montfort M, et al. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. *Genome Biol.* 2015;16:184 Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4556409&tool=pmcentrez&rendertype=abstract>.
- 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 Genomics.* 2017;18(1):1–19.
- 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.* 2016;7:11708 Available from: <http://www.nature.com/doi/10.1038/ncomms11708>.

6. 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*. 2017;18(1):395 Available from: <http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3757-8>.
7. 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*. 2015;10(7):1–15. <https://doi.org/10.1371/journal.pone.0132628>.
8. Koepfli K-P, Paten B, O'Brien SJ. The Genome 10K Project: A Way Forward. *Annu Rev Anim Biosci*. 2015;3(1):57–111.
9. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol*. 2019;20(1):1–13.
10. Wyman D, Balderrama-gutierrez G, Reese F, Jiang S, Rahmanian S, Zeng W, et al. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv*. 2019.
11. Križanović K, Echchiki A, Roux J, Šikić M. Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics*. 2018;34(5):748–54.
12. Ono Y, Asai K, Hamada M. PBSIM: PacBio reads simulator—toward accurate genome assembly. *Bioinformatics*. 2013;29(1):119–21 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23129296>. Cited 2014 Oct 8.
13. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun*. 2020;11(1):1–12. <https://doi.org/10.1038/s41467-020-15171-6>.
14. Kuosmanen A, Sobih A, Rizzi R, Mäkinen V, Tomescu AI. On using longer RNA-seq reads to improve transcript prediction accuracy. In: *Bioinforma 2016 - 7th Int Conf Bioinforma Model Methods Algorithms, Proceedings; Part 9th Int Jt Conf Biomed Eng Syst Technol BIOSTEC 2016*. 2016. p. 272–7.
15. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. *F1000Research*. 2020;9:304.
16. Salmela L, Rivals E. LoRDEC: Accurate and efficient long read error correction. *Bioinformatics*. 2014;30(24):3506–14.
17. Yao J, Qin Y, Wu DC, Nottingham RM, Lambowitz AM, Hunnicke-Smith S. RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. *Rna*. 2016;22(4):597–613.
18. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. *Nucleic Acids Res*. 2018;46(D1):D754–61.
19. Cartolano M, Huettel B, Hartwig B, Reinhardt R, Schneeberger K. cDNA library enrichment of full length transcripts for SMRT long read sequencing. *PLoS One*. 2016;11(6):1–10. <https://doi.org/10.1371/journal.pone.0157779>.
20. Hu R, Sun G, Sun X. LSCplus: A fast solution for improving long read accuracy by short read alignment. *BMC Bioinformatics*. 2016;17(1):1–9. <https://doi.org/10.1186/s12859-016-1316-y>.
21. Holmes I, Durbin R. Dynamic programming alignment accuracy. *J Comput Biol*. 1998;5(3):493–504.
22. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. *Nature*. 2012;489(7414):101–8 Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3684276&tool=pmcentrez&rendertype=abstract>. Cited 2014 Jul 9.
23. Sahlin K, Tomaszewicz M, Makova KD, Medvedev P. Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. *Nat Commun*. 2018;1–12. <https://doi.org/10.1038/s41467-018-06910-x>.
24. 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).
25. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. *Nucleic Acids Res*. 2014;43(October 2014):662–9 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25352552>. Cited 2014 Nov 25.
26. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792–7.
27. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res*. 2013;41(6):e74 Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3616698&tool=pmcentrez&rendertype=abstract>. Cited 2015 Feb 19.
28. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res*. 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>.
29. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*. 2020; 21(1):1–16.
30. Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094–100.
31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
32. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2.
33. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. *BMC Bioinformatics*. 2009;10:1–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



3.3 Discussion

In this study, I wanted to illustrate three main aspects of long read RNA sequencing. The first aspect was to look at what is the top limit in terms of producing long read RNA sequencing data. I felt that the use of a very standardized and well documented sample such as the UHRR in conjunction with the most accurate long read RNA sequencing method truly created the limit of expectation. Basically, all other studies should probably not expect to generate data that has better characteristics with respect to RNA degradation, sample processing, and sequencing. With this set as an assumption, I felt it was useful to dive into the data processing to see what issues and uncertainties still remained. Thus the second aspect I wanted to cover is the complexity that we still see even after we have established such a high bar of sample quality, processing, and sequencing technology. The third aspect that I cover in this article is the amount of variability in final results that can be produced by simple changes in the data processing methodology. I wanted to explain how small changes in the algorithms can lead to large differences in total numbers of transcripts in the final annotation as well as how these models present. Overall, these three aspects were meant to give a foundation for what types of assumptions we could make and how to assess the choices we make in data processing.

One major issue that I see with current studies using long read RNA sequencing from transcriptome annotation is the lack of understanding of how the data processing pipelines bias the final outcome. This is especially problematic when special characteristics of either PacBio or Nanopore sequencing are not taken into account. For example, there were studies that utilized Nanopore direct RNA sequencing for the purpose of general transcript discovery and annotation despite this method having major disadvantages to cDNA based methods including higher error rate and issues with RNA structure interference with Nanopore sequencing[130]. In many cases, Nanopore long read RNA sequencing data is often mapped to the genome without removing adapters or poly-A tails[131]. Leaving the adapters and poly-A tail sequences

as part of the read sequence to be mapped results in poorer mapping performance and the potential for mapping to a completely different loci.

There are still efforts to use short read data to error correct long read data despite major flaws in this general approach[132]. The inherent issue with using short read data to correct long read data is that there is a compromise between rescuing low quality reads and erroneously “correcting” reads. Essentially, if the short read aligns well to a long read then the number of errors will be minimal. However, if a short read has significant differences in sequence to the aligned long read then it could be due to improper alignment. Despite providing proof of this inter-read alignment issue, many studies still use pipelines which include the use of clustering algorithms for error correction[133][86]. This is most likely due to the fact that this method is part of the official PacBio Iso-Seq software package.

I also wanted to highlight that many of the criticisms that people have of long read RNA sequencing and the results that have been generated so far are based on a lack of understanding key underlying principles of sample processing that were essentially hidden by the lack of clarity that comes with short read RNA sequencing. There were many issues that plagued short read methods that were simply overlooked because the indicators for these issues were not visible given the limitation of short reads. However, the perception seems to be changing with more articles supporting the use of long read RNA sequencing. This includes advocating the use of long read RNA sequencing for medical genetics[134] and cancer research[135].

By using a dataset that should represent high quality RNA and state of the art PacBio sequencing, I found different methods for processing the sequencing reads to produce optimal compromises between sensitivity and specificity of transcript model detection. I found that there are ways to filter signal from noise without as much reliance on read coverage. I also explore the concept of wobble in determining true splice junctions in the predicted transcript

models. The TAMA software is optimized for the pursuit of low abundance novel RNA sequences and thus can be used to improve lncRNA discovery.

Chapter 4: Identifying lncRNA in the chicken transcriptome

4.1 Introduction and aims

In the previous results chapters, I focused on the development of methods and pipelines to enhance the use of long read RNA sequencing for lncRNA discovery. While I performed some preliminary analysis to gain some insight into lncRNA, I did not perform in depth analysis of the lncRNA predictions. In this chapter, I have added further Iso-Seq data from different tissue types. These include spleen, macrophage, testes, and ovaries. These new datasets are predominantly comprised of 5' cap selected and normalized cDNA libraries. The cDNA normalization improves the sampling of low abundance genes. Since lncRNA are typically lowly expressed, this technique should allow for the detection of more lncRNA. The 5' cap selection was performed to increase the proportion of reads representing full-length RNA sequences. Increasing the likelihood of picking up full-length reads, especially from spliced transcripts, reduces the representation of genomic contamination and makes it easier to identify real transcript models. This is especially true for the elusive sense exonic class of lncRNA. The logic behind this is that if we have more assurance that the reads are full length, we can rule out the possibility that sense exonic models are not simply the fragments of an incompletely processed RNA from that region.

The addition of the other tissue types adds to the diversity of the sequencing data and allows for more cross sample comparison. This makes it easier to see which lncRNA are shared across tissues and which seem to be more tissue specific. These tissues were selected based on their expected transcriptome complexity.

I also incorporate short read RNA sequencing for gene level quantification to provide additional evidence of lncRNA functionality. The short read data comes from two lines of chicken (J line and broiler) with 21 different tissue types from each line. These data provide a comprehensive snapshot of RNA quantification across the whole chicken transcriptome with respect to tissue

types and breeds. In this analysis, the Iso-Seq transcriptome is used as a reference to improve the quantification accuracy for lncRNA from the short read data. This approach essentially combines the strengths of long read and short read technology.

4.2 Methods

4.2.1 Creating cDNA libraries for Iso-Seq

We selected 6 tissue/sample types based on transcript diversity: Brain, embryo, spleen, macrophage, testes, and ovaries. The brain tissue was collected from an adult J-Line chicken (brown leghorn) bred at the Edinburgh Poultry Research Centre. The embryos were at Hamburger-Hamilton stage 26 and from an ISA Brown chicken bred at the Edinburgh Poultry Research Centre. The spleen and ovaries came from an adult broiler female chicken. The testes came from an adult male broiler chicken. The macrophages came from LPS+ and LPS- cell lines.

The embryo, spleen, macrophage, testes, and ovaries RNA were 5' cap selected and normalized via the hydroxyapatite column method to reduce over abundant transcripts. The brain tissue Iso-Seq data was generated from different cDNA library preparation strategies. The first method of preparation involved a column based cDNA normalization. The second method of preparation included 5' cap selection using the Teloprime kit from Lexogen but had no normalization. The third method also included the 5' cap selection method and involved duplex specific nuclease (DSNase) normalization that was performed using the Trimmer-2 kit from Evrogen.

The embryo and column normalized brain samples were sequenced on PacBio RSII machines while the spleen, macrophage, testes, ovary, non-normalized brain, and DSNase normalized brain samples were sequenced on PacBio Sequel machines.

4.2.2 Processing Iso-Seq data

The Iso-Seq data was processed using Iso-Seq2 software developed by PacBio and TAMA. Iso-Seq2 CCS was used for Circular Consensus Sequence (CCS) calling. Iso-Seq2 Classify was used for adapter removal, poly-A removal, non-full length (NFL) read removal, and artificial concatemer removal. Iso-Seq2 Cluster was used to error correct reads by clustering reads via sequence alignment and performing inter-read alignment correction. GMAP was used to map cluster sequences to the Gallus_gallus-5.0 genome assembly (GCA_000002315.3). TAMA collapse was used to create unique transcript models based on the reference sequence. For the 5' cap selected libraries TAMA Collapse with these parameters were used “-x capped -a 100 -z 100”. For the libraries that did not undergo 5' cap selection these parameters were used “-x no_cap -a 100 -z 100”.. TAMA merge was used to merge the TAMA collapse results from each sample using the parameters “-a 100 -z 100”.

4.2.3 Short read RNA sequencing

The short read RNA-seq was performed on two different sets of samples. The first set of samples were collected from 9 16/17 weeks old female J-line chickens and included 21 tissue types (Table 4.4). Sequencing was performed using Illumina Total RNA Stranded kit generating pair-end, stranded reads. Samples were pooled from the 9 chickens for each tissue type. The second set of samples were collected from adult broiler chickens and included 21 different tissue types (Table 4.4). Sequencing was performed using Illumina Total RNA Stranded kit generating pair-end, stranded reads.

4.2.4 Short read RNA sequencing data processing

The first step in processing the short read RNA-seq data was to run FastQC (Babraham Institute, Cambridge, UK). The reads were then trimmed using Cutadapt [111]. After trimming the reads were checked again using FastQC to confirm that trimming had been performed optimally. The reads were then pseudo-mapped to the Iso-Seq based transcriptome using Kallisto [136].

4.2.5 Coding predictions and lncRNA classification

I used three different methods to estimate coding potential: CPAT [137], CPC2 [138], and BLASTX [139] of transcript sequences against the Uniref90 protein database [140]. CPAT and CPC2 were run using default parameters. The predicted lncRNA were then divided into positional classes using `tama_ncrna_classification_positional.py`.

4.2.6 Comparing Iso-Seq annotation to Ensembl

The merged Iso-Seq annotation from the 6 different samples were compared with the Ensembl v87 chicken annotation using TAMA Merge with these parameters “-a 100 -z 100”. This produced a set of files that show the transcript models that are shared between the annotations and those that are different.

4.3 Results and Discussion

4.3.1 Samples, sequencing, and raw data processing

Six tissues were selected based on transcript diversity: brain, embryo, spleen, macrophage, testes, and ovaries. Three different library preparation methods on brain tissue to test the effects of cDNA normalization and 5' cap selection on transcriptome coverage and sequenced the libraries on PacBio platform as follows. The first was prepared with column cDNA normalization and sequencing on the PacBio RSII machines. The second was prepared using Lexogen Teloprime 5' cap selection and no normalization and sequenced on a PacBio Sequel machine. The third was prepared using Teloprime 5' cap selection and Trimmer-2 DSNase cDNA normalization and sequenced on a PacBio Sequel machine.

The embryo, spleen, macrophage, testes, and ovaries RNA were 5' cap selected by GATC Biotech (Konstanz, Germany) and normalized via the hydroxyapatite column method to reduce over abundant transcripts. These samples were sequenced on the RSII machines.

There were large differences in the total number of reads attained for each tissue type due to differences in the number of sequencing runs per sample, loading efficiency of SMRT cells, and the version of sequencing platform used.

It is also important to understand the relationship between the different stages of read processing. While the terminology has changed since the inception of PacBio sequencing I have defined these terms in line with current usage. The raw reads that are produced are typically called polymerase reads. These represent the full-length read that is generated from each Zero Mode Wave guide (ZMW). Within the PacBio sequencing software these are then separated into their individual passes of the initial template. This form of the read representation is called subreads. The subreads for each polymerase read are then multiple sequence aligned to generate an intra-molecule error corrected read that is known as the circular consensus read (CCS). The CCS read is the first step where the original transcript/template is represented as a single sequence. However, the CCS reads still contain all adapter sequences as well as the poly-A tail of the original RNA. Thus the CCS reads are further processed to remove the adapters/poly-A tails and orient the reads from 5' to 3'. This processing step also removes reads that do not contain both adapters or contain a combination of adapters that would suggest the reads are actually comprised of concatemerization of multiple templates. Typically reads which do not contain both 5' and 3' adapters are not full length. The resulting sequences from this step are referred to as full length non-chimeric (FLNC) reads. The FLNC reads represent the original sequence of the RNA without any adapters or poly-A tails and thus can be directly mapped to the genome.

Through each processing step, the number of reads remaining diminishes as the unusable read are removed. Thus, the number of FLNC reads is typically the most useful metric to identify as

there can be significant numbers of discarded reads prior to this stage. Usually the most significant drop in reads occurs between the CCS and FLNC stage. The ratio of removed reads between CCS and FLNC depends on the sequencing library preparation and the relationship between the polymerase read lengths and the template reads lengths. For example, if a polymerase read were 2 kb in length for a 5 kb template, then the read would not represent the full length of the original RNA and thus be discarded.

The dataset had a range of between 145,527 (Cap Norm Brain) CCS reads to 805,606 (original brain) CCS reads per sample and a range of 124,620 (Cap Norm Brain) FLNC reads to 515,175 (original brain) FLNC reads per sample (Table 4.1). In general, a reduction of between 14% (Cap Norm Brain) and 52% (Ovary) was observed. Note that Table 4.1 lists the samples in chronological order of when they were sequenced which also correlates to improvements in read lengths for the PacBio platforms. Thus, the ratio of discarded reads in general decreases due to longer read lengths resulting in fewer non-full length reads.

Table 4.1 : Summary of reads, genes, and transcripts from the Iso-Seq data

Tissue	CCS	FLNC	Mapped	Genes	Transcripts	Platform	cDNA Prep
Brain	805,606	515,175	499,517	28,021	78,233	RSII	Col. Norm.
Embryo	247,626	138,266	130,117	5,598	12,410	RSII	5' Cap & Col. Norm.
Ovary	310,868	148,832	135,107	21,567	78,562	RSII	5' Cap & Col. Norm.
Testes	316,947	152,263	140,092	25,650	80,305	RSII	5' Cap & Col. Norm.
Spleen	309,652	210,576	97,452	10,208	32,063	RSII	5' Cap & Col. Norm.
Macrophage	230,498	153,611	140,953	9,095	37,736	RSII	5' Cap & Col. Norm.
No Norm							
Brain	566,307	422,163	390,788	11,934	39,909	Sequel	5' Cap Telo.

Cap Norm							5' Cap Telo. &
Brain	145,527	124,620	117,347	19,849	49,465	Sequel	DSNase Norm.

4.3.2 Identifying Gene and Transcript Models in Gal5

Most of the analyses were performed using the *Gallus_gallus*-5.0 genome assembly (GCA_000002315.3). While this is not the most up-to-date assembly, this assembly was chosen since later assemblies were released with annotations from Ensembl that used our Iso-Seq data set. Thus in order to compare the gene models from these Iso-Seq data to the public annotation it was necessary to use a version that did not include these data. However, we also performed analysis on the Gal6 genome assembly to compare the annotation methods as opposed to the data type used.

The FLNC and Cluster reads were mapped to the *Gallus_gallus*-5.0 genome assembly using GMAP [141]. The resulting BAM files were processed using TAMA Collapse [129] to form non-redundant transcriptome annotations for each sample. The sample annotations were then merged into a single annotation using TAMA Merge [129].

The highest number of predicted genes was observed in the non-5'-cap selected normalized brain with 28,021 genes. However, this sample also benefitted from the highest read coverage. The embryo was the sample with the lowest number of predicted genes at 5,598. This is likely due to the relatively low number of reads generated and possibly a reduction in RNA complexity from the 5' cap selection method (tobacco acid pyrophosphatase) used.

Comparing the two 5' cap selected brain libraries it can be seen that sampling efficiency for gene discovery is much higher for the normalized library. The 5' cap selected normalized library

allowed the detection of 159 unique genes per 1000 FLNC reads while the non-normalized library revealed only 28 unique genes per 1000 FLNC reads.

However, using the DSNase approach for normalization resulted in a strange phenomenon where the top ranking genes in terms of expression levels seemed to be missed. This seems counter intuitive given the proposed mechanism of normalization where highly expressed genes should not be reduced to levels lower than the lowly expressed genes. However, it may be that there is a mechanism by which this can occur in the DSNase method of normalization.

After merging the transcript models from all samples, we identified 52,519 gene loci that corresponded to 300,494 transcripts.

4.3.3 Coding RNA analyses

Three different methods for identifying evidence for coding potential were used: CPAT [137], CPC2 [138], and BLASTX [139] of transcript sequences against the Uniref90 protein database [140]. CPAT infers coding potential purely from assessing potential open reading frames and looking for codon-like patterns that fit into protein coding model. CPC2 also uses open reading frame analysis but uses different methods of analyzing coding potential. The BLASTX approach allows for the direct comparison of each ORF to the largest database of known protein sequences. The idea behind these three methods is to use orthogonal methods for picking up any potential for encoding protein within the predicted transcript models. Given that the primary goal was to identify lncRNA a transcript was classified as protein coding if it showed evidence from any one of these methods.

Using these metrics, 26,064 gene loci were identified that had protein-coding evidence from at least one of these prediction software tools. These gene loci contain 187,272 predicted isoforms with evidence of protein coding potential. Of the total predicated protein coding gene loci, 18,763 contained multi-exonic transcript models. The remaining 7,301 gene loci were

comprised solely of mono-exonic transcript models. Thus, these may be processed pseudogenes as opposed to functional protein coding genes.

4.3.4 Non-coding RNA pipeline

Long non-coding RNA were identified by looking for coding potential with three different methods as described above: CPC2, CPAT, and BLASTX of transcript sequences against the Uniref90 protein database. If a transcript had no evidence for coding potential from any of these methods, it was classified as a non-coding RNA. CPC2 and CPAT use sequence motif based methods, however BLASTX with Uniref90 provides information on direct hits to known protein coding genes. So if a transcript has a large enough overlap with a coding region, a Uniref90 hit will likely occur for that region thus providing coding evidence for that transcripts. However, the transcript would not need to contain a viable ORF in order for this to occur. Thus, the prediction pipeline is designed to be more specific and less sensitive for identifying lncRNA.

