2016

Characterizing Alternative Polyadenylation in Male Germ Cells Using Poly(A)-seq

Holly Tran
University of Rhode Island, hollylhtran@gmail.com

CHARACTERIZING ALTERNATIVE POLYADENYLATION IN MALE GERM CELLS USING POLY(A)-SEQ

BY

HOLLY TRAN

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

CELL AND MOLECULAR BIOLOGY

UNIVERSITY OF RHODE ISLAND

2016

MASTER OF SCIENCE THESIS

OF

HOLLY TRAN

APPROVED:

Thesis Committee:

Major Professor: Becky L. Sartini

Marta Gomez-Chiarri

Ying Zhang

Navindra Seeram

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2016

ABSTRACT

Post-transcriptional processing of mRNA subsequently determines its longevity, transport, and gene expression. Alternative polyadenylation (APA), one form of post-transcriptional processing, is the use of a variant polyadenylation signal and polyA-site for transcript cleavage and addition of adenine residues. The occurrence of this phenomenon in the 3'-untranslated region (3'UTR) can lead to transcript isoforms differing in 3'UTR length and as a result, alter the downstream cis-regulatory elements in use. Transcript modification at the 3'-end in alternative polyadenylation has been shown to foster a number of diseases in altering what would otherwise be normal protein expression and is furthermore emerging as a driver of spermatogenesis.

To date, the global mechanisms that control the post-transcriptional processing in male germ cells remain unknown. PolyA-seq is a strand-specific, quantitative method for the high-throughput sequencing of 3'-ends of transcripts post-transcriptionally modified in polyadenylation. It has the ability to accurately and globally map polyA-sites. To study the molecular regulation of sperm production, an in-depth bioinformatics analysis was performed on available PolyA-seq data to identify male germ cell transcripts that uniquely use alternative polyadenylation, a novel method was developed to isolate and purify male germ cells from testicular tissue, and libraries were prepared for PolyA-seq from isolated male germ cells.

Findings show no significant global difference in polyadenylation signal use between testicular and liver tissues when the same polyadenylation site is compared for human PolyA-seq data. Further annotation suggests a conservation in polyadenylation signal and polyA-site use across tissue types in the same species. Transcripts in the liver

were more likely to use the canonical polyadenylation signal in comparison to those in the testis, lending further evidence of increased variant polyadenylation signal use in male germ cells attributed to alternative polyadenylation. Moreover, manual identification of known alternatively polyadenylated transcripts in testis from mice suggests that PolyA-seq is a reliable method for transcriptome characterization. Isolation and purification of male germ cells was successful using DRAQ5 nuclear stain. Using the verified isolated male germ cells, PolyA-seq libraries were generated. Comparison of different polyadenylation sites for the same transcripts between testis and liver PolyA-Seq still needs to be conducted. Also, increasing yields of polyA+ RNA for PolyA-Seq library prep will facilitate successful sequencing of 3'-ends. Further investigation of male germ cell-specific transcripts associated with alternative polyadenylation will lead to an improved understanding of molecular regulation involved in spermatogenesis and factors that cause male infertility.

ACKNOWLEDGMENTS

This work would not be possible without funding from the National Institutes of Health (NIH) grant #HD072553. Many thanks to Kevin Carlson of Brown University's Flow Facility for processing our flow samples, Christopher Thibeault for his assistance in writing the Perl script, and Dr. Tomas Babak of Queen's University for fielding my numerous questions.

Additionally I would like to express gratitude to my colleagues and dear friends—especially Elizabeth Anderson, Chris Card, Carly Barone, Charles Telekes, Mak Thakur, and Andrea Lesinski—for their scientific guidance, moral support, and willingness to sit through and read through reiterations of this work.

Special thanks to my committee members Dr. Becky Sartini, Dr. Marta Gomez-Chiarri, Dr. Ying Zhang, and Dr. Navindra Seeram for their incredible wealth of knowledge and insights throughout the research and thesis process.