Analyses of the datasets predicted 113,222 lncRNA transcripts corresponding to 38,828 gene loci. However, this classification of genes is based on a group of transcripts with sense exonic overlap of at least 1 base pair. Thus by this definition multiple lncRNA transcripts can be from the same gene even though they do not have any overlap because of overlapping coding transcripts. If the same gene definition is used but overlapping lncRNA are considered, then 42,709 lncRNA genes are predicted. The additional lncRNA genes comes from sense exonic lncRNA that overlap the same protein-coding gene but do not overlap each other.

4.3.5 Comparison of all genes to the Ensembl annotation

The annotation generated from the combined Iso-Seq data as described above was compared to the Ensembl Release 87 (December 2016) of the Gallus_gallus-5.0 (GCA_000002315.3)

assembly using TAMA Merge. Version 87 was the last version before the incorporation of our Iso-Seq data into the public annotation and thus later Ensembl Releases could not be used in order to assess the benefits of incorporating the long-read Iso-Seq data into the annotation pipelines.

TAMA Merge defines gene loci by grouping transcript models on the same strand where if two transcript models overlap by at least 1 base pair they are combined into the same gene locus. By this definition, 14,541 genes were found which overlapped between the Ensembl annotation and the Iso-Seq annotation. 8,562 genes were found only in the Ensembl annotation and 36,184 genes were found only in the Iso-Seq annotation. Of the total Iso-Seq only genes, 7,129 were found in multiple samples. Thus, the overwhelming majority of Iso-Seq only genes (29,055) were only found in a single sample type.

There were 11,431 multi-exonic genes found only in the Iso-Seq annotation. These contained 33,203 multi-exonic transcript models.

At the transcript level, there were 12,522 transcript models that matched between the Ensembl annotation and the Iso-Seq annotation. Matching transcripts are defined by having the same exon structure and no more than 100 bp difference for the transcription start site and end site. For the genes with overlap between the two annotations, 14,684 transcript models were found only in Ensembl while 216,113 were found only in the Iso-Seq annotation. This result represents a massive expansion of transcript models over the Ensembl annotation. Of these transcript models, 158,195 were multi-exonic. The mono-exonic models have lower evidence of being real transcripts.

Positional classification of long non-coding RNA

The predicted lncRNA were then classified by their positional relationship to the nearest protein coding gene. This includes intergenic (lincRNA), intronic, antisense, and sense exonic (Figure 4.1). I also looked at another positional class not previously investigated before which I call lncRNA to lncRNA antisense (Figure 4.2). This class refers to lncRNA that are antisense exonic to another lncRNA.

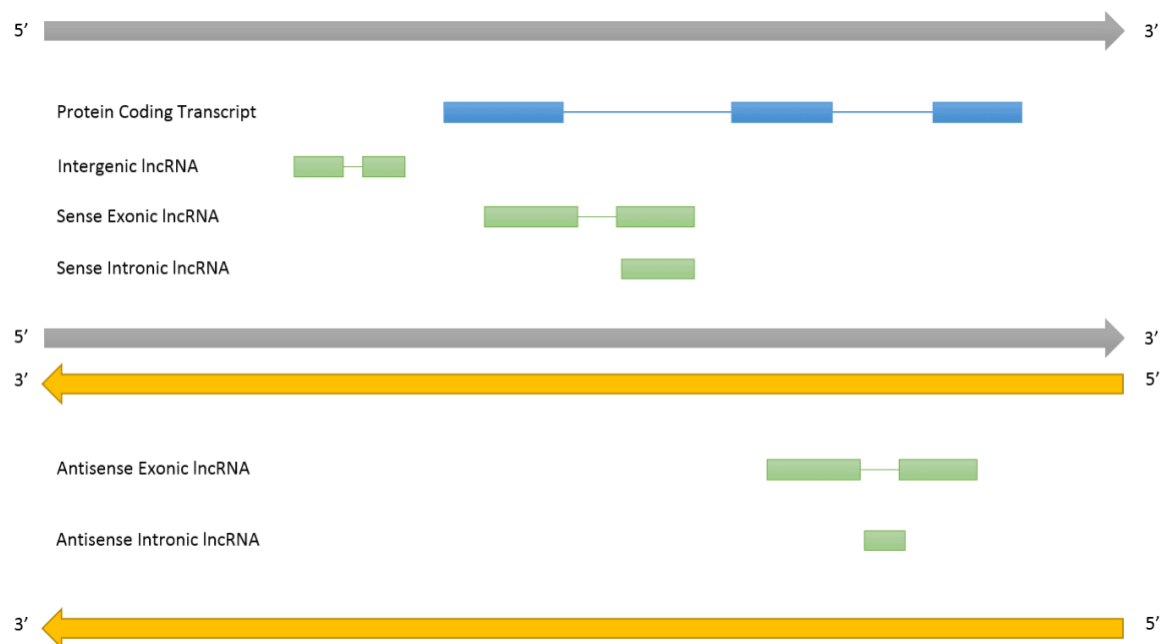


Figure 4.1 : Diagram of standard long non-coding RNA classes. These include intergenic, sense exonic, sense intronic, antisense exonic, and antisense intronic lncRNA as shown in the diagram. The forward strand of the genome is depicted by the region between the grey arrows while the reverse strand is the region between the yellow arrows.

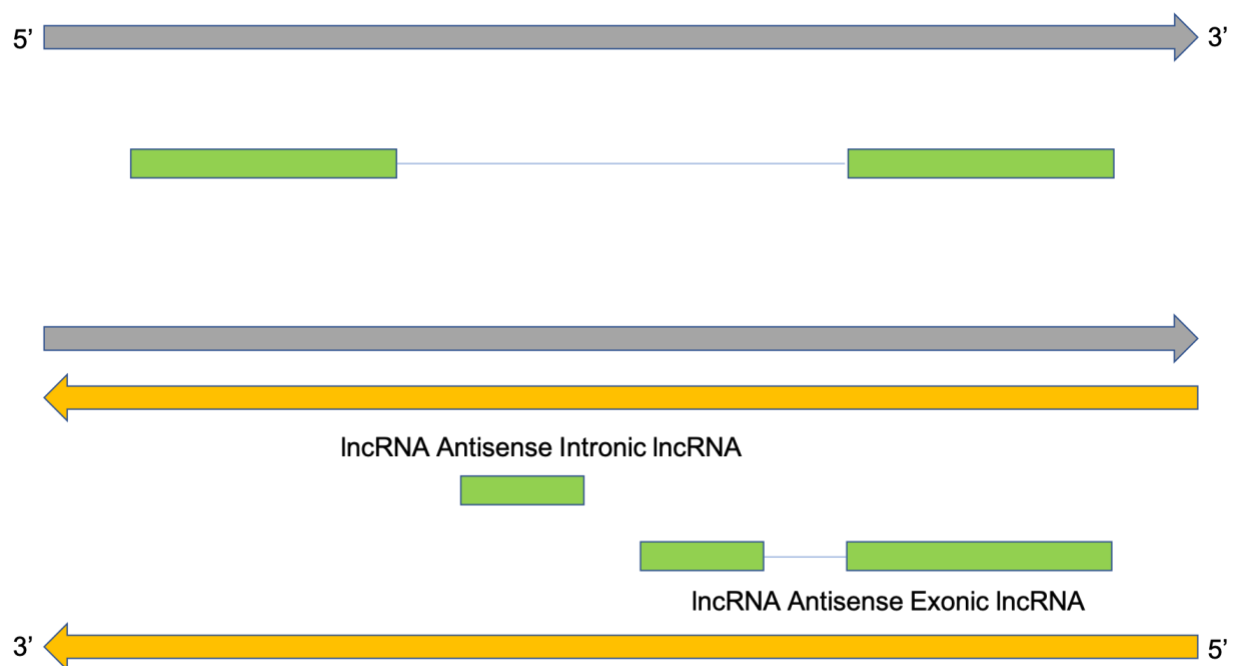


Figure 4.2 : Diagram of lncRNA classes that are based on relationship to nearest lncRNA. These include lncRNA antisense exonic lncRNA and lncRNA antisense intronic lncRNA. The forward strand of the genome is depicted by the region between the grey arrows while the reverse strand is the region between the yellow arrows.

Table 4.2 : Number of predicted long non-coding RNA by positional class from the Iso-Seq annotation

Positional Class	Number of Transcripts
LincRNA	33,625
Sense Exonic	68,025
Sense Intronic	23,290
Antisense Exonic	22,542
Antisense Intronic	13,085
LncRNA Antisense Exonic	16,531

Sense exonic lncRNA represented the positional class with the most transcripts (68,025). While lincRNA was the second most abundant with 33,625 transcripts (Table 4.2). This is in stark contrast to public annotations for the human and mouse genome annotations (Ensembl v96) where lincRNA represent the most abundant class with 7,690 lincRNA for human and 5,854 lincRNA for mouse. Ensembl v96 is the last annotation version to separate out the lincRNA positional class. These numbers were calculated by looking at the gene biotype information included in the Ensembl gtf format annotation files. However, there are many possible causes for this high number of sense exonic lncRNA. The first explanation is that these are simply errors in predictions caused by the confounding factor of the overlapping protein-coding gene. In essence, these sense exonic lncRNA could represent unprocessed RNA from the protein-coding gene, fragmented RNA, or noisy transcription. It is nearly impossible to distinguish, in a high throughput manner, the true source of these models given the limitations of current technology. Thus, it would require targeted experimentation using biochemical methods to either prove or disprove the existence and function of these seemingly novel transcripts.

It is also interesting to note that there are more antisense exonic lncRNA (22,542) than antisense intronic lncRNA (13,085). In theory, it should be the other way around with antisense intronic lncRNA being far more common than antisense exonic lncRNA. Since antisense exonic lncRNA have at least a portion of their sequence governed by the antisense gene, one would expect this to be very limiting and thus occur far less frequently. However, it could be that the function is directly tied to this sequential relationship. For example, it has been hypothesized that antisense exonic RNA could be involved in regulating their antisense counterpart by binding to it via the overlapping sequence region or by being incorporated in protein complexes to target binding in those genomic locations. If we assume that there is a larger need for this type of regulation, then it might actually be that the observed frequency of the sub-types of antisense lncRNA are actually more accurate across species. However, the counterpoint would be that antisense exonic lncRNA or rather the signals indicating this type of RNA are often mixed up with sequencing noise caused either by biological transcriptional noise or by simply the wrong interpretation of strand orientation from sequencing read information.

Long read RNA sequencing data are generally more robust to errors in identifying the strand orientation of the reads. This is due to the ability to see the 3' end sequence that is expected to contain a poly-A tail and have a direction connection with the 5' end of the RNA molecule. Thus, if the poly-A tail is observed on the 3' end and the correct adapter configuration on both ends, then there is a stronger case for having identified the correct strand orientation. Thus, I believe that these results provide a fair amount of evidence in support of the relatively high abundance of antisense exonic lncRNA.

4.3.6 Single exon long non-coding RNA

Mono-exonic transcript models are often considered suspicious given that they could be the result of sequencing reads that represent RNA/cDNA truncation or genomic contamination. To investigate the validity of these models in general, I looked at the distribution for the number of exons for each lncRNA positional class. For all classes of lncRNA, single exon transcripts were the most abundant (Figure 4.3). This contrasts with the current public annotations for the human genome annotation (Ensembl 106) in which 2 exon transcripts are the most dominant. Also, sense exonic lncRNA have the highest numbers of mono-exonic models with lincRNA as the second highest. Sense exonic models have the greatest probability of being the result of sequencing noise from the overlapping protein coding gene. LincRNA have the highest probability of being the result of genomic contamination in the sequencing library. Thus, it would seem that the numbers are in favour of the interpretation that many of these models are the result of some type of sequencing noise.

However, I looked at the number of possible internal truncation from 3' stretches of genomic A nucleotides and found that only 5% of the total used reads had possible poly-A truncation.

I also looked at the number of transcript models from each positional class that had support from at least one poly-A truncation possible read (Table 4.3). The percentage of poly-A possible

transcript models from each class ranged from 7% to 26%. Sense intronic had the highest percentage of 26%. However, sense exonic only had 8% and lincRNA had only 12%. These numbers are at odds with the expectation that the majority of these models come from some type of noise.

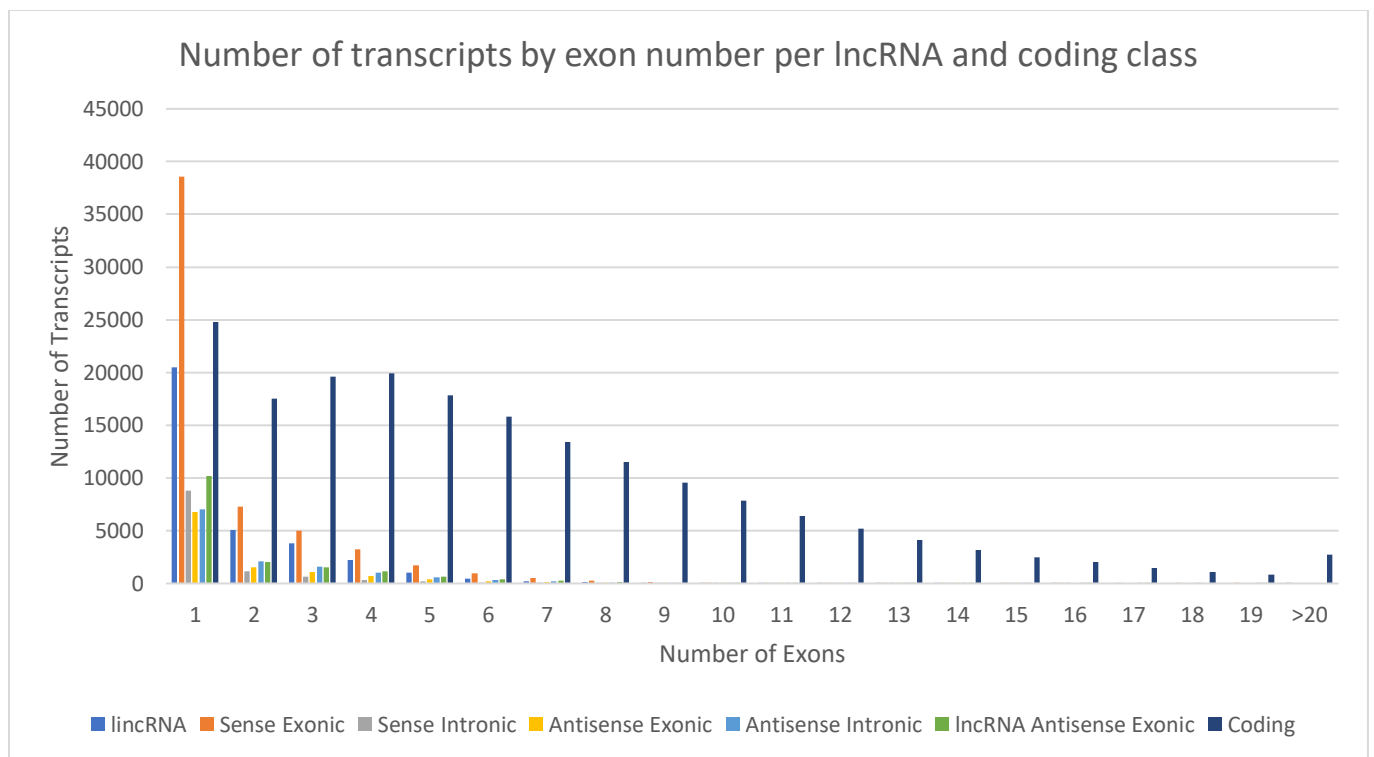


Figure 4.3 : Breakdown of the number of exons per transcript divided by lincRNA class. Each lincRNA class is represented by a unique colour and plotted beside each other for each number of exon bin. Mono-exonic transcripts are the largest group by exon number across all lincRNA classes. Protein coding transcripts are included for comparison.

Table 4.3 : Percentage of transcript models per class with poly-A motifs in the 3' genomic region

LncRNA Class	Poly-A Models	Total Transcripts	Percentage Poly-A
Antisense Exonic	731	11107	7%
Lncrna Antisense Exonic	1215	16531	7%
LincRNA	3913	33625	12%
Antisense Intronic	1322	13085	10%
Sense Intronic	2892	11337	26%
Sense Exonic	4654	57883	8%

4.3.7 Length of lncRNA

I then looked at the length distribution based on number of exons (Figure 4.4) and positional class to see if there were any trends (Figure 4.5). LncRNA across all positional classes and number of exons had peaks within their length distribution around the 1 kb length. There were comparatively few lncRNA transcript models at 200 bp or shorter. Since it is expected that noise would present as shorter models, these length distributions provide evidence in favour of the interpretation that the majority of the lncRNA classes are not from sequencing noise.

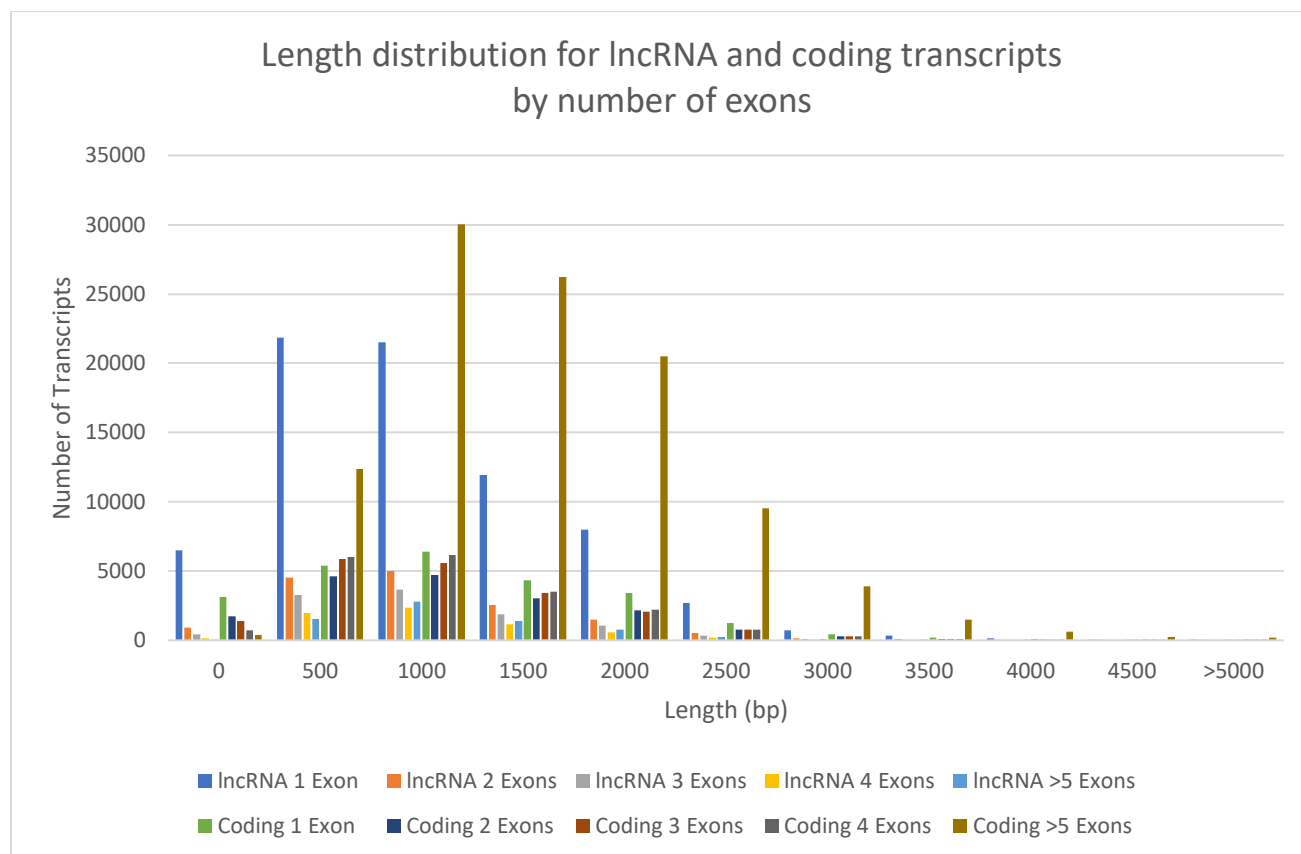


Figure 4.4 : Length distribution of lncRNA when grouping lncRNA by number of exons. The most prominent peak for all exon numbers is around 1000 base pairs. Protein coding transcripts are included for reference.

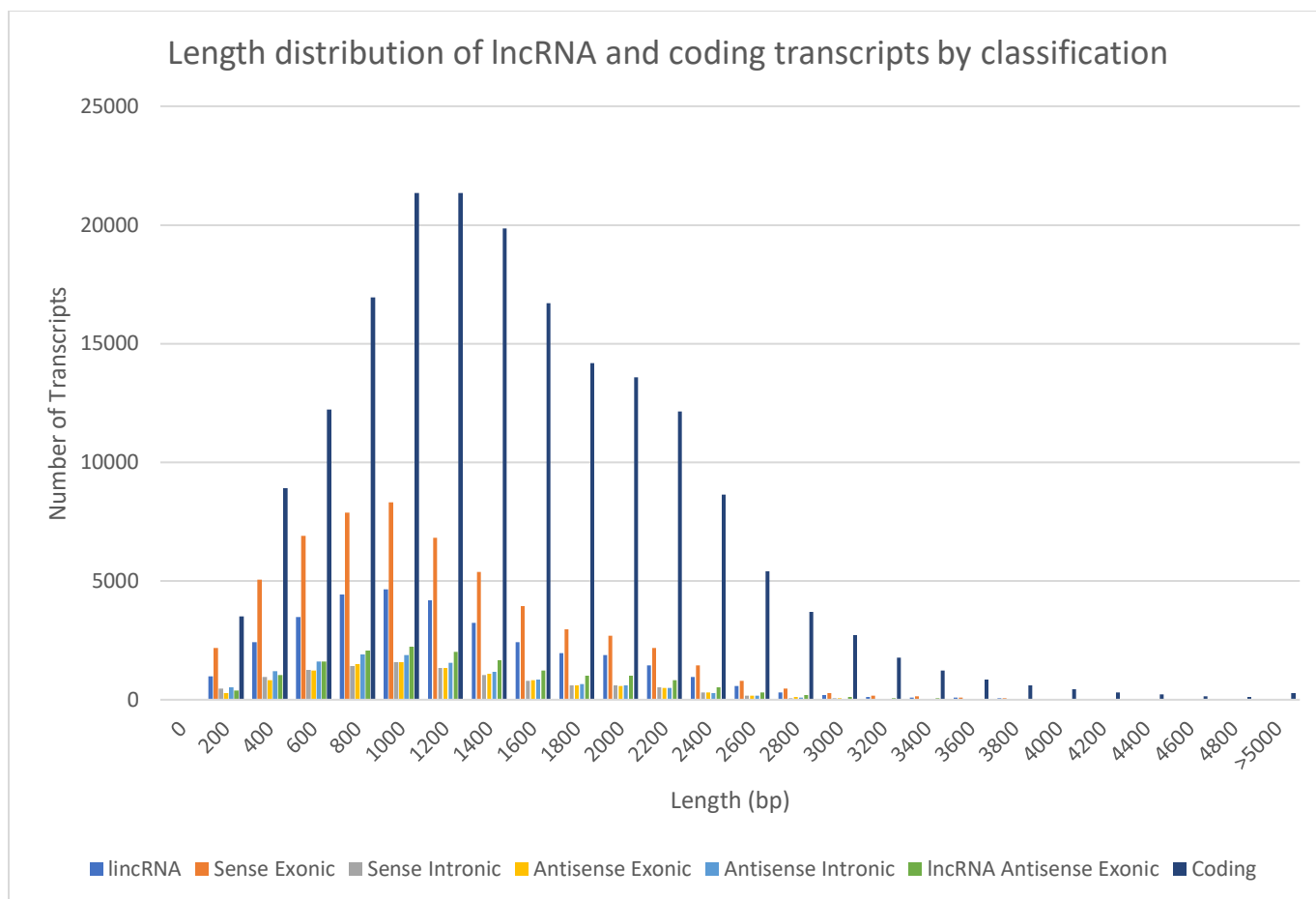


Figure 4.5 : Length distribution of lncRNA when grouping by class. The most prominent peak for all lncRNA classes is around 1000 base pairs. Protein coding transcripts are included for reference.

4.3.8 Additional evidence for lncRNA

For all positional classes of lncRNA, single source support and single read support is dominant. However, out of 113,167 total lncRNA 22,236 had more than 1 source support (Figure 4.6). The relative amount of source support was similar to the protein coding transcripts. Out of the total lncRNA, 39,622 had more than 1 read support (Figure 4.7). Having multiple reads from the same sequencing run that support the same transcript model gives some evidence that these models are biologically real. However, it is possible that they are the result of duplicate artefact sequences from the cDNA library generation. Seeing the same transcript model in different

samples provides a much higher level of evidence for the authenticity of these lncRNA since it is much less likely to see the same models across samples.

It is also interesting to see that there are 15,796 sense exonic lncRNA which are seen in multiple samples. This could indicate a large number of sense exonic lncRNA that are expressed across different tissue types. However, it is possible that these are the result of some kind of common fragmentation/degradation or slow step in the RNA processing for the overlapping gene.

I then looked at separating the lncRNA by number of exons per transcript to see if there were trends for mono-exonic and multi-exonic lncRNA. When I compared the number of long reads supporting lncRNA with different numbers of exons it was interesting to note that the number of mono-exonic lncRNA with read counts of 20 or more dwarfed the other exonic classes. While the sheer number of predicted mono-exonic lncRNA accounts for it dominating in each read count class it was unexpected to see so many with such high read counts. So while the mono-exonic lncRNA with single read support seem very much like they could be the product of some noise, it is harder to explain away the mono-exonic lncRNA which seem to have so much read support.

I also looked at the breakdown of lncRNA prediction focusing on the source sample. It is interesting to note that the ovary and testes samples seemed to dominate for lincRNA (Figure 4.9), sense exonic lncRNA (Figure 4.10), sense intronic lncRNA (figure 4.11), antisense intronic lncRNA (Figure 4.13), and lncRNA antisense exonic lncRNA (Figure 4.14). Even for the antisense exonic class the ovary and testes samples were in the top three sources (Figure 4.12). For the more established lincRNA class the ovary sample had the highest number of predicted lncRNA with the testes coming in second.

When I compared lncRNA class with the supporting source, I found that the brain, testes, and ovary Iso-Seq datasets had both the highest numbers of lncRNA and the highest intersection of lncRNA between them. These organs are known to be very transcriptionally diverse which may

explain why we see such high numbers of lncRNA among them, however, it is fascinating to see the overlap of lncRNA. However, the overlap between any two of the grouping of three (brain, ovary, and testes) shows much high intersecting than between all three. So it may be that this result only indicates the coincidence of overlapping lncRNA when tissues exhibit complex RNA expression.

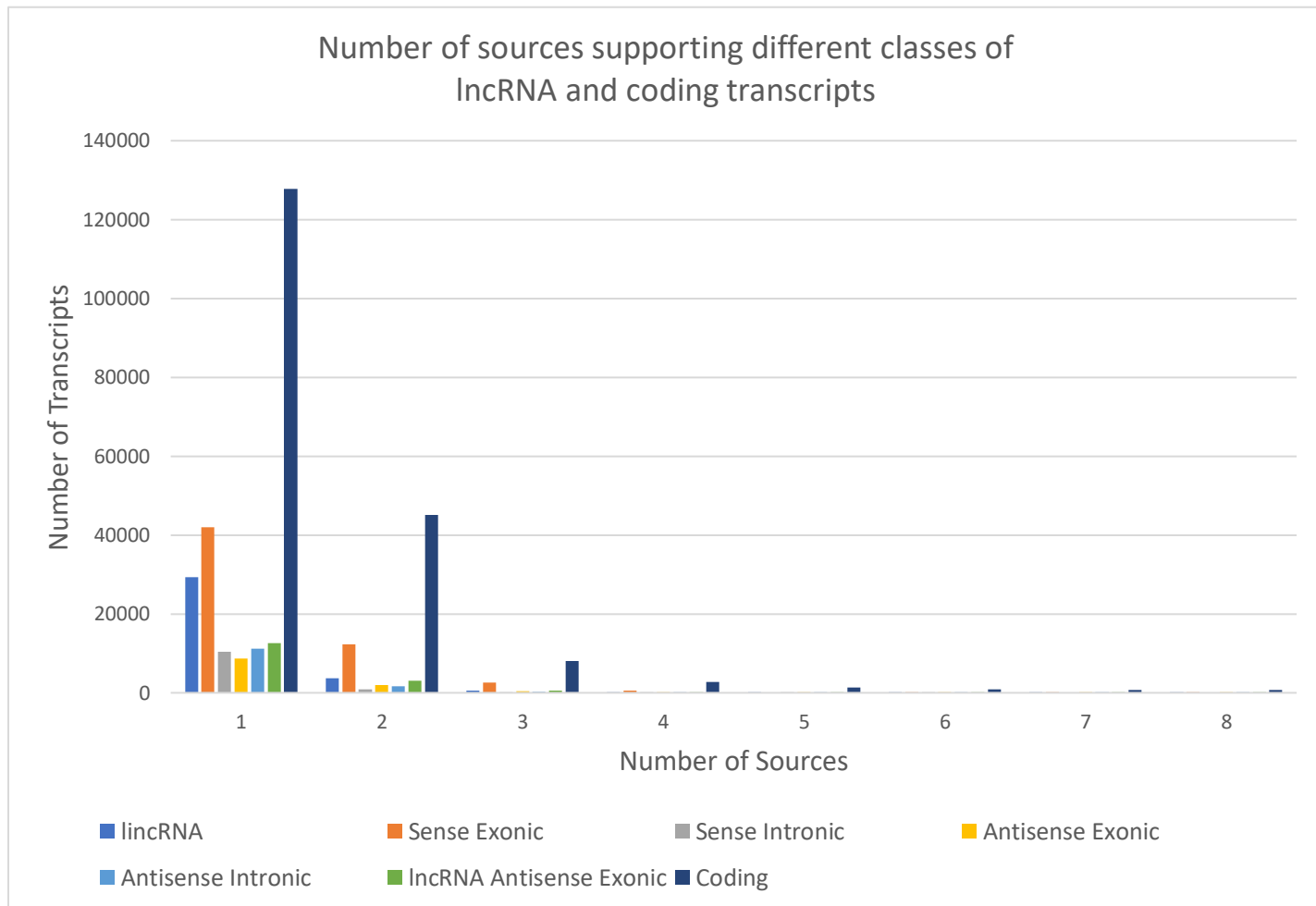


Figure 4.6 : Number of source sample types supporting each lncRNA class and coding transcripts. All classes have the majority of their transcripts supported by only 1 source. Sense exonic lncRNA still have a large number of multiple source supported transcripts.

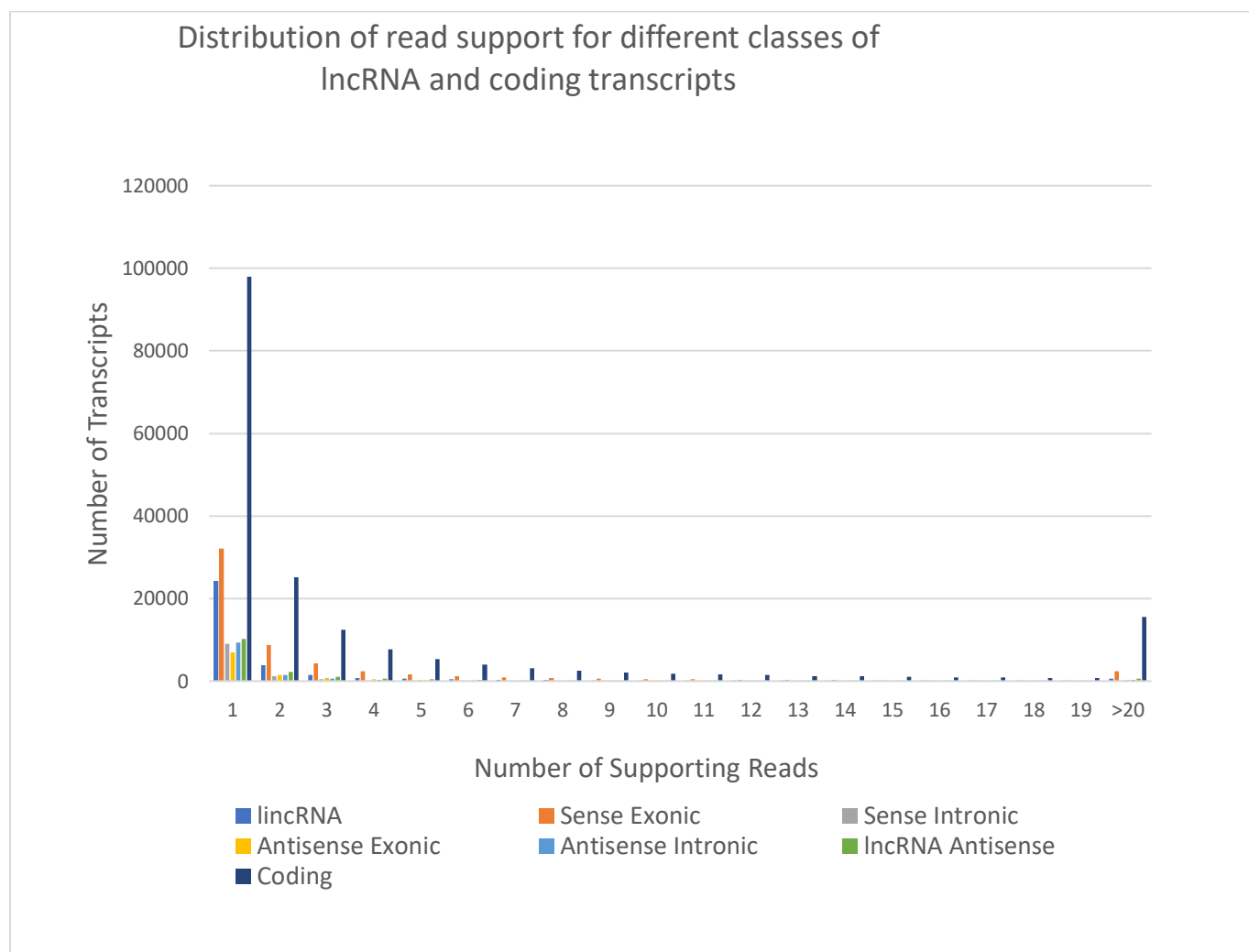


Figure 4.7 : Distribution of the number of transcripts in each lncRNA and coding class. While all lncRNA classes have the majority of their transcripts supported by a single read, sense exonic lncRNA still have over ten thousand transcripts with multiple reads support and thousands of transcripts with over 20 reads support.

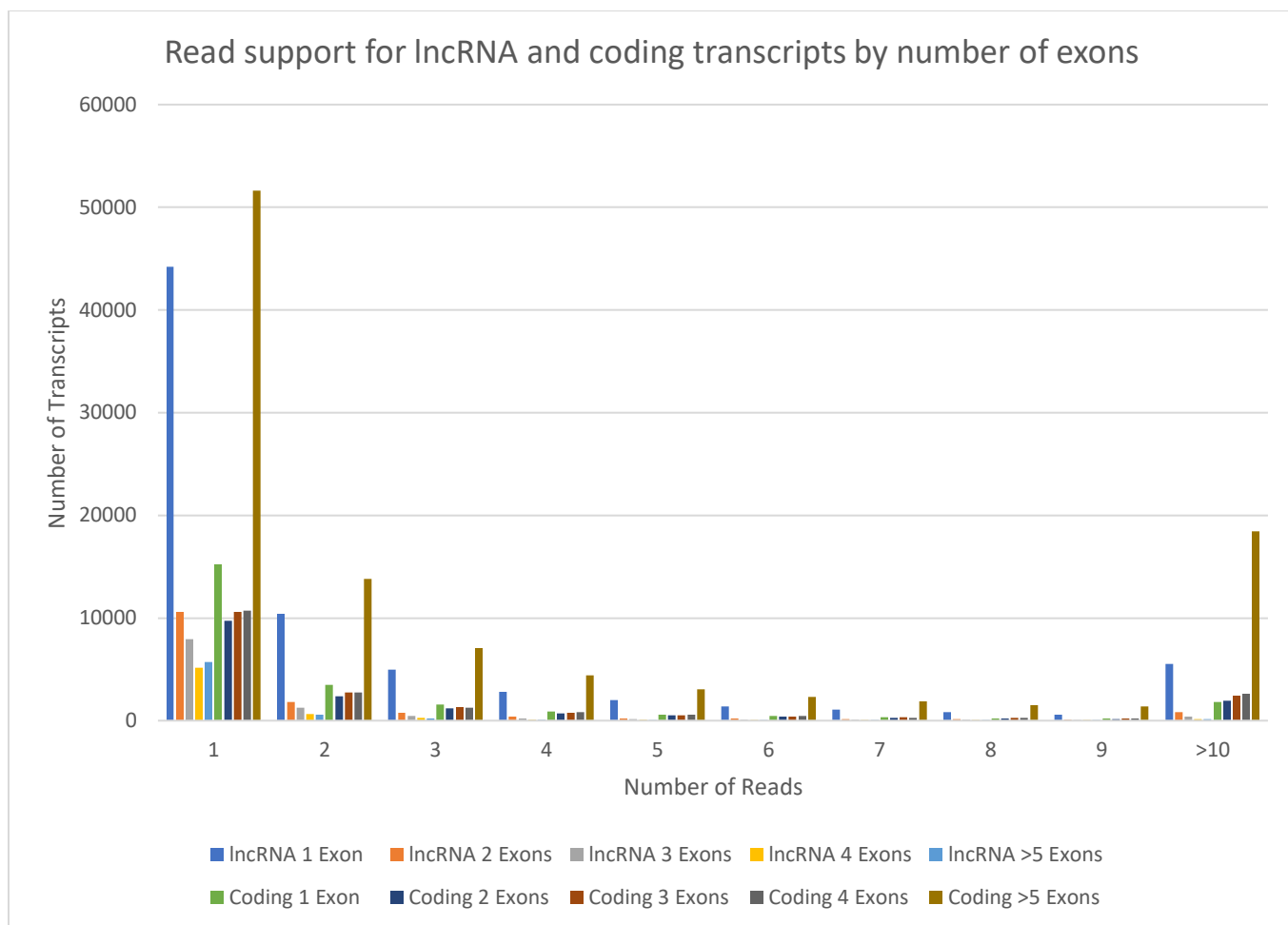


Figure 4.8 : Read counts for lncRNA when grouped by number of exons. Mono-exonic lncRNA have over ten thousand transcripts with multiple reads support. Protein coding transcripts are included for reference.