Finally, I would like to especially thank my advisor, Dr. Becky Sartini for not only this opportunity, but also for her constant support, mentorship, and scientific brilliance through the years.

TABLE OF CONTENTS

ABSTRACT ii

ACKNOWLEDGEMENTS iv

TABLE OF CONTENTS v

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER 1: REVIEW OF LITERATURE 1

CHAPTER 2: BIOINFORMATICS ANALYSIS OF DIFFERENTIAL ALTERNATIVE POLYADENYLATION SITE USE IN TESTIS AND LIVER 13

CHAPTER 3: METHOD REFINEMENT OF GERM CELL ISOLATION BY FLOW CYTOMETRY 23

CHAPTER 4: PRODUCTION OF THE POLYA-SEQ LIBRARIES FROM ISOLATED MALE GERM CELL POPULATIONS 30

APPENDICES: I-X 33

FIGURES 61

TABLES 67

BIBLIOGRAPHY 105

LIST OF TABLES

TABLE PAGE

Table 1. Testis-specific genes with alternative poly(A) signals in the 3'UTR. 67

Table 2. Human transcript numerical data 68

Table 3. Human PAS statistics 68

Table 4. Human 1 Honed PAS spread 69

Table 5. Human 2 Honed PAS spread 69

Table 6. Mouse comprehensive numerical data 70

Table 7. Mouse PAS spread for all chromosomes 70

Table 8. Honed mouse PAS spread 70

Table 9. Literature-mined transcripts in PolyA-seq data 71

Table 10. Mouse comprehensive transcript annotation 73

Table 11. Mouse honed transcript subset 80

Table 12. Individual testis and liver tissue transcript annotation numerical data from mouse 81

Table 13. Testis tissue annotation from mouse identifying transcript isoforms 82

Table 14. Liver tissue annotation from mouse identifying transcript isoforms 92

Table 15. Gene ontology for biological processes in mouse annotated transcripts .. 102

Table 16. Isolated male germ cell purity validated with germ cell-specific primers 103

Table 17. Germ cell-specific primers used in PCR validation 103

Table 18. Production of polyA+ mRNA 104

LIST OF FIGURES

FIGURE PAGE

Figure 1. Spermatogenesis. 61

Figure 2. Eukaryotic messenger RNA. 61

Figure 3. Polyadenylation 51

Figure 4. Alternative polyadenylation in the 3'UTR creates transcript isoforms 62

Figure 5. Bioinformatics analysis workflow 63

Figure 6. Graphical representation of PAS use in mouse 64

Figure 7. Size distribution of male germ cells and gating for live cells 65

Figure 8. Two distinct cell populations were identified with DRAQ5 nuclear stain .. 66

Figure 9. PolyA-seq libraries from mouse 67

CHAPTER 1: LITERATURE REVIEW

Male Infertility

Infertility, clinically defined as the inability to bear a child after a year of unprotected intercourse, affects 15% of all couples looking to conceive (Agarwal et al., 2008; Rockett et al., 2004). Over 40% of cases are afflicted by male factor infertility (Jamsai, 2011; Patel, 2011). Male factor infertility can be acquired and/or congenital. Acquired male infertility may result from hormonal imbalances, lifestyle, and environmental factors that affect male sperm count and quality (Saner-Amigh & Halvorson, 2011). Underlying genetic aberrations, such as chromosomal abnormalities and microdeletions in the Y chromosome, account for 10-15% of cases (Ferlin et al., 2007). Genetic mutations impacting the development and function of the endocrine and urogenital system can lead to defective erection, ejaculation, gametogenesis, fertilization, or embryonic development, and results in idiopathic male infertility, which impacts 25% of all infertile men (Roy et al., 2007; Matzuk & Lamb, 2008).