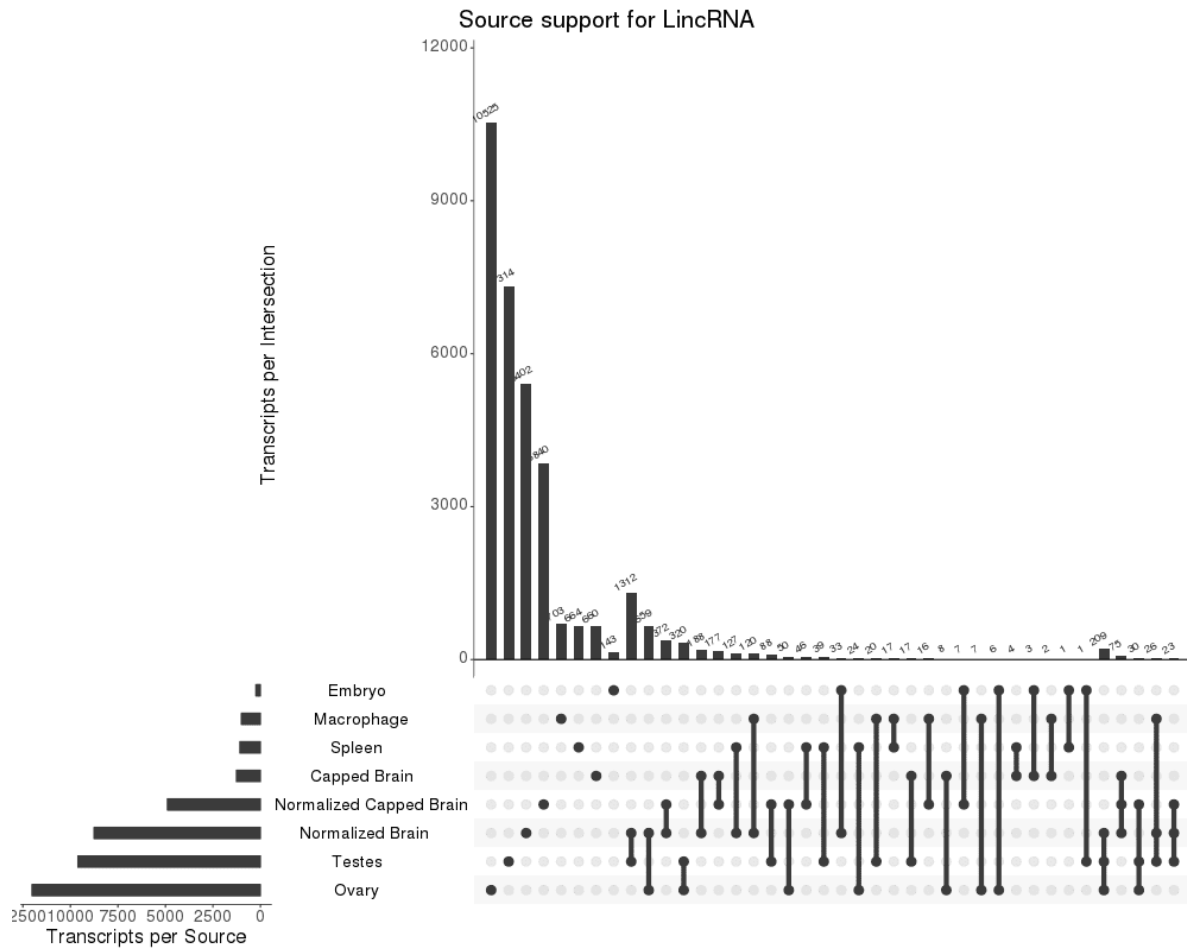


Figure 4.9 : Upset plot showing combination of intersections across the different samples supporting the intergenic lncRNA class. This type of plot allows for higher dimensional representation of intersections across multiple datasets. The relationships are ordered first with single source transcripts and then by intersection grouping with a sub-ordering of high to low numbers of transcripts. The ovary dataset contained the highest number of unique intergenic lncRNA but the largest intersection is between testes and the normalized brain sample.

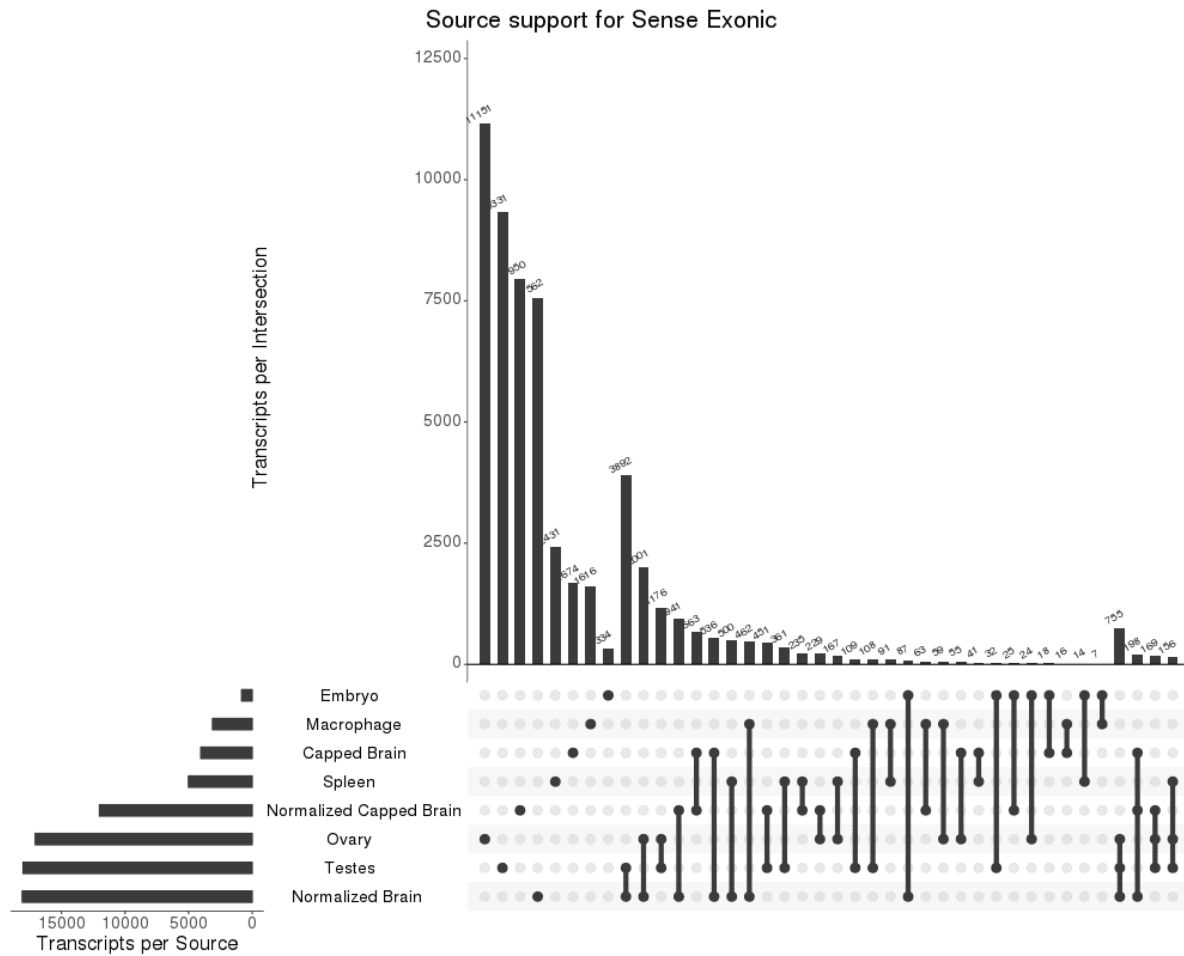


Figure 4.10 : Upset plot showing combination of intersections across the different samples supporting the sense exonic lncRNA class. The ovary dataset contained the highest number of unique sense exonic lncRNA but the largest intersection is between testes and the normalized brain sample.

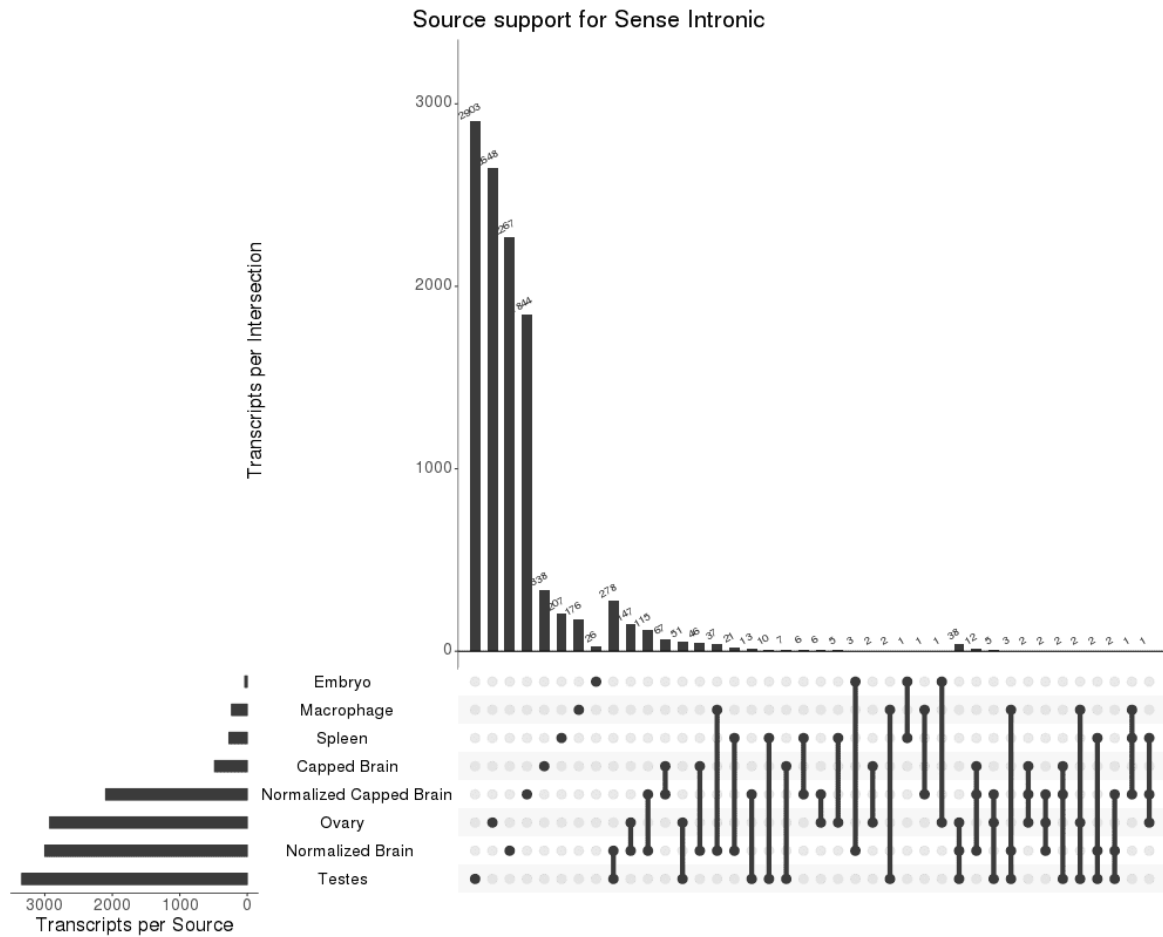


Figure 4.11 : Upset plot showing combination of intersections across the different samples supporting the sense intronic lncRNA class. The testes dataset contained the highest number of unique sense intronic lncRNA and the largest intersection is between testes and the normalized brain sample.

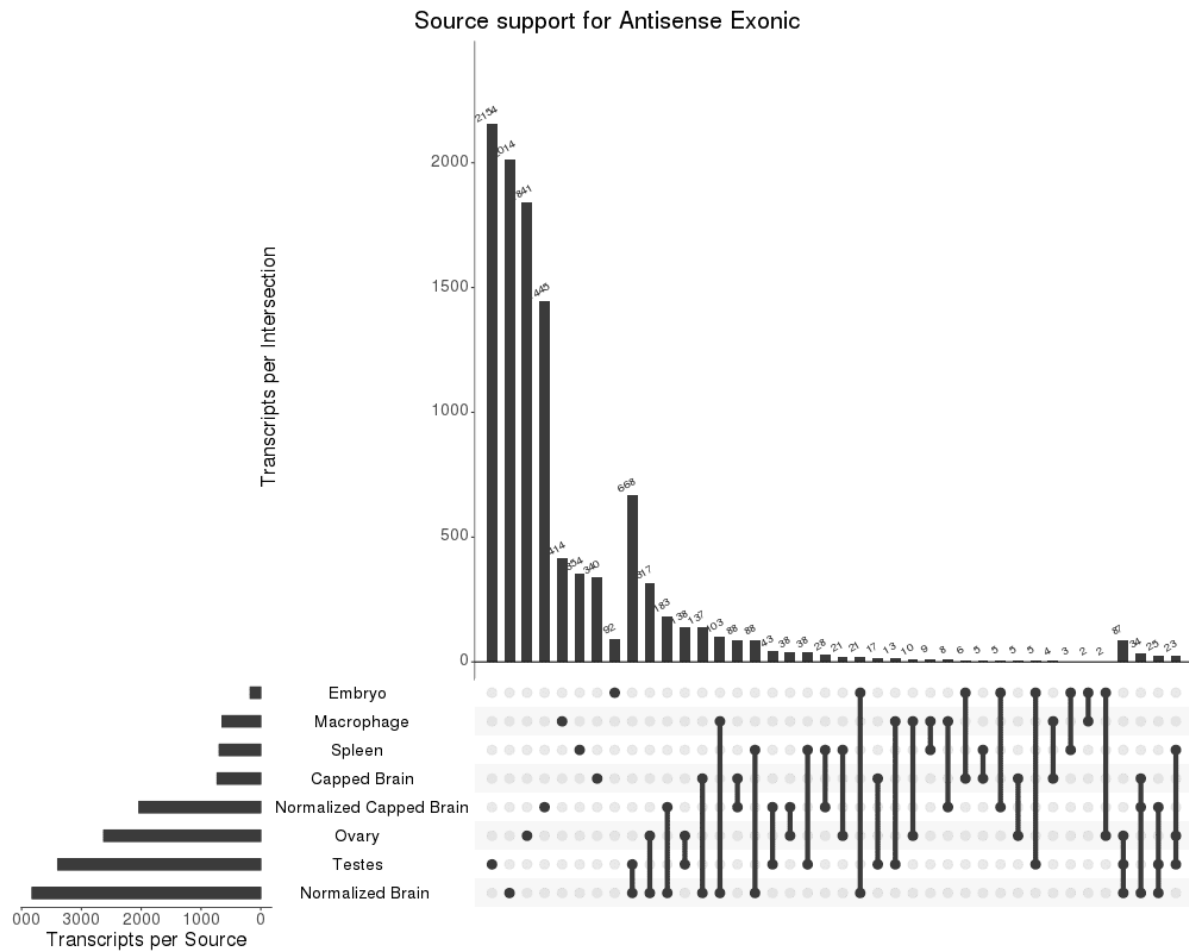


Figure 4.12 : Upset plot showing combination of intersections across the different samples supporting the antisense exonic lncRNA class. The testes dataset contained the highest number of unique antisense exonic lncRNA and the largest intersection is between testes and the normalized brain sample.

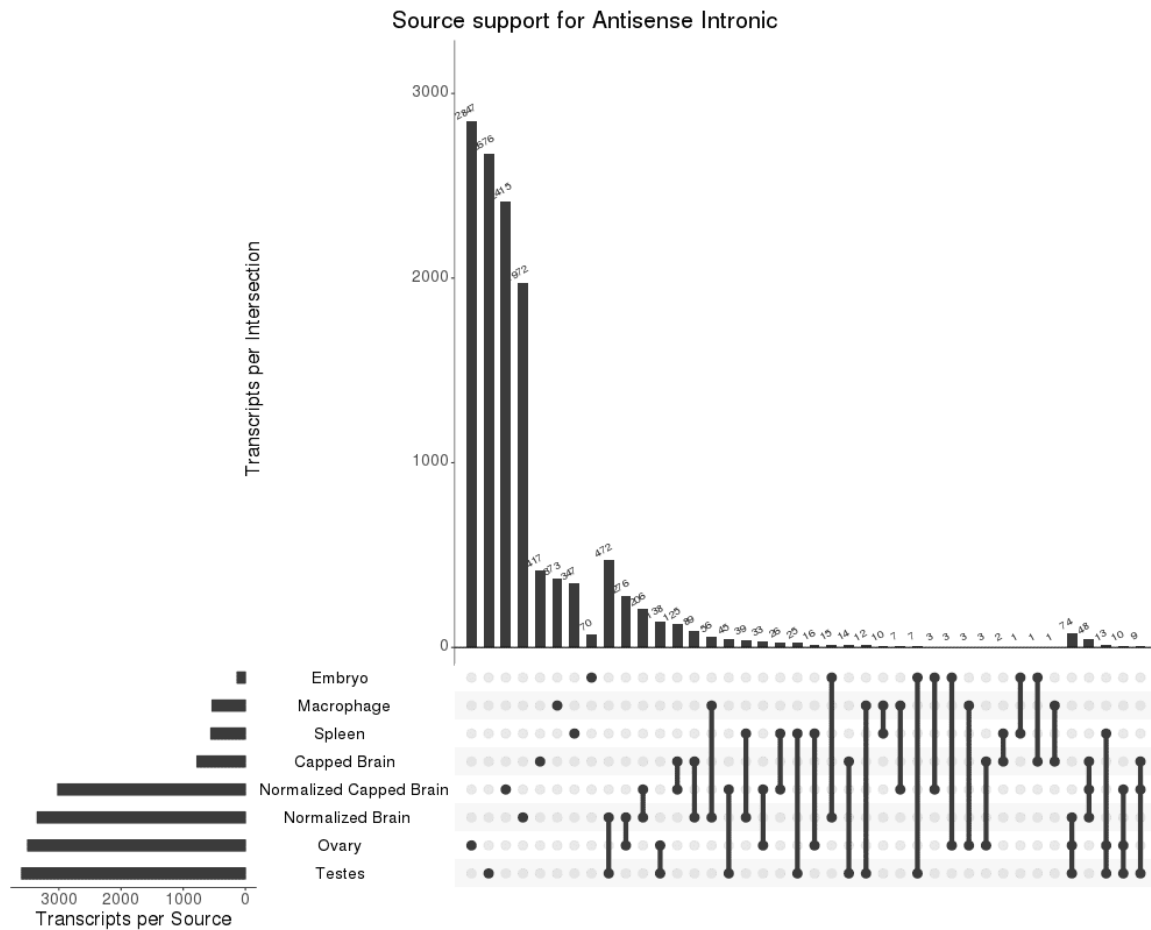


Figure 4.13 : Upset plot showing combination of intersections across the different samples supporting the antisense intronic lncRNA class. The ovary dataset contained the highest number of unique antisense intronic lncRNA but the largest intersection is between testes and the normalized brain sample.

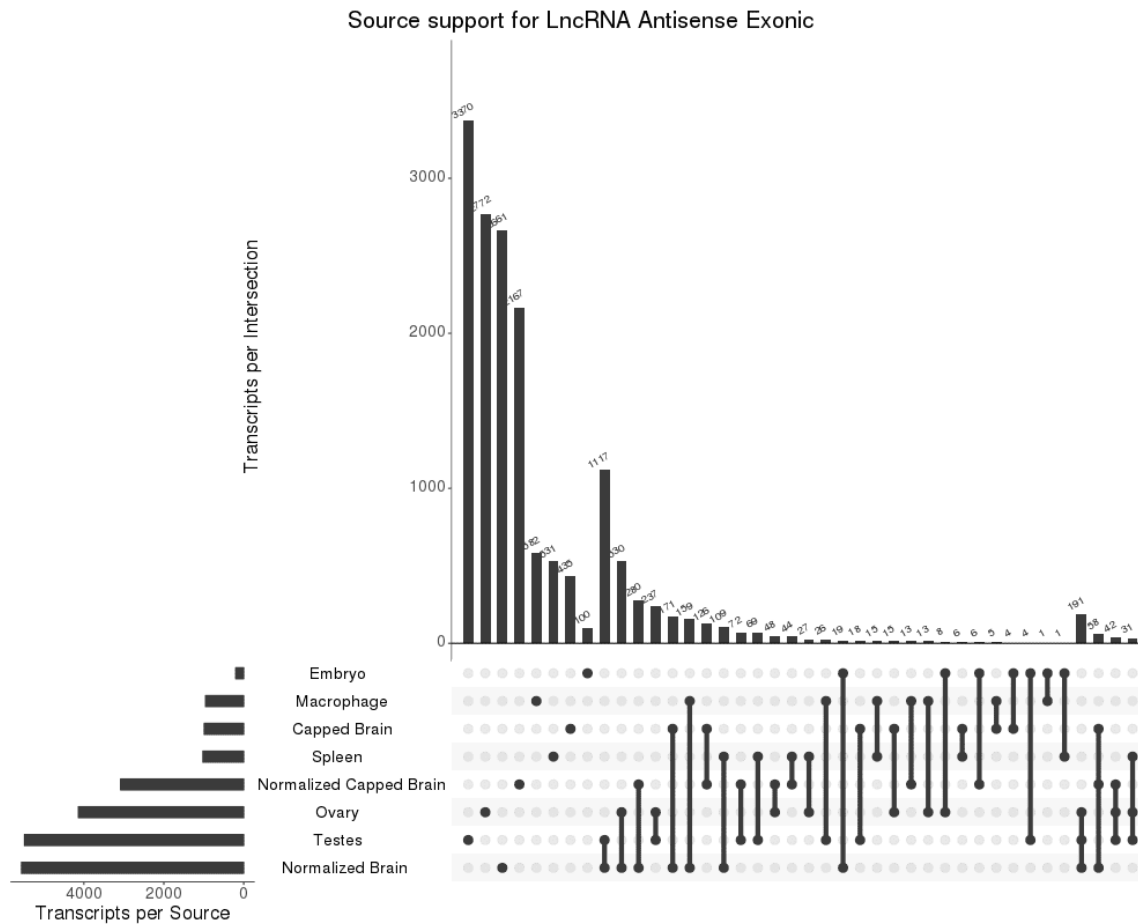


Figure 4.14 : Upset plot showing combination of intersections across the different samples supporting the lncRNA antisense exonic lncRNA class. The testes dataset contained the highest number of unique lncRNA antisense exonic lncRNA and the largest intersection is between testes and the normalized brain sample.

4.3.9 Gene expression patterns for lncRNA

I used short read RNAseq datasets from J-line and broiler chickens to identify expression patterns for the transcript models in the Iso-Seq annotation (Table 4.4). These were comprised of 21 tissue types each, thus totaling 42 different samples.

Since lncRNA have been observed to have generally low expression as compared to protein-coding genes, I looked at the expression of the predicted lncRNA across the short read RNAseq datasets. Due to issues with transcript level quantification when using transcriptome annotations with complex alternative transcripts, I only report the gene level expression. I looked at both the maximum expression level (Figure 4.15) and average expression level (Figure 4.16) in transcripts per million (TPM) for the different classes of lncRNA across the broiler and layer tissue RNAseq datasets. I also looked at comparing the maximum and average levels of expression between coding (Figure 4.17) and non-coding genes (Figure 4.18). Maximum expression level was calculated as the highest expression level for each gene seen in at least one RNAseq dataset.

lncRNA, sense exonic lncRNA, and sense intronic lncRNA had most of their genes showing an average TPM of less than 1. Antisense exonic, antisense intronic, and lncRNA antisense exonic lncRNA showed roughly the same number of genes with average expression levels less than 1 TPM and greater than 10 TPM. In general, it appears that all classes of lncRNA are mostly comprised of genes with a low average expression level.

However, all classes of lncRNA except for sense intronic had their highest number of genes with maximum TPM at above 10 TPM. This would suggest that while lncRNA in general are expressed at low levels, they show tissue specific expression where in the right tissue they exhibit high expression levels.

Sense exonic lncRNA had the highest number of transcripts with a maximum of over 10 TPM in at least one RNAseq dataset. However, the quantification of sense exonic lncRNA is difficult to ascertain due to the exonic overlap with a protein-coding gene.

Table 4.4 Tissue types that were sequenced using short read RNAseq

Sample Number	J Line	Broiler
1	Breast Muscle	Breast Muscle
2	Bursa	Bursa
3	Caecal Tonsil	Caecal Tonsil
4	Cerebellum	Cerebellum
5	Duodenum	Duodenum
6	Gizzard Fat	Gizzard Fat
7	Harderian Gland	Gizzard Muscle
8	Heart Muscle	Heart Muscle
9	Ileum	Ileum
10	Kidney	Ovary
11	Left Optic Lobe	Left Optic Lobe
12	Liver	Liver
13	Lung	Lung
14	Ovary	Ovary
15	Pancreas	Pancreas
16	Proventriculus	Spleen
17	Skin	Thymus
18	Spleen	Thyroid
19	Thymus	Trachea
20	Thyroid	Testes
21	Trachea	Meckel's diverticulum

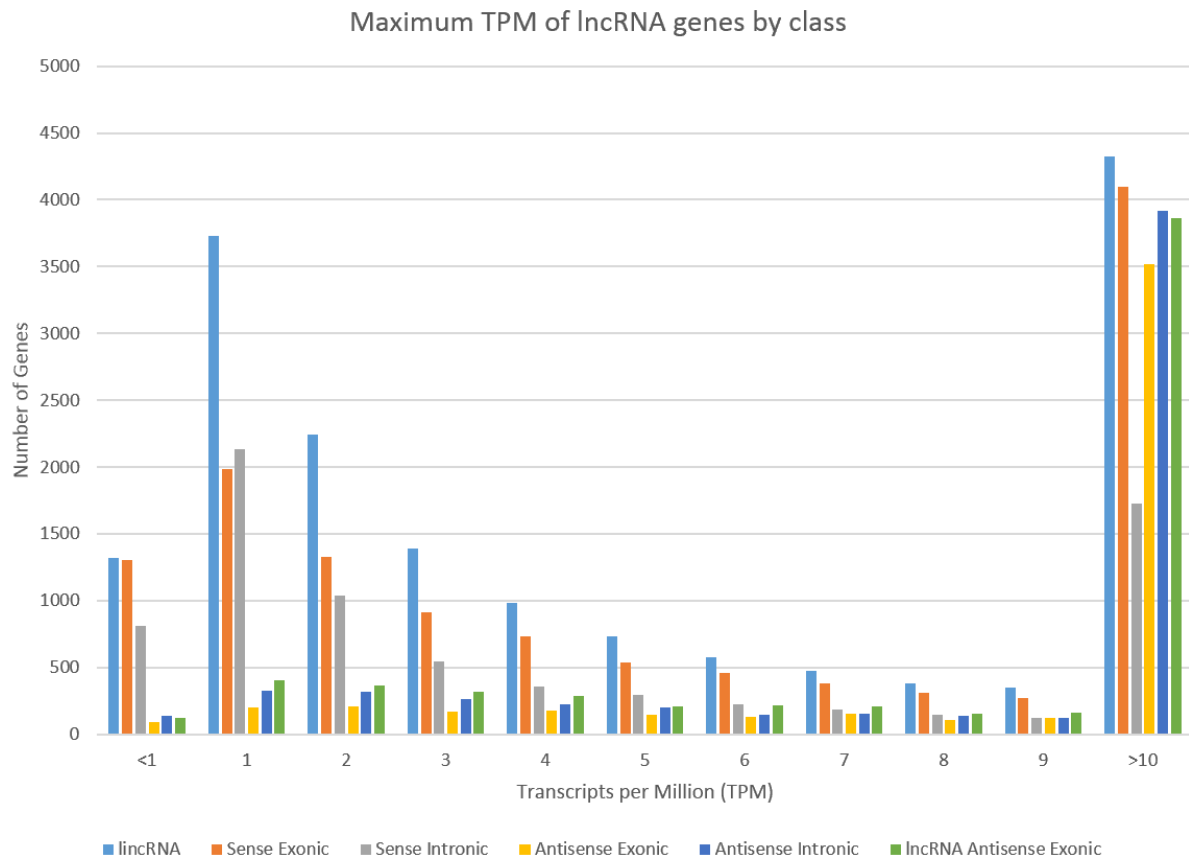


Figure 4.15 : Distribution of maximum TPM per gene based on lncRNA classes. The maximum TPM is chosen from the short read RNAseq sample with the highest TPM for each individual lncRNA gene. All lncRNA classes except sense intronic have their highest peak at 10 or greater TPM.

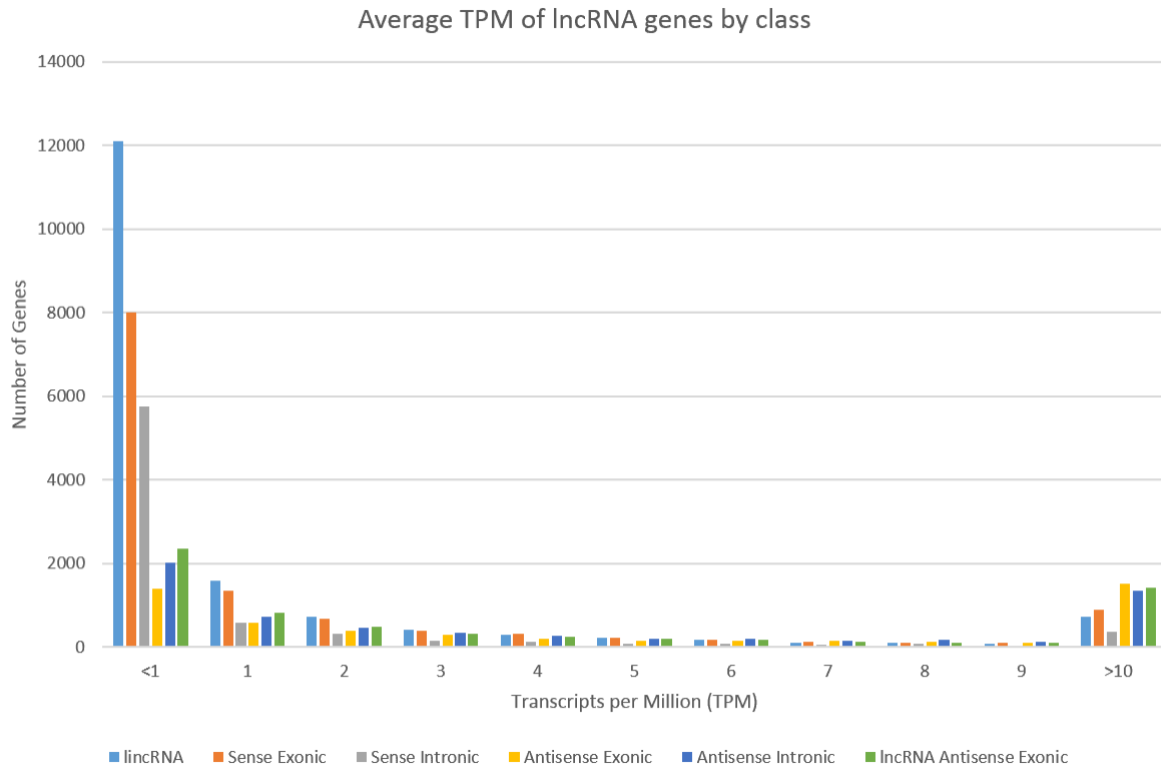


Figure 4.16 : Distribution of average TPM per gene based on lncRNA classes. The average TPM is calculated across all 42 short read RNAseq datasets. All lncRNA classes except for antisense exonic have their highest peak at less than 1 TPM thus supporting the observation from previous studies that lncRNA are in general expressed at lower levels.

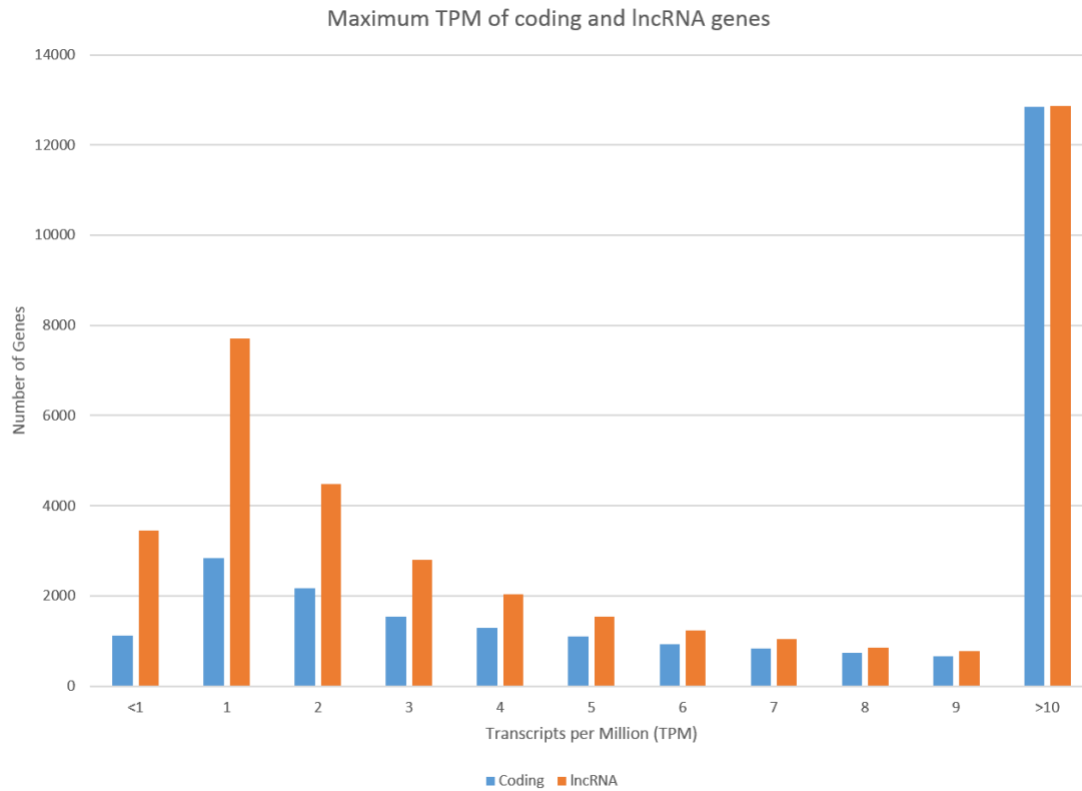


Figure 4.17 : Maximum TPM per gene comparison between coding and non-coding genes. Maximum TPM for both coding and lncRNA genes are chosen from the short read RNAseq datasets with the highest expression for each specific gene. There roughly the same number of coding genes as lncRNA genes with a maximum TPM of 10 or greater.

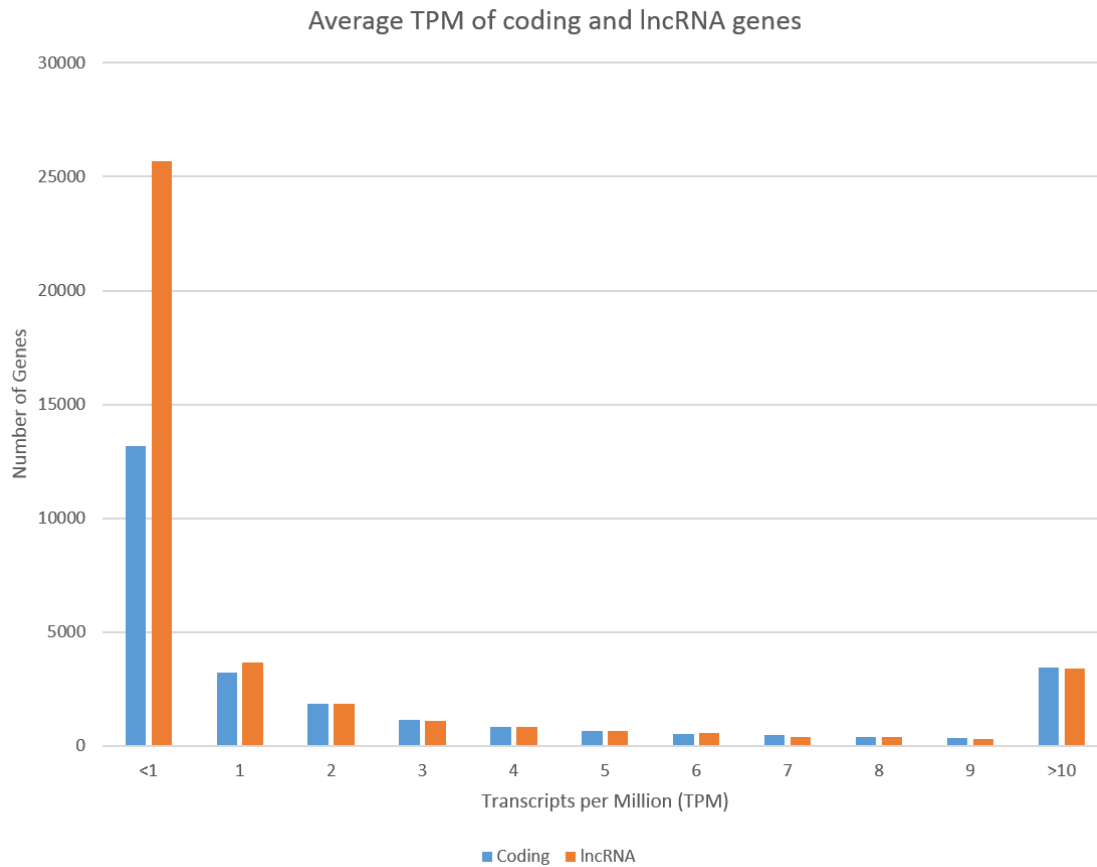


Figure 4.18 : Average TPM per gene comparison between coding and non-coding genes.
Average TPM for both coding and lncRNA genes is calculated across all 42 short read RNAseq datasets. There are almost twice the number of lncRNA genes with lower than 1 TPM as compared to coding genes.

4.4 Conclusion

Since the beginning of my PhD programme there have been many others searching for lncRNA within the chicken transcriptome[142][143][144][145][146]. These entities include the major public annotation organizations: NCBI and Ensembl. I submitted the data generated from this chapter and chapter 2 to the European Nucleotide Archive (ENA). NCBI and Ensembl access the data from ENA to improve their annotations. The Ensembl v88 and higher annotation versions have incorporated both the long read and short read data presented in this thesis. Since those versions of the Ensembl chicken annotation use the data from my thesis, the most recent version that I could use to compare the Iso-Seq annotation to is version 87. NCBI are not as transparent with the data usage for their RefSeq annotation so I am not sure exactly when they began using the data from my thesis, however, the RefSeq annotation versions that use the Gallus Gallus 6 assembly should have incorporated the long read and short read data from this thesis.

The Ensembl v94 chicken annotation that uses the Iso-Seq data from this thesis contains predictions for 5,506 lncRNA genes. The NCBI RefSeq v104 chicken annotation contains predictions for 6,534 non-coding RNA genes with 8,233 lncRNA transcripts from those ncRNA genes. They do not specify the number of lncRNA genes. So while both public annotators have used the Iso-Seq data in their chicken annotations, they are still predicting far fewer lncRNA genes than I have predicted in my analyses. This disparity is likely due to the difference in filtering criteria between the different annotations.

There have also been other groups involved in the identification of lncRNA. One of the most notable recent chicken lncRNA studies was published in 2020 from members of the Functional Annotation of Animal Genomes consortium (FAANG) [147]. In this 2020 study, information was integrated from different public databases as well as from short read RNAseq generated from 25 tissues. From the integration of these different datasets, they predicted 30,084 lncRNA genes including the genes that had already been annotated in the source public annotations.

Using their large short read RNAseq dataset, they predicted 13,009 novel lncRNA genes. Of these novel lncRNA, 7,265 were mono-exonic and 5,744 were multi-exonic. Thus, their results seem to indicate that there are many more mono-exonic lncRNA than were previously represented in the public annotations. While there have been other studies focusing on chicken lncRNA, they were mostly concerned with specific experimental conditions as opposed to general transcriptome annotation [144]-[143].