Male infertility is routinely evaluated by semen analysis, where an ejaculate is used to assess abnormal sperm morphology, concentration, and motility. Alone these methods do not serve as an accurate indicator for fertility potential as many other sperm functional components are not assayed. For a significant number of men with normal semen analysis results, full-term pregnancy is still out of reach and the underlying cause of infertility cannot be diagnosed (Garrido et al., 2009; Matzuk & Lamb, 2008; Lee & Foo, 2014). The complex multifactorial etiologies involved in male factor infertility and the limitations of current routine analyses emphasize the need for a more accurate and

universal test. A distinctive aspect of gene expression in male germ cells during spermatogenesis holds promise as a marker for male infertility.

Spermatogenesis

Spermatogenesis is the highly-ordered and precisely timed process of male germ cell development and maturation. The process, which takes place in the testes, occurs uniformly throughout the lifespan of a male. During spermatogenesis, primordial germ cells develop into diploid stem cell spermatogonia in the testicular seminiferous epithelium. After numerous cycles of mitosis and cell renewal, spermatogonia develop into spermatocytes. Spermatocytes complete two meiotic divisions, undergoing genetic recombination and reduction of chromosomes, to produce haploid round spermatids. During the post meiotic phase, round spermatids experience transcriptional arrest as histones are replaced by transition proteins, followed by protamines. The process of unique chromatin remodeling enables for the compaction of chromatins in the sperm head. Spherical round spermatids then enter spermiogenesis, a four-step process where the sperm head and tail terminally differentiate to form haploid spermatozoa with highly condensed DNA (Figure 1). Spermatozoa empty into the lumen of the seminiferous tubules where they go through final maturation and are stored in the cauda epididymis until ejaculation (Eddy, 1998; Gilbert, 2000; Braun, 2001; Senger, 2003; White-Cooper & Davidson, 2011). Development of the male germ cell from spermatogonia to fully differentiated spermatozoa occurs in close association with testicular somatic cells including the supporting Sertoli cells, testosterone-producing Leydig cells, and structural peritubular myoid cells.

Unlike somatic cells, male germ cell function and development is regulated by unique gene expression and endocrine signaling (Eddy, 1998). A number of genes encoding for proteins specific to the structural and functional development of spermatogenic cells, including those involved in meiosis and the specialized differentiation process of spermiogenesis (e.g. transition proteins 1 and 2, protamines 1 and 2, and synaptonemal complex protein 1 genes), are upregulated in testicular tissue and downregulated in somatic tissue—earning the name, "Chauvinist genes" (Eddy, 1998). Unique to male germ cell development is also the increased frequency of alternatively polyadenylated transcripts through the use of variant polyadenylation sites, which is the focus of this research.

Polyadenylation and alternative polyadenylation

In eukaryotes, the nascent RNA is transcribed from DNA containing coding exons and intervening introns. Introns are excised via RNA splicing and the transcript is post-transcriptionally modified with the addition of a 5' cap, transcript cleavage, and addition of a string of adenine residues (poly-A tail) to generate the mature messenger RNA (mRNA). Structurally, mRNA contains a coding region, an area composed of exons that is translated into protein, and untranslated regions (UTRs), non-coding regions that flank both the 5'- and 3'-ends (Figure 2). The post-transcriptional addition of a poly-A tail to a transcript is known as polyadenylation and is required for the synthesis, processing, nuclear export, and stability of eukaryotic mRNA.

Use of variant cleavage sites within a 3'UTR generates transcript isoforms in a process called alternative polyadenylation. To generate an alternatively polyadenylated transcript, selection of different poly(A) signal (PAS) within the 3'UTR by recognition

proteins directs the site of mRNA cleavage and addition of the poly-A tail. The process begins with recognition and binding of the PAS, typically the canonical hexamer sequence AAUAAA, by the protein complex Cleavage and Polyadenylation Specificity Factor (CPSF). After CPSF recognition and binding in the 3'UTR, cleavage stimulation factor (CstF) binds downstream in a GU rich region and CPSF-73, a subunit nuclease of CPSF, cleaves at the polyA-site about 20 nucleotides downstream of the PAS. The polyadenylate polymerase (PAP) enzyme adds about 250-300 adenine nucleotides to the 3'end of the transcript to form the poly-A tail (Di Giammartino et al., 2011; Mandel et al., 2006; Figure 3). Use of a PAS and polyA-site located proximal to the transcript coding region can generate transcript isoforms with short 3'UTRs, compared to use of a PAS and polyA-site distal to the coding region (Figure 4).

The 3'UTR is a non-coding region that contains many sequence elements, such as RNA-binding protein and microRNA (miRNA) binding sites, which are targeted by post-transcriptional regulatory elements. As such, cleavage in the 3'UTR from alternative polyadenylation can result in transcripts with the same coding region and thus production of the same protein, but result in varied 3'UTR lengths. The resulting variation of regulatory element composition on the 3'UTR can alter downstream transcript regulation, stability, and localization (Di Giammartino et al., 2011; Gupta et al., 2014). Alternative polyadenylation can occur in a tissue or cell-specific manner and produce transcripts expressing different 3'UTR lengths in different tissues (Di Giammartino et al., 2011). This research focuses on male germ cell transcripts with a short 3'UTR generated by alternative polyadenylation compared to somatic cell transcripts.

The impact of alternative polyadenylation on transcript use

It is estimated that over 50% of genes in humans and 32% of target genes in mice undergo alternative polyadenylation (Shi, 2012; Tian et al., 2005). The differential regulation of mRNA transcripts can alter gene expression necessary for specific biological processes. For example, short 3'UTR mRNA isoforms have been observed to produce as much as 40-fold more protein by excluding the regulatory impact of miRNAs and other RNA-binding elements (Lianoglou et al., 2013). miRNAs are non-coding RNA that complementarily antisense bind to target mRNA, usually in the 3'UTR, for translational repression in blocking mRNA target site or induction of mRNA degradation through cleavage. Their binding often changes the subsequent timing of translation and transcript expression levels (Bartel, 2004; Mayr & Bartel, 2009; Liaw et al., 2013). Loss of miRNA regulation is estimated to account for 25-67% of increased protein expression related to truncated 3'UTRs, leaving the remainder of difference in protein expression between long and short mRNAs attributed to repression from other regulatory elements lost in alternative polyadenylation (Mayr & Bartel, 2009).

Alterations in 3'UTR length facilitate the exclusion or inclusion of sequence motifs that can affect the lifespan of an mRNA transcript. Short mRNA isoforms were on average 2.6 times more stable than their long mRNA isoforms in the different tissues and genes (Mayr & Bartel, 2009) while isoforms differing in a single nucleotide in the 3'-end dramatically affected the decay rate (Gupta et al., 2014). Different polyA-site selection in alternative polyadenylation can furthermore affect transcript localization patterns. Sequence motifs present in the 3'UTR attract RNA-binding proteins that signal molecular motor transport of transcripts. In neurons, learning-related plasticity and

memory storage require specific subcellular transport of transcripts to axons or dendrites (Mitra et al., 2015; Martin et al., 1997). Brain-derived neurotrophic factor (BDNF) transcripts are expressed as long isoforms in the cell soma and dendrite. Mice expressing only short isoforms of BDNF experienced long-lasting impairment in dendritic potentiation not seen in the cell soma (Martin et al., 1997).