The discrepancy in the numbers and class breakdown of predicted lncRNA between the public annotations and community studies, highlights the inherent difficulty in lncRNA annotation. While more work needs to be done to verify these predicted lncRNA genes, I believe that a large number of these genes are real. By real, I mean that RNA is produced which matches those transcript models. However, it is much more difficult to answer the question of how many of these predicted lncRNA are functional. There seems to be some controversy over what a gene must do to be considered functional. The more traditional biologist tends to focus on protein coding genes and only acknowledge the well-established non-coding RNA such as ribosomal RNA and transfer RNA[148]. Perhaps the more modern take is that there exist a large set of non-coding RNA with important biological functions[149]. From my perspective, the more important question is how these lncRNA relate to different biological pathways. For example, it may be that a specific lncRNA does not in itself serve any function but the act of transcribing it serves some function. In this case the function could simply be to open up the genomic region for the transcription of neighboring genes. It could be that a subset of these lncRNA are simply by-products of some messy transcription of an upstream protein-coding gene. In this case the lncRNA exists because the RNA polymerase may continue on from the upstream gene to then transcribe the lncRNA gene. The lncRNA portion of the RNA may then be cleaved during processing. However, the fact that it had to be produced to allow for the production of the protein-coding RNA seems to me to be “functional”. Unfortunately, many of these types of cases are much harder to investigate. Therefore, I suspect it will be some time before we start seeing evidence of these phenomena showing up in the literature. However, I believe it is

important to keep an open mind about the complexity of lncRNA so as not to make baseless assumptions that may hinder the way we continue to explore the transcriptome.

The use of expression information across tissue types can provide some early clues as to which lncRNA we should investigate further. As long read sequencing becomes more affordable, it will also be more feasible to look at transcript level expression quantification. This will also mean that we will be able to quantify the expression of tricky classes of lncRNA such as the sense exonic class. The sense exonic lncRNA class represents a truly challenging class to investigate. Because they overlap on the same strand with the exons of a protein-coding gene, it is difficult to use short read information to distinguish sense exonic lncRNA from the overlapping protein coding gene. While some may argue that the vast number of sense exonic lncRNA identified in this study are likely to be artifacts, the way these lncRNA present with respect to their overlapping coding genes does not appear to me to be the result of noisy transcription or the by-products of RNA processing. Since many of these sense exonic lncRNA have read support from the 5' capped datasets it seems likely that they are indeed 5' capped. If this is true, then it would be very strange for the majority of them to be noise since it would not make much sense for the cell to add a 5' cap to transcript fragments that are meant to be destroyed. However, it is possible that the 5' cap selection method is much less stringent than expected and/or that 5' capping may be a biologically noisy event. It is possible that some number of genomic DNA fragments made it through library preparation resulting in the sense exonic models, however, it seems unlikely that their numbers would be this high with a 5' cap selection method. Thus, I believe there is strong evidence in support of these sense exonic lncRNA. With their high numbers in these datasets and their low numbers in public annotation, this class of lncRNA is probably the most under-annotated classes of lncRNA today.

I hope the result from this study can help build the foundation for the next generation of research that will one day uncover the truth behind lncRNA within the chicken transcriptome.

Chapter 5: Discussion

In this thesis I have used relatively new long read RNA sequencing technology to identify thousands of potentially novel long non-coding RNA genes within the chicken transcriptome. I have developed new bioinformatic tools as well as creating new data analysis pipelines to integrate different types of data harnessing the strengths of different types of technologies to shed light on the existence and biological significance of different classes of lncRNA in chicken. I have also experimented with different sample preparation methods with the potential of improving long read RNA sequencing for lncRNA discovery.

While the use of these new technologies has incredible potential to illuminate the world of lncRNA, there is still more progress that can be made to fully harness these new technologies. There have also been other important developments in RNA sequencing that could help to clarify the functionality of newly discovered lncRNA.

In the following sections, I will elaborate on the unrealized improvements and identify areas for future work.

5.1 Improving long read sequencing for lncRNA discovery

Long read RNA sequencing provides much needed information on the full-length sequence of transcripts and thus can improve lncRNA discovery [150]. However, this information is not a given. The sequencing data are the result of multiple biochemical processing steps and biological realities that can lead to an incomplete view of the transcriptome. In order for long read RNA sequencing to see its full potential, better methods for RNA sample processing will need to be developed.

RNA is well known for its instability. The relatively fragile nature of RNA coupled with abundant environmental RNases means that much of the RNA within a sample is fragmented or degraded from the 5' end [151]. Thus even though the sequencing technology has the potential to provide full-length sequences, it cannot do this when the full-length transcript molecules are no longer present in the sample. Indeed, within the data I analyzed, there is abundant evidence for a high percentage of degraded RNA. For example, it can be extremely difficult to differentiate degraded RNA from intact transcripts that have alternative transcript starts or ends. This issue could be ameliorated by new techniques that either remove degraded RNA or allow for the differentiation between complete and incomplete RNA sequences. 5' cap selection is a promising method for this [152], however, current 5' cap selections operate more as a 5' cap bias than true selection. If a new method of 5' cap selection could have a higher rate of distinguishing between 5' capped and non-capped RNA, this would improve the overall data significantly.

However, it is important to recognize that not all RNA will be naturally 5' capped [153][154][155]. In light of this, other methods will need to be developed to help identify the full-length mature transcripts for these types of RNA.

Just as not all RNA are 5' capped, not all RNA have poly-A tails [156]-[157]. Most of the long-read RNA sequencing pipelines rely on oligo-dT primers in order to select transcripts with a poly-A tail for reverse transcription. Thus, any RNA without a poly-A tail are essentially removed from the library before sequencing. While this can be avoided by ligating adapters to the 3' ends of RNA before reverse transcription [158][159], this will also increase the amount of 3' truncated RNA representation in the cDNA library. Therefore, this problem presents a tricky compromise on each side. Given that many researchers suspect that many lncRNA are not polyadenylated [160][161], we still do not have a robust method for identifying these types of lncRNA.

Another issue that affects all RNA sequencing is the presence of incompletely processed RNA [162][163]. These are RNA molecules that have been captured during their transcription and/or processing phases. In these cases the RNA are not mature and the sequences would only indicate an ephemeral form that in itself has no biological function. Sequencing reads from these RNA often look like retained intron events since not all the introns have been spliced out yet. The issue this poses for transcriptome annotation is that true retained introns do exist which makes differentiating real retained introns from incompletely processed RNA currently highly challenging [164], [165][166].

As previously covered in depth in the first results chapter, the highly differential abundance of RNA within a sample often means that low abundance RNA never get sequenced. These represent a major blind spot for lncRNA discovery since many lncRNA are often present in relatively low amounts within any given sample.

Another biochemical processing issue with current long read RNA sequencing is that both PacBio and ONT RNA sequencing require the use of reverse transcription. While ONT does allow for direct RNA sequencing, reverse transcription is still used to create a complementary strand which prevents structural formations and stabilizes the kinetics of the RNA molecule while traveling through the pore [167]. Reverse transcription has several issues and limitations [168][169]. One of the major limitations for reverse transcription is the length of sequence that can be reliably reverse transcribed [170]. At present, the maximum reliable reverse transcription length is roughly 10-15 kb. Thus, even though long RNA sequencing technologies are capable of producing reads far longer than this, we do not yet have a way of creating the templates for sequencing.

It also now well known that RNA nucleotides can undergo modifications such as methylation and the conversion of adenosine to inosine [171][172]. These types of modifications still require further development of long read RNA sequencing to be able to robustly identify at a high throughput level [173][174].

Given the evidence that lncRNA are often highly tissue specific in their expression, it stands to reason that there is a significant subset that is also highly cell specific. However, the current cost and throughput of long read RNA sequencing makes it impractical for single cell sequencing methods. In the case of ONT sequencing, the error rates are still too high to reliably identify the cell barcodes.

5.2 Improving analysis of long read RNA sequencing

While there is still much work to do on the sample processing side of long read RNA sequencing, there is perhaps even more that needs to be done to develop the analysis of long read RNA sequencing data [175]. The development of more sample processing techniques will require complementary analysis methods to be developed.

Long read RNA sequencing provides additional types of information that previous forms of RNA detection were incapable of producing. This includes the potential to identify the transcription start site, end site, and full exon chaining of a single transcript. However, these pieces of information are not guaranteed and there are many reasons why a long read can obscure the truth behind any of these features. As such, there have been some development of methods for distinguishing signal from noise. Some of these methods depend on orthogonal information [168][176][87][177], some methods use strategies originally developed for short read analysis [90], and some use information from purely long read datasets [129].

Each of these philosophies comes with specific advantages and limitations. The use of orthogonal information such as comparing long read datasets to reference annotations can be incredibly powerful for validating the more well established genes, however this does not lend itself to discovery as these methods are essentially limiting their scope to what is known

instead of truly exploring the unknown. Thus these methods seem to diminish some of the potential of long read data.

Given the amount of time that short read RNA sequencing seemed to dominate transcriptomics and the number of bioinformatic tools that were developed to harness the power of short read sequencing [178], it comes as no surprise that many bioinformaticians approached long read RNA sequencing with a very short read perspective. However, many of the principles of short read simply do not translate when working with long read data. For example, some major bioinformatic tools for processing long read RNA sequencing data have algorithms that attempt to merge reads to create longer transcripts [90]. However, this action is at odds with what the long reads represent and actually results in more erroneous transcript models.

Many long read tools focus primarily on read coverage for discerning between signal and noise transcript models. While it is generally accepted that higher read coverage equates to more evidence, there is an underlying challenge of identifying which reads should be considered to be supporting a specific transcript. So while read coverage is important, calculating read coverage is non-trivial with long read data. The issue here is that there is almost always some slight differences between any 2 reads within long read data. Usually the differences are on the 5' and 3' ends but sometimes there are slight differences in the splice junctions. There is a very real challenge here for selecting the appropriate thresholds for inclusion or exclusion of reads from a grouping. It is not very clear how this should be done and in most cases it really depends on what the data will be used for downstream. However, many bioinformaticians seem to just choose some arbitrary threshold and move on with their analyses. I believe that the best way of handling these situations is to move away from the linear representations of transcripts and to develop transcript models that have these feature variations built in. Imagining how these RNA are being produced in the cell, it seems likely to me that transcription is more complex and more messy than simple start and stop sites. We have certainly seen evidence of this from large studies on transcription start sites [179]. Thus, it does not make sense to me as to why many researchers still subscribe to a linear representation of transcripts that mask these real

complexities. However, the rise of graph based representation in genome representation will likely carry over to transcriptome representation and once the major public annotators adopt graph based genomes they may choose to jump to graph based transcriptome annotations in the same leap.

In order to understand how to improve long read RNA sequencing analyses in more depth, we must have some workable definition of what transcriptional noise is. Actually I prefer to call this sequencing noise as the former term can be a cause for miscommunication and confusion due to how such noise is generated. I define transcriptional/sequencing noise as reads or read characteristics that result in the production of a transcript model which is significantly different from any real and meaningful transcript within the sample of origin. However, there is another definition of transcriptional noise which is defined as spurious transcription of the genome resulting in superfluous RNA production. Essentially this other definition is concerned with the concept of messy transcription where by not all RNA produced in a cell is functional or relevant. I believe that our current knowledge of the transcriptome is too limited for us to really understand what the relevance of an RNA that so far has no annotated function actually is. Non-coding RNA are a prime example of this uncertainty. If an RNA is not translated into protein, what function does it serve? While we know the function of a small set of ncRNA, there could be many more unimagined roles for ncRNA that we have yet to uncover.

Given that it is so difficult to define what exactly constitutes transcriptional noise, I side more with the idea that we should not discount RNA when we have clear evidence that it is transcribed with some consistency. However, if we use the definition of “transcribed with some consistency” as the threshold, then we must explore how consistency could be created by artificial circumstances.

A matter that has come to light or at least has had more light shed upon it with the advent of high throughput long read sequencing, is the effect that sample processing has on the resulting sequencing data. I believe that biases or changes to the original RNA composition during

sample processing can be classed under sources of transcriptional noise or rather sequencing noise. This is to say that sample processing can result in sequencing reads that do not accurately represent the original relevant RNA library. The issue with noise coming from sample processing is that it can often be repeatable even across samples. Thus using read count or cross sample comparison to filter out these artifacts can sometimes be ineffective and lead to a false sense of confidence in the final curated transcriptome annotation.

Given the relative novelty of long read RNA sequencing, I believe that more effort should be spent on the development of long read centric methods. By this I mean methods where long read data is assessed without orthogonal information and without integrating with short read datasets. There have been a few developments in this area which are promising [180][181]. By focusing on this more simplistic pursuit we can establish more robust methods of handling this type of data whereas efforts to integrate other types of data seem to be missing some points.

Many of these issues seem to stem from the relatively limited exposure that many researchers have had with long read RNAseq data. However, the popularity of long read RNAseq seems to be growing rapidly. With this growth, there also seems to be more people interested in understanding the nuances of long read data analysis. This can be evidenced by the increasing number of papers citing the TAMA software [182]-[183].

One challenging aspect of predicting true splice junction locations is the complexity surrounding small shifts in either splice site starts or ends. For example, the NAGNAG phenomenon is well documented in protein coding genes which basically represent a tandem repeat of the intron acceptor site which allows for the use of two different splice junction end sites that are adjacent [184]. There is some evidence that NAGNAG occurs in lncRNA as well [185]. This particular challenge is one that TAMA does not handle particularly well due to the reliance on using wobble assessment and groupings. While this could be overcome by using specific parameters with TAMA Collapse, I am not satisfied with the current method of dealing with these scenarios. In essence, to be able to identify these small shifts in splice junctions, the user

of TAMA would need to use a no wobble setting which would mean a likely over inflation of the prediction of small shifts in splice junctions that were caused by sequence or mapping errors. With the type of data presented in this thesis the error rates were too high for this approach to work well. However, as long read sequencing becomes more accurate, new methods could be used to take advantage of the improved accuracy for identifying these short splice junction shifts with greater accuracy. It is my hope to continue the development of TAMA to create a new version which will have a more sophisticated way of handling these scenarios.

While the field of long read RNAseq is still somewhat nascent, the bioinformatic ecosystem is growing with new tools being released by academic groups every year [168], [177], [180][186]-[187]. It is also good to see that the long-read ecosystem has a good diversity of different tools to fill the specific niches of data analysis requirements. Hopefully this trend continues and is embraced by the community since a one size fits all mentality can be problematic when it comes to data analysis and a wider selection of commonly used tools should encourage researchers to explore the best tool for their experiments as opposed to just choosing the most popular ones.

5.3 Future work for chicken lncRNA annotation

Long non-coding RNA annotation in chickens requires three important stages: generation of high quality sequencing data, availability of appropriate analysis software, and functional validation. I focused on the first two parts within this thesis and attempted to identify the best methods for optimizing those aspects of lncRNA discovery in chicken.

There have been other efforts to identify and functionally annotate chicken lncRNA. Much of this work has been carried out by members of animal genomics consortia like the Vertebrate Genome Project (VGP) and FAANG as well academic groups that are not associated with larger consortia. For instance, members of FAANG integrated a massive short read RNAseq dataset (including 25 tissues) with the major genomic, transcriptomic, and protein annotation resources

to identify large numbers of novel lncRNA within chicken [147]. Another subset of FAANG performed a more focused search for lncRNA in chicken liver and adipose tissues [69]. There was also a similar study focusing on chicken liver through different developmental stages [188]. A group of FAANG members in the United States of America took an even more general approach and looked for lncRNA in chicken, cattle, and pig [189]. FAANG members even developed their own software (FEELnc) for identifying lncRNA from predicted transcript models [190]. There have been a number of considerable studies on chicken lncRNA outside of FAANG as well [70][72], [191]-[192]. Among other studies, researchers have looked at the role of epigenetic regulation of lncRNA expression within the avian immune system [193], the roles of lncRNA in feed efficiency [191], lncRNA functionality in Marek's disease tumorigenesis [194], roles of lncRNA in the differentiation of abdominal preadipocytes [145], lncRNA involved in the domestication of chicken [70], and roles of lncRNA in chicken liver development [188].

All of these works have provided some clues as to the underlying complexity and breadth of lncRNA that are likely to exist within the chicken transcriptome. However, they all relied on short read RNA sequencing and did not perform *in vitro* or *in vivo* tests for lncRNA functionality.

The results from my PhD study demonstrate the first use of Iso-Seq long read RNA sequencing for the identification of lncRNA in chicken. The use of this novel technology came with both drawbacks and advantages. One of the drawbacks include having to learn about the properties of this new sequencing technology without a significant foundation of knowledge from previous publications to rely on. Learning how best to use Iso-Seq required communication with PacBio's research staff as well as others in the Iso-Seq community who were also just starting to learn. The software resources at the start were limited and most of them were hacks of other tools to serve a non-intended purpose of processing Iso-Seq data [195][196][197]. The PacBio software for running the CCS and Cluster software were designed to work on their own machines that meant that they did not port smoothly over to the university servers. The Roslin Institute and University of Edinburgh IT departments had to setup special computing environments to allow for the installation of the first generation PacBio software. It was only after years of requests

from the PacBio user community that they began making their software easier to install on other computing environments.

When I finally had the chance to look at the transcript model results from the Iso-Seq data, I saw something that looked strange and new. The transcript models contained large expansions of alternative transcripts as well as long mono-exonic models. There were alternative splicing events that did not match anything that was in the public chicken annotations and many of them looked like they must be the result of some noise. After witnessing those results, I had two options: resort to conventional short read RNA sequencing ideas or develop methods specific to the characteristics of long read RNA sequencing. I believed that many of the assumptions used in short read methodologies did not apply to long read data and that many characteristics of long read data could be used to improve the data analysis. The only software for creating transcript models from Iso-Seq reads at the time was the software from PacBio. This set of tools was called Tofu and developed by Elizabeth Tseng at PacBio [198]. However, Tofu was not designed to handle Iso-Seq performed on 5' cap selected or normalized cDNA libraries. Tofu also had limitations on what could be done with the analysis. In response to this, I developed the Transcriptome Annotation by Modular Algorithms (TAMA) suite of tools to meet the needs for using Iso-Seq with different cDNA library preparation methods for the identification of lowly expressed genes in organisms with low complexity public genome annotations.

As part of the TAMA suite of tools, I also developed methods of lncRNA prediction that were specific to long read data. I developed these tools so that they could take into consideration issues with possible 5' incompleteness of transcript models and errors in splice junctions that could cause frame shifts in open reading frames. I also developed a lncRNA classification tool that identified positional lncRNA classes based on their positional relationship to other lncRNA as opposed to in relation to the nearest protein coding gene.

By finding solutions to the challenges of working with the relatively new Iso-Seq technology, I was able to find some major advantages of Iso-Seq for the prediction of lncRNA. By reducing the complication of transcript assembly, the Iso-Seq reads allow for more assured exon chaining and also for identifying mono-exonic transcripts. Situations that would be difficult if not impossible to resolve with short read data, such as transcripts with overlapping exons and different exon chains, are typically much clearer in a long read dataset. Since there is still much that is unknown with respect to lncRNA representation on the genome, these advantages of long read RNA sequencing unlock a new level of discovery.

Despite the large step forward that long read RNA sequencing has produced for lncRNA discovery, there is still the matter of verifying the functionality. While *in silico* analysis has provided some indications of functionality, uncovering the biological roles of each lncRNA will require further work. This work should comprise biochemical approaches to elucidating the exact nature of each potential novel lncRNA gene.

There are two types of functional investigation modes that can be applied for lncRNA: *in vitro* and *in vivo* experimentation. *In vitro* experiments are typically higher throughput as compared to *in vivo* experiments but they do not provide the same level of information.

There are newer *in vitro* methodologies have been created to test thousands of genomic modifications in a high throughput manner. These high throughput *in vitro* methods now often use CRISPR gene editing [199][200]. However, they mostly serve to provide a further narrowing down of the specific lncRNA that are involved in certain biological mechanisms. To really explore the mechanisms behind lncRNA function, lower throughput *in vivo* studies will be required. I propose that a ranking system needs to be developed to help rank lncRNA as best candidates for further investigation with *in vivo* methods. The lncRNA rank should take into account several characteristics of the potential functional lncRNA. Characteristics such as proximity to well annotated protein coding genes, neighboring protein coding genes involved in traits which are important in biological applications, ease of discerning between the lncRNA and neighboring genes, and expression patterns that correlate to known pathways. Generally, these

characteristics can be broken down into likelihood of being functional, likelihood of being valuable, and ease of testing functionality. This idea is similar to the way many researchers handle differential gene expression information from RNA sequencing. With so many pieces in the puzzle, we need a way of building the foundations first.

With the rise of more advanced gene editing technologies such as CRISPR-Cas9 [201], we now have the ability to probe gene functionality with greater precision, ease, and throughput. These new technologies coupled with the ease of access to chickens at different developmental stages [202] provides a fantastic opportunity to design faster and better experiments for testing the functional roles of novel lncRNA genes. We can now effectively turn on and off genes within the chicken genome in a robust way that will allow us to see their effects through embryonic development and into adulthood. These methods have already been employed for lncRNA functionality investigation in other species [203]-[204]. A few of these studies used zebrafish due to the ease of applying CRISPR to this species to investigate development. The chicken egg presents a similar experimentally tractable system and with the development of more advanced gene editing technologies in birds [205][206][207], there is a prime opportunity to combine these technologies to directly investigate the functional roles of newly predicted lncRNA.

Thus, an exciting next step would be to choose a set of the top ranking predicted lncRNA genes and perform gene editing experiments to either knockout their expression or enhance their expression. Some ideal characteristics would be lncRNA that are highly expressed in embryos. Knocking out these genes would allow for a direct test to see if they indeed serve some function in the embryonic developmental stages. Targeting lncRNA that are potentially crucial for embryonic development also allows for an overall less complicated experiment because there would be less overall time between gene editing and observable phenotypic differences.

With enough funding and time, it would be ideal to run longer experiments that would allow for the exploration of lncRNA that are potentially involved in traits such as growth rate, feed efficiency, and health.

5.4 General conclusions

With the advent of these new sequencing technologies, the world of long noncoding RNA has suddenly become far more approachable. While there is still much unknown about this mysterious class of RNA, I believe we will soon make massive strides in understanding different sub-classes of lncRNA that serve pivotal biological roles. Given that we still have not found the mechanisms and actors that shape species diversity, I suspect that at least some part of that answer lies within lncRNA. I hope that soon even more exciting technologies will be developed to help us find these answers.

Chapter 6: Bibliography

1. Godfray HCJ. Challenges for taxonomy. *Nature*. 2002;417(6884):17–9.
2. Raven PH, Berlin B, Breedlove DE. The origins of taxonomy. *Science* (80-). 1971;174(4015):1210–3.
3. Godfray HCJ. Linnaeus in the information age. *Nature*. 2007;446(7133):259–60.
4. Darwin C. *On the Origin of Species*. Cambridge University Press; 1859.
5. Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc Natl Acad Sci U S A*. 2015;112(41):12764–9.
6. Davis RH. The age of model organisms. *Nat Rev Genet*. 2004;5(1):69–76.
7. Müller B, Grossniklaus U. Model organisms - A historical perspective. *J Proteomics* [Internet]. 2010;73(11):2054–63. Available from: <http://dx.doi.org/10.1016/j.jprot.2010.08.002>
8. Hedges SB. The origin and evolution of model organisms. *Nat Rev Genet*. 2002;3(11):838–49.
9. Phifer-Rixey M, Nachman MW. Insights into mammalian biology from the wild house mouse *Mus musculus*. *Elife*. 2015;2015(4):1–13.
10. Roberts DB. *Drosophila melanogaster*: The model organism. *Entomol Exp Appl*. 2006;121(2):93–103.
11. Goldstein B, King N. The Future of Cell Biology: Emerging Model Organisms. *Trends Cell Biol* [Internet]. 2016;26(11):818–24. Available from: <http://dx.doi.org/10.1016/j.tcb.2016.08.005>
12. Leung MCK, Williams PL, Benedetto A, Au C, Helmcke KJ, Aschner M, et al. *Caenorhabditis elegans*: An emerging model in biomedical and environmental toxicology. *Toxicol Sci*. 2008;106(1):5–28.
13. Kaletta T, Hengartner MO. Finding function in novel targets: *C. elegans* as a model organism. *Nat Rev Drug Discov*. 2006;5(5):387–99.
14. Willemsen R, Padje SVT, Van Swieten JC, Oostra BA. Zebrafish (*Danio rerio*) as a model

- organism for dementia. *Neuromethods*. 2011;48:255–69.
15. Fumihito A, Miyake T, Sumi SI, Takada M, Ohno S, Kondo N. One subspecies of the red junglefowl (*Gallus gallus gallus*) suffices as the matriarchic ancestor of all domestic breeds. *Proc Natl Acad Sci U S A*. 1994;91(26):12505–9.
 16. Burt DW. Emergence of the chicken as a model organism: Implications for agriculture and biology. *Poult Sci [Internet]*. 2007;86(7):1460–71. Available from: <http://dx.doi.org/10.1093/ps/86.7.1460>
 17. Mullikin JC. The evolution of comparative genomics. *Mol Genet Genomic Med*. 2014;2(5):363–8.
 18. Brown WRA, Hubbard SJ, Tickle C, Wilson SA. The chicken as a model for large-scale analysis of vertebrate gene function. *Nat Rev Genet*. 2003;4(2):87–98.
 19. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*. 2004;432(7018):695–716.
 20. Cogburn LA, Porter TE, Duclos MJ, Simon J, Burgess SC, Zhu JJ, et al. Functional genomics of the chicken - A model organism. *Poult Sci*. 2007;86(10):2059–94.
 21. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. Vol. 346, *Science*. 2014. 1320–1331 p.
 22. Jarvis ED, Ye C, Liang S, Yan Z, Zepeda ML, Campos PF, et al. A Phylogeny of Modern Birds. *Science (80-)*. 2014;346(6215):1126–38.
 23. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science (80-)*. 2014;346(6215):1311–21.
 24. Hubbard SJ, Grafham D V., Beattie KJ, Overton IM, McLaren SR, Croning MDR, et al. Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags. *Genome Res*. 2005;15(1):174–83.
 25. Carre W, Wang X, Porter TE, Nys Y, Tang J, Bernberg E, et al. Chicken genomics resource: Sequencing and annotation of 35,407 ESTs from single and multiple tissue cDNA libraries and CAP3 assembly of a chicken gene index. *Physiol Genomics*. 2006;25(3):514–24.
 26. Nagaraj SH, Gasser RB, Ranganathan S. A hitchhiker’s guide to expressed sequence tag

- (EST) analysis. *Brief Bioinform.* 2007;8(1):6–21.
27. Parkinson J. BM. Expressed Sequence Tags: An Overview. In: *Methods in Molecular Biology*. 2009. p. 10–27.
 28. 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* (80-). 1991;252(5013):1651–6.
 29. Stapleton M, Carlson J, Brokstein P, Yu C, Champe M, George R, et al. A *Drosophila* full-length cDNA resource. *Genome Biol.* 2002;3(12):1–8.
 30. Rnas N, Riken T, Gene M, Project E, When U, For U, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature.* 2002;420(5):563–73.
 31. Smith EJ, Shi L, Drummond P, Rodriguez L, Hamilton R, Ramlal S, et al. Expressed sequence tags for the chicken genome from a normalized 10-day-old White Leghorn whole embryo cDNA library: 1. DNA sequence characterization and linkage analysis. *J Hered.* 2001;92(1):1–8.
 32. Boardman PE, Sanz-Ezquerro J, Overton IM, Burt DW, Bosch E, Fong WT, et al. A comprehensive collection of chicken cDNAs. *Curr Biol.* 2002;12(22):1965–9.
 33. Reyes A, Huber W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* 2018;46(2):582–92.
 34. Bush SJ, Chen L, Tovar-Corona JM, Urrutia AO. Alternative splicing and the evolution of phenotypic novelty. *Philos Trans R Soc B Biol Sci.* 2017;372(1713):1–7.
 35. Chen L, Tovar-Corona JM, Urrutia AO. Alternative Splicing: A Potential Source of Functional Innovation in the Eukaryotic Genome. *Int J Evol Biol.* 2012;2012:1–10.
 36. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbelt JO, et al. What is a gene, post-ENCODE? History and updated definition. *Genome Res.* 2007;17(6):669–81.
 37. Ying SY, Lin SL. Current perspectives in intronic micro RNAs (miRNAs). *J Biomed Sci.* 2006;13(1):5–15.
 38. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet*

- [Internet]. 2019; Available from: <http://www.nature.com/articles/s41576-019-0150-2>
39. Hitzemann R, Bottomly D, Darakjian P, Walter N, Iancu O, Searles R, et al. Genes, behavior and next-generation RNA sequencing. *Genes, Brain Behav.* 2013;12(1):1–12.
 40. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 2006;34(WEB. SERV. ISS.):435–9.
 41. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res.* 2012;40(D1):130–5.
 42. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, et al. The Ensembl automatic gene annotation system. *Genome Res* [Internet]. 2004 May [cited 2014 Sep 29];14(5):942–50. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=479124&tool=pmcentrez&rendertype=abstract>
 43. Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC Genomics.* 2020;21(1):1–20.
 44. Tamura K. Origins and early evolution of the tRNA molecule. *Life.* 2015;5(4):1687–99.
 45. R. Brimacombe and W. Stiege. Structure and function of ribosomal RNA. *Biochem J.* 1985;229:1–17.
 46. Zeng Y. Principles of micro-RNA production and maturation. *Oncogene.* 2006;25(46):6156–62.
 47. Dana H, Chalbatani GM, Mahmoodzadeh H, Karimloo R, Rezaiean O, Moradzadeh A, et al. Molecular Mechanisms and Biological Functions of siRNA. *Int J Biomed Sci* [Internet]. 2017;13(2):48–57. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28824341>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5542916>
 48. Hernandez N. Small Nuclear RNA Genes: A Model System to Study Fundamental Mechanisms of Transcription. *J Biol Chem* [Internet]. 2001;276(29):26733–6. Available from: <http://dx.doi.org/10.1074/jbc.R100032200>

49. Statello L, Guo CJ, Chen LL, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol* [Internet]. 2021;22(2):96–118. Available from: <http://dx.doi.org/10.1038/s41580-020-00315-9>
50. Chang YF, Imam JS, Wilkinson MF. The Nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem*. 2007;76:51–74.
51. Brogna S, Wen J. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat Struct Mol Biol*. 2009;16(2):107–13.
52. Chan WK, Huang L, Gudikote JP, Chang YF, Imam JS, MacLean JA, et al. An alternative branch of the nonsense-mediated decay pathway. *EMBO J*. 2007;26(7):1820–30.
53. Hentze MW, Kulozik AE. A Perfect Message: RNA Surveillance and Nonsense-Mediated Decay. *Cell*. 1999;96(3):307–10.
54. Rao MRS. *Advances in Experimental Medicine and Biology: Long Non Coding RNA Biology* [Internet]. 2017. 323 p. Available from: <http://www.springer.com/series/5584>
55. Noller HF. Structure of ribosomal RNA. *Annu Rev Biochem*. 1984;53:119–62.
56. Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, et al. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*. 1991;349(6304):38–44.
57. Ballabio A, Willard HF. Mammalian X-chromosome inactivation and the XIST gene. *Curr Opin Genet Dev*. 1992;2(3):439–47.
58. Brown CJ, Hendrich BD, Rupert JL, Lafrenière RG, Xing Y, Lawrence J, et al. The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*. 1992;71(3):527–42.
59. Marahrens Y, Loring J, Jaenisch R. Role of the Xist gene in X chromosome choosing. *Cell*. 1998;92(5):657–64.
60. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, et al. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell*. 2007;129(7):1311–23.
61. Mozdarani H, Ezzatizadeh V, Rahbar Parvaneh R. The emerging role of the long non-coding RNA HOTAIR in breast cancer development and treatment. *J Transl Med*

- [Internet]. 2020;18(1):1–15. Available from: <https://doi.org/10.1186/s12967-020-02320-0>
62. Rajagopal T, Talluri S, Akshaya RL, Dunna NR. HOTAIR LncRNA: A novel oncogenic propellant in human cancer. *Clin Chim Acta* [Internet]. 2020;503(October 2019):1–18. Available from: <https://doi.org/10.1016/j.cca.2019.12.028>
 63. Bhan A, Mandal SS. LncRNA HOTAIR: A master regulator of chromatin dynamics and cancer. *Biochim Biophys Acta - Rev Cancer* [Internet]. 2015;1856(1):151–64. Available from: <http://dx.doi.org/10.1016/j.bbcan.2015.07.001>
 64. Johnsson P, Lipovich L, Grandér D, Morris K V. Evolutionary conservation of long non-coding RNAs; Sequence, structure, function. *Biochim Biophys Acta - Gen Subj* [Internet]. 2014;1840(3):1063–71. Available from: <http://dx.doi.org/10.1016/j.bbagen.2013.10.035>
 65. Deng P, Liu S, Nie X, Weining S, Wu L. Conservation analysis of long non-coding RNAs in plants. *Sci China Life Sci*. 2018;61(2):190–8.
 66. Wu Z, Liu X, Liu L, Deng H, Zhang J, Xu Q, et al. Regulation of lncRNA expression. *Cell Mol Biol Lett*. 2014;19(4):561–75.
 67. Lee C, Kikyo N. Strategies to identify long noncoding RNAs involved in gene regulation. *Cell Biosci*. 2012;2(37):1–6.
 68. Mattick JS, Rinn JL. Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol*. 2015;22(1):5–7.
 69. Muret K, Klopp C, Wucher V, Esquerré D, Legeai F, Lecerf F, et al. Long noncoding RNA repertoire in chicken liver and adipose tissue. *Genet Sel Evol*. 2017;49(1):1–17.
 70. Wang Y-M, Xu H-B, Wang M-S, Otecko NO, Ye L-Q, Wu D-D, et al. Annotating long intergenic non-coding RNAs under artificial selection during chicken domestication. *BMC Evol Biol* [Internet]. 2017;17(1):192. Available from: <http://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-017-1036-6>
 71. Karimi P, Bakhtiarzadeh MR, Salehi A, Izadnia HR. Transcriptome analysis reveals the potential roles of long non-coding RNAs in feed efficiency of chicken. *Sci Rep* [Internet]. 2022;12(1):1–17. Available from: <https://doi.org/10.1038/s41598-022-06528-6>
 72. Gardner PP, Fasold M, Burge SW, Ninova M, Hertel J, Kehr S, et al. Conservation and

- losses of non-coding RNAs in avian genomes. *PLoS One*. 2015;10(3):1–19.
73. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics* [Internet]. 2016;107(1):1–8. Available from: <http://dx.doi.org/10.1016/j.ygeno.2015.11.003>
74. Robert W. Holley, Jean Apgar, George A. Everett, James T. Madison, Mark Marquisee, Susan H. Merrill, John Robert Penswick AZ. Structure of a Ribonucleic Acid. *Science* (80-). 1965;147:1462–1465.
75. Sanger F, Brownlee GG, Barrell BG. A two-dimensional fractionation procedure for radioactive nucleotides. *J Mol Biol* [Internet]. 1965;13(2):373–98. Available from: [http://dx.doi.org/10.1016/S0022-2836\(65\)80104-8](http://dx.doi.org/10.1016/S0022-2836(65)80104-8)
76. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74(12):5463–7.
77. Osborne MA, Furey WS, Klenerman D, Balasubramanian S. Single-molecule analysis of DNA immobilized on microspheres. *Anal Chem*. 2000;72(15):3678–81.
78. Osborne MA, Balasubramanian S, Furey WS, Klenerman D. Optically biased diffusion of single molecules studied by confocal fluorescence microscopy. *J Phys Chem B*. 1998;102(17):3160–7.
79. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. *Science* (80-). 2009;323(5910):133–8.
80. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol*. 2009;4(4):265–70.
81. 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>
82. 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>
83. Tseng E, Nowakowski TJ, Penn O, Underwood JG, Munson KM, Eichler EE, et al.

- Transcriptional fates of human-specific segmental duplications in brain. *Genome Res.* 2018;28(10):1566–76.
84. Veiga DFT, Nesta A, Zhao Y, Mays AD, Huynh R, Rossi R, et al. A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. *Sci Adv.* 2022;8(3):1–15.
 85. Zuo C, Blow M, Sreedasyam A, Kuo RC, Ramamoorthy GK, Torres-Jerez I, et al. Revealing the transcriptomic complexity of switchgrass by PacBio long-read sequencing. *Biotechnol Biofuels* [Internet]. 2018;11(1):1–15. Available from: <https://doi.org/10.1186/s13068-018-1167-z>
 86. Ali A, Thorgaard GH, Salem M. PacBio Iso-Seq Improves the Rainbow Trout Genome Annotation and Identifies Alternative Splicing Associated With Economically Important Phenotypes. *Front Genet.* 2021;12(July):1–20.
 87. 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>
 88. Hoang N V., 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://bmcbgenomics.biomedcentral.com/articles/10.1186/s12864-017-3757-8>
 89. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* [Internet]. 2020;11(1):1–12. Available from: <http://dx.doi.org/10.1038/s41467-020-15171-6>
 90. Kovaka S, Zimin A V., Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* 2019;20(1):1–13.
 91. Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Forner S, et al. A

- technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv* [Internet]. 2020;672931. Available from: <https://www.biorxiv.org/content/10.1101/672931v2%0Ahttps://www.biorxiv.org/content/10.1101/672931v2.abstract>
92. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015;33(3):290–5.
 93. Mikheenko A, Prjibelski AD, Joglekar A, Tilgner HU. Sequencing of individual barcoded cDNAs using Pacific Biosciences and Oxford Nanopore Technologies reveals platform-specific error patterns. *Genome Res*. 2022;32(4):726–37.
 94. 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(9):1760–74.
 95. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res*. 2012;22(9):1775–89.
 96. VanderNoot VA, Langevin SA, Solberg OD, Lane PD, Curtis DJ, Bent ZW, et al. cDNA normalization by hydroxyapatite chromatography to enrich transcriptome diversity in RNA-seq applications. *Biotechniques*. 2012;53(6):373–80.
 97. Andrews-Pfannkoch C, Fadrosch DW, Thorpe J, Williamson SJ. Hydroxyapatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. *Appl Environ Microbiol*. 2010;76(15):5039–45.
 98. Van Hoof A, Parker R. Messenger RNA degradation: Beginning at the end. *Curr Biol*. 2002;12(8):285–7.
 99. Houseley J, Tollervey D. The Many Pathways of RNA Degradation. *Cell* [Internet]. 2009;136(4):763–76. Available from: <http://dx.doi.org/10.1016/j.cell.2009.01.019>
 100. Opitz L, Salinas-Riester G, Grade M, Jung K, Jo P, Emons G, et al. Impact of RNA degradation on gene expression profiling. *BMC Med Genomics*. 2010;3.
 101. Gallego Romero I, Pai AA, Tung J, Gilad Y. RNA-seq: Impact of RNA degradation on

- transcript quantification. *BMC Biol.* 2014;12:1–13.
102. Schmidt WM, Mueller MW. CapSelect: a highly sensitive method for 5' CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. *Nucleic Acids Res.* 1999;27(21):2–5.
 103. Edery I, Chu LL, Sonenberg N, Pelletier J. An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). *Mol Cell Biol.* 1995;15(6):3363–71.
 104. Bai L, Wang Q, Li H mei, Cheng M, Zhang N bo, Li H. 5'-Cap selection methods and their application in full-length cDNA library construction and transcription start site profiling. *J Shanghai Jiaotong Univ.* 2014;19(5):580–6.
 105. Moreton J, Dunham SP, Emes RD. A consensus approach to vertebrate de novo transcriptome assembly from RNA-seq data: Assembly of the duck (*Anas platyrhynchos*) transcriptome. *Front Genet.* 2014;5(JUN):1–6.
 106. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 2011;29(7):644–52.
 107. Hölzer M, Marz M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *Gigascience.* 2019;8(5):1–16.
 108. Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics.* 2011;12 Suppl 1(Suppl 14).
 109. Lu BX, Zeng ZB, Shi TL. Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Sci China Life Sci.* 2013;56(2):143–55.
 110. Laine VN, Gossmann TI, Van Oers K, Visser ME, Groenen MAM. Exploring the unmapped DNA and RNA reads in a songbird genome. *BMC Genomics.* 2019;20(1):1–12.
 111. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal.* 2011;17:10–2.
 112. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and

- transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc.* 2012;7(3):562–78.
113. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3).
 114. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* [Internet]. 2015 Feb 18 [cited 2015 Feb 18];33(3). Available from: <http://www.nature.com/doifinder/10.1038/nbt.3122>
 115. Martin J a, Wang Z. Next-generation transcriptome assembly. *Nat Rev Genet* [Internet]. 2011 Oct [cited 2014 Jul 11];12(10):671–82. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21897427>
 116. Salzberg SL. Next-generation genome annotation: We still struggle to get it right. *Genome Biol.* 2019;20(1):19–21.
 117. Zhang R, Calixto CPG, Tzioutziou N a, James AB, Simpson CG, Guo W, et al. AtRTD - a comprehensive reference transcript dataset resource for accurate quantification of transcript-specific expression in *Arabidopsis thaliana*. *New Phytol* [Internet]. 2015 Oct [cited 2015 Sep 3];208(1):96–101. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26111100>
 118. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* 2016;17(1):1–19.
 119. Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley and SLS. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 2013;14(R36).
 120. Trapnell C, Williams B a, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* [Internet]. 2010 May [cited 2014 May 23];28(5):511–5. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3146043&tool=pmcentrez&rendertype=abstract>

121. 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 Genomics*. 2017;18(1):1–19.
122. Sessegolo C, Cruaud C, Da Silva C, Cologne A, Dubarry M, Derrien T, et al. Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Sci Rep [Internet]*. 2019;9(1):1–12. Available from: <http://dx.doi.org/10.1038/s41598-019-51470-9>
123. Furlan M, Delgado-Tejedor A, Mulroney L, Pelizzola M, Novoa EM, Leonardi T. Computational methods for RNA modification detection from nanopore direct RNA sequencing data. *RNA Biol [Internet]*. 2021;18(S1):31–40. Available from: <https://doi.org/10.1080/15476286.2021.1978215>
124. Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol*. 2021;22(1):1–24.
125. Paoli-iseppi R De, Gleeson J, Clark MB. Isoform Age - Splice Isoform Profiling Using Long-Read Technologies. *Front Mol Biosci*. 2021;8(August):1–16.
126. Miller AR, Wijeratne S, McGrath SD, Schieffer KM, Miller KE, Lee K, et al. Pacific Biosciences Fusion and Long Isoform Pipeline for Cancer Transcriptome–Based Resolution of Isoform Complexity. *J Mol Diagnostics [Internet]*. 2022;24(12):1292–306. Available from: <https://doi.org/10.1016/j.jmoldx.2022.09.003>
127. Guan D, Halstead MM, Islas-Trejo AD, Goszczynski DE, Cheng HH, Ross PJ, et al. Prediction of transcript isoforms in 19 chicken tissues by Oxford Nanopore long-read sequencing. *Front Genet*. 2022;13(October):1–13.
128. Ren J, Li Q, Zhang Q, Clinton M, Sun C, Yang N. Systematic screening of long intergenic noncoding RNAs expressed during chicken embryogenesis. *Poult Sci [Internet]*. 2021;100(7):101160. Available from: <https://doi.org/10.1016/j.psj.2021.101160>
129. Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, et al. Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics*. 2020;21(1):1–22.

130. Roach NP, Sadowski N, Alessi AF, Timp W, Taylor J, Kim JK. The full-length transcriptome of *C. Elegans* using direct RNA sequencing. *Genome Res.* 2020;30(2):299–312.
131. Massaiu I, Songia P, Chiesa M, Valerio V, Moschetta D, Alfieri V, et al. Evaluation of oxford nanopore minion rna-seq performance for human primary cells. *Int J Mol Sci.* 2021;22(12).
132. Lima L, Marchet C, Caboche S, da Silva C, Istace B, Aury JM, et al. Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data. *Brief Bioinform.* 2019;21(4):1164–81.
133. Huang J, Chen W, Wang Q, Zhang Y, Liu Q, Yang D. Iso-Seq assembly and functional annotation of full-length transcriptome of turbot (*Scophthalmus maximus*) during bacterial infection. *Mar Genomics [Internet].* 2022;63(April):100954. Available from: <https://doi.org/10.1016/j.margen.2022.100954>
134. Mantere T, Kersten S, Hoischen A. Long-read sequencing emerging in medical genetics. *Front Genet.* 2019;10(MAY):1–14.
135. Hong M, Tao S, Zhang L, Diao LT, Huang X, Huang S, et al. RNA sequencing: new technologies and applications in cancer research. *J Hematol Oncol [Internet].* 2020;13(1):1–16. Available from: <https://doi.org/10.1186/s13045-020-01005-x>
136. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol [Internet].* 2016;34(5):525–7. Available from: <http://www.nature.com/doifinder/10.1038/nbt.3519>
137. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: 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>
138. Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L, et al. CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* 2017;45(W1):W12–6.
139. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:

- Architecture and applications. *BMC Bioinformatics*. 2009;10:1–9.
140. 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>
 141. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* [Internet]. 2005 May 1 [cited 2014 Apr 30];21(9):1859–75. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15728110>
 142. Liu Y, Sun Y, Li Y, Bai H, Xue F, Xu S, et al. Analyses of Long Non-Coding RNA and mRNA profiling using RNA sequencing in chicken testis with extreme sperm motility. *Sci Rep* [Internet]. 2017;7(1):3–10. Available from: <http://dx.doi.org/10.1038/s41598-017-08738-9>
 143. Li H, Cui P, Fu X, Zhang L, Yan W, Zhai Y, et al. Identification and analysis of long non-coding RNAs and mRNAs in chicken macrophages infected with avian infectious bronchitis coronavirus. *BMC Genomics*. 2021;22(1):1–13.
 144. Liu J, Zhou Y, Hu X, Yang J, Lei Q, Liu W, et al. Transcriptome Analysis Reveals the Profile of Long Non-coding RNAs During Chicken Muscle Development. *Front Physiol*. 2021;12(May):1–12.
 145. Zhang T, Zhang X, Han K, Zhang G, Wang J, Xie K, et al. Genome-wide analysis of lncRNA and mRNA expression during differentiation of abdominal preadipocytes in the chicken. *G3 Genes, Genomes, Genet*. 2017;7(3):953–66.
 146. Zhang M, Ma X, Zhai Y, Zhang D, Sui L, Li W, et al. Comprehensive Transcriptome Analysis of lncRNAs Reveals the Role of lncAD in Chicken Intramuscular and Abdominal Adipogenesis. *ACS Appl Mater Interfaces*. 2020;
 147. Jehl F, Muret K, Bernard M, Boutin M, Lagoutte L, Désert C, et al. An integrative atlas of chicken long non-coding genes and their annotations across 25 tissues. *Sci Rep* [Internet]. 2020;10(1):1–17. Available from: <https://doi.org/10.1038/s41598-020-77586-x>
 148. Stein L. Genome Annotation: From Sequence to Biology. *Nat Rev Genet*. 2001;2(July):493–503.

149. Klapproth C, Sen R, Stadler PF, Findeiß S, Fallmann J. Common Features in lncRNA Annotation and Classification: A Survey. *Non-Coding RNA*. 2021;7(4):77.
150. Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet* [Internet]. 2018;19(9):535–48. Available from: <http://dx.doi.org/10.1038/s41576-018-0017-y>
151. Sachs AB. Messenger RNA Degradation in Eukaryotes. *Cell Press*. 1993;74:413–21.
152. Cartolano M, Huettel B, Hartwig B, Reinhardt R, Schneeberger K. cDNA library enrichment of full length transcripts for SMRT long read sequencing. *PLoS One*. 2016;11(6):1–10.
153. Keiper BD. Cap-Independent mRNA translation in germ cells. *Int J Mol Sci*. 2019;20(1):17–21.
154. Shatsky IN, Terenin IM, Smirnova V V., Andreev DE. Cap-Independent Translation: What's in a Name? *Trends Biochem Sci* [Internet]. 2018;43(11):882–95. Available from: <https://doi.org/10.1016/j.tibs.2018.04.011>
155. Russell PJ, Hambidge SJ, Kirkegarrrd K. Direct introduction and transient expression of capped and non-capped RNA in *Saccharomyces cerevisiae*. *Nucleic Acids Res*. 1991;19(18):4949–53.
156. Schmidt MJ, Norbury CJ. Polyadenylation and beyond: Emerging roles for noncanonical poly(A) polymerases. *Wiley Interdiscip Rev RNA*. 2010;1(1):142–51.
157. Yu S, Kim VN. A tale of non-canonical tails: gene regulation by post-transcriptional RNA tailing. *Nat Rev Mol Cell Biol* [Internet]. 2020;21(9):542–56. Available from: <http://dx.doi.org/10.1038/s41580-020-0246-8>
158. Mäki A, Tirola M. Directional high-throughput sequencing of RNAs without gene-specific primers. *Biotechniques*. 2018;65(4):219–23.
159. Wongsurawat T, Jenjaroenpun P, Taylor MK, Lee J, Tolardo AL, Parvathareddy J, et al. Rapid sequencing of multiple RNA viruses in their native form. *Front Microbiol*. 2019;10(FEB):1–8.
160. Zhang XO, Yin QF, Chen LL, Yang L. Gene expression profiling of non-polyadenylated RNA-seq across species. *Genomics Data* [Internet]. 2014;2:237–41. Available from:

<http://dx.doi.org/10.1016/j.gdata.2014.07.005>

161. Livyatan I, Harikumar A, Nissim-Rafinia M, Duttagupta R, Gingeras TR, Meshorer E. Non-polyadenylated transcription in embryonic stem cells reveals novel non-coding RNA related to pluripotency and differentiation. *Nucleic Acids Res.* 2013;41(12):6300–15.
162. Yeom KH, Pan Z, Lin CH, Lim HY, Xiao W, Xing Y, et al. Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring. *Genome Res.* 2021;31(6):1106–19.
163. Herzel L, Straube K, Neugebauer KM. Long-read sequencing of nascent RNA reveals coupling among RNA processing events. *Genome Res.* 2018;28(7):1008–19.
164. Zheng JT, Lin CX, Fang ZY, Li HD. Intron Retention as a Mode for RNA-Seq Data Analysis. *Front Genet.* 2020;11(July):1–7.
165. Li HD, Funk CC, Price ND. IREAD: A tool for intron retention detection from RNA-seq data. *BMC Genomics.* 2020;21(1):1–11.
166. Broseus L, Ritchie W. Challenges in detecting and quantifying intron retention from next generation sequencing data. *Comput Struct Biotechnol J [Internet].* 2020;18:501–8. Available from: <https://doi.org/10.1016/j.csbj.2020.02.010>
167. Schulz L, Torres-Diz M, Cortés-López M, Hayer KE, Asnani M, Tasian SK, et al. Direct long-read RNA sequencing identifies a subset of questionable exons likely arising from reverse transcription artifacts. *Genome Biol.* 2021;22(1):1–12.
168. Tardaguila M, De La Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, et al. SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* 2018;28(7):1096.
169. Fungtammasan A, Tomaszewicz M, Campos-Sánchez R, Eckert KA, Degiorgio M, Makova KD. Reverse Transcription Errors and RNA-DNA Differences at Short Tandem Repeats. *Mol Biol Evol.* 2016;33(10):2744–58.
170. Hawkins PR, Jin P, Fu GK. Full-length cDNA synthesis for long-distance RT-PCR of large mRNA transcripts. *Biotechniques.* 2003;34(4):768–73.
171. Slotkin W, Nishikura K. Adenosine-to-inosine RNA editing and human disease. *Genome*

- Med. 2013;5(11):1–13.
172. Frye M, Jaffrey SR, Pan T, Rechavi G, Suzuki T. RNA modifications: What have we learned and where are we headed? *Nat Rev Genet.* 2016;17(6):365–72.
 173. Leger A, Amaral PP, Pandolfini L, Capitanchik C, Capraro F, Miano V, et al. RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat Commun.* 2021;12(1):1–18.
 174. Parker MT, Knop K, Sherwood A V., Schurch NJ, Mackinnon K, Gould PD, et al. Nanopore direct RNA sequencing maps the complexity of arabidopsis mRNA processing and m6A modification. *Elife.* 2020;9:1–35.
 175. Amarasinghe SL, Ritchie ME, Gouil Q. Long-Read-Tools.Org: an Interactive Catalogue of Analysis Methods for Long-Read Sequencing Data. *Gigascience.* 2021;10(2):1–7.
 176. Wyman D, Balderrama-gutierrez G, Reese F, Jiang S, Rahmanian S, Zeng W, et al. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv.* 2019;
 177. Chen Y, Davidson NM, Kei Wan Y, Patel H, Yao F, Meng Low H, et al. A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines Authors. *bioRxiv.* 2021;1:23.
 178. Chen G, Wang C, Shi TL. Overview of available methods for diverse RNA-Seq data analyses. *Sci China Life Sci.* 2011;54(12):1121–8.
 179. Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Lassmann T, et al. A promoter-level mammalian expression atlas. *Nature [Internet].* 2014 Mar 27 [cited 2014 Jul 9];507(7493):462–70. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24670764>
 180. Ivan de la Rubia, Akanksha Srivastava, Wenjing Xue, Joel A Indi, Silvia Carbonell-Sala, Julien Lagarde, M Mar Albà EE. RATTLE: Reference-free reconstruction and quantification of transcriptomes from Nanopore sequencing. *bioRxiv Prepr.* 2022;1–28.
 181. Sahlin K, Sipos B, James PL, Medvedev P. Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nat Commun [Internet].* 2021;12(1):1–13. Available from: <http://dx.doi.org/10.1038/s41467-020-20340-8>
 182. Medina CA, Samac DA, Yu LX. Pan-transcriptome identifying master genes and regulation

- network in response to drought and salt stresses in Alfalfa (*Medicago sativa* L.). *Sci Rep* [Internet]. 2021;11(1):1–16. Available from: <https://doi.org/10.1038/s41598-021-96712-x>
183. Lienhard M, van den Beucken T, Timmermann B, Hochradel M, Boerno S, Caiment F, et al. Long-read transcriptome sequencing analysis with IsoTools. *bioRxiv* [Internet]. 2021;2021.07.13.452091. Available from: <http://biorxiv.org/content/early/2021/07/14/2021.07.13.452091.abstract>
 184. Chern TM, Van Nimwegen E, Kai C, Kawai J, Carninci P, Hayashizaki Y, et al. A simple physical model predicts small exon length variations. *PLoS Genet*. 2006;2(4):606–13.
 185. Sun X, Lin SM, Yan X. Computational evidence of NAGNAG alternative splicing in human large intergenic noncoding RNA. *Biomed Res Int*. 2014;2014.
 186. Rautiainen M, Durai D, Chen Y, Xin L, Low HM, Goeke J, et al. AERON: Transcript quantification and gene-fusion detection using long reads. *bioRxiv* [Internet]. 2020;2020.01.27.921338. Available from: <http://biorxiv.org/content/early/2020/01/27/2020.01.27.921338.abstract>
 187. Sahlin K, Mäkinen V. Accurate spliced alignment of long RNA sequencing reads. *Bioinformatics*. 2021;37(24):4643–51.
 188. Ning C, Ma T, Hu S, Xu Z, Zhang P, Zhao X, et al. Long Non-coding RNA and mRNA Profile of Liver Tissue During Four Developmental Stages in the Chicken. *Front Genet*. 2020;11(June):1–11.
 189. Kern C, Wang Y, Chitwood J, Korf I, Delany M, Cheng H, et al. Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. *BMC Genomics*. 2018;19(1):1–14.
 190. Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, et al. FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res*. 2017;45(8):1–12.
 191. Hu G, Do DN, Davoudi P, Miar Y. Emerging Roles of Non-Coding RNAs in the Feed Efficiency of Livestock Species. *Genes (Basel)*. 2022;13(2).
 192. Gao W, Zhang C, Jin K, Zhang Y, Zuo Q, Li B. Analysis of Incrna expression profile during

- the formation of male germ cells in chickens. *Animals*. 2020;10(10):1–19.
193. Chen X, Abdalla BA, Li Z, Nie Q. Epigenetic regulation by non-coding RNAs in the avian immune system. *Life*. 2020;10(8):1–24.
 194. Zhang Z, Zhang S, Wang G, Feng S, Han K, Han L, et al. Role of microRNA and long non-coding RNA in Marek's disease tumorigenesis in chicken. *Res Vet Sci* [Internet]. 2021;135(January):134–42. Available from: <https://doi.org/10.1016/j.rvsc.2021.01.007>
 195. Zhang G, Sun M, Wang J, Lei M, Li C, Zhao D, et al. PacBio full-length cDNA sequencing integrated with RNA-seq reads drastically improves the discovery of splicing transcripts in rice. *Plant J*. 2019;97(2):296–305.
 196. Dohm JC, Peters P, Stralis-Pavese N, Himmelbauer H. Benchmarking of long-read correction methods. *NAR Genomics Bioinforma*. 2020;2(2):1–12.
 197. Au KF, Underwood JG, Lee L, Wong WH. Improving PacBio Long Read Accuracy by Short Read Alignment. *PLoS One*. 2012;7(10):1–8.
 198. 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*. 2015;10(7):1–15.
 199. Joung J, Engreitz JM, Konermann S, Abudayyeh OO, Verdine VK, Aguet F, et al. Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. *Nature* [Internet]. 2017;548(7667):343–6. Available from: <http://dx.doi.org/10.1038/nature23451>
 200. Haswell JR, Mattioli K, Gerhardinger C, Maass PG, Foster DJ, Peinado P, et al. Genome-wide CRISPR interference screen identifies long non-coding RNA loci required for differentiation and pluripotency. *PLoS One* [Internet]. 2021;16(11 November):1–24. Available from: <http://dx.doi.org/10.1371/journal.pone.0252848>
 201. Knott GJ, Doudna JA. CRISPR-Cas guides the future of genetic engineering. *Science* (80-). 2018;361(6405):866–9.
 202. Panda SK, McGrew MJ. Genome editing of avian species: implications for animal use and welfare. *Lab Anim*. 2022;56(1):50–9.
 203. Goudarzi M, Berg K, Pieper LM, Schier AF. Individual long non-coding RNAs have no overt

- functions in zebrafish embryogenesis, viability and fertility. *Elife*. 2019;8:1–17.
204. S. Zibitt M, Hartford CCR, Lal A. Interrogating lncRNA functions via CRISPR/Cas systems. *RNA Biol* [Internet]. 2021;18(12):2097–106. Available from: <https://doi.org/10.1080/15476286.2021.1899500>
205. Woodcock ME, Gheyas AA, Mason AS, Nandi S, Taylor L, Sherman A, et al. Reviving rare chicken breeds using genetically engineered sterility in surrogate host birds. *Proc Natl Acad Sci U S A*. 2019;116(42):20930–7.
206. Davey MG, Balic A, Rainger J, Sang HM, McGrew MJ. Illuminating the chicken model through genetic modification. *Int J Dev Biol*. 2018;62(1–3):85–92.
207. Idoko-Akoh A, Taylor L, Sang HM, McGrew MJ. High fidelity CRISPR/Cas9 increases precise monoallelic and biallelic editing events in primordial germ cells. *Sci Rep*. 2018;8(1):1–14.