Alternative polyadenylation is emerging as a characteristic trait of highly proliferative cells. In cancer cells and neuronal stem cells, this observation primarily manifests as short 3'UTR leading to loss of miRNA binding sites and an increase in protein production. In this case, truncated 3'UTRs result in a degree of miRNA exclusion that would otherwise keep cell growth in check, resulting in enhanced stability and protein production (Di Giammartino et al., 2011; Mayr & Bartel, 2009). Direct sequencing has revealed that cancer tissues preferentially express isoforms with the shortest 3'UTR (Lin et al., 2012). In a subset of mantle cell lymphoma, short 3'UTR isoforms in *CCNDI* lead to a 1.6-fold increase in Cyclin D1 expression, correlating with increased proliferation of lymphoma cell (Mayr & Bartel, 2009). Furthermore, short 3'UTR isoforms of proto-oncogene insulin-like growth factor 2 mRNA binding protein 1 (*IGF2BP1/IMP-1*) have been shown to exhibit increased oncogenic transformation in comparison to its full-length counterpart (Mayr & Bartel, 2009).

Short 3'UTRs were also more commonly found in highly expressed genes in a study analyzing murine C2C12 myoblasts in their differentiated and proliferative states. The same phenomenon of high prevalence of short 3'UTRs associated with upregulation is present in the analyses of breast cancer cell lines compared with normal breast tissue,

and tumor necrosis factor alpha (TNFα)-treated lymphoblasts compared with untreated cells (Ji et al., 2011).

IMP-1 is an RNA-binding protein heavily expressed in numerous types of human cancers, including lung, colon, and breast cancer. It is believed to play a role in stabilizing target mRNAs such as β-catenin, c-myc, and β-TrCP1 (Mayr & Bartel, 2009). Expression vectors with full-length *IMP-1* isoforms, designed with mutations to prevent cleavage from alternative polyadenylation, produced results relative to an empty vector. In contrast, expression of the short isoforms significantly promoted cellular transformation due to loss of miRNAs. The short isoform was able to transform human breast epithelia and fibroblast cell lines, demonstrating that loss of repressive regulatory 3'UTR elements through different polyA-site use can enhance protein expression and translationally promote tumorigenesis and the pathogenesis of cancer (Mayr & Bartel, 2009).

Alternative polyadenylation in male germ cells

As with highly proliferative cells, alternative polyadenylation is emerging as a driver for male germ cell gene expression. Previous studies have reported variation in the use of specific polyadenylation signals between pre-meiotic spermatogonia and post-meiotic round spermatids, with round spermatids using the canonical polyadenylation signal AA(U/T)AAA less frequently (Lui et al., 2007). Moreover, round spermatids had a shorter median length for the 3'UTR (Lui et al., 2007). High usage of non-canonical polyadenylation signal and poly-A site in spermatogenic cells result in an overexpression of the short 3'UTR isoform compared to somatic cells. This suggests that alternative polyadenylation functions to regulate male germ cell transcripts

(Liu et al., 2007; McMahon et al., 2006; Sartini et al., 2008). Using ESTs, differences in 3'-processing site usage and median 3'UTR lengths were found for male germ cells: spermatogonia (220 nt), spermatocytes (260 nt), and round spermatids (150 nt), while somatic Sertoli cells had a higher median length (400 nt) (Liu et al., 2006).

Studies of individual transcripts show that several testicular transcripts exist in a short 3'UTR isoform when compared to somatic transcript expression (Table 1). Several transcripts are expressed as testis-specific isoforms with a short 3'UTR in addition to longer 3'UTR isoforms that are also expressed in testis and somatic tissues (i.e. Bzw1, Cpsf6, Nudt21, RNF4, and RanGAP1). Two 3'UTR isoforms can also been expressed exclusively in male germ cells. For example, the transcript Deleted in AZoospermia Associated Protein 1 (DAZAP1) exists as two different 3'UTR isoforms in male germ cells. DAZAP1 is a heterogeneous ribonucleoprotein particle highly expressed in late stage spermatocytes and post-meiotic spermatids that regulates translation during spermatogenesis. In the absence of *Dazap1*, mice experience arrest in spermatogenesis and growth retardation. *Dazap1* 3'UTR transcript isoforms are generated through APA--2.4-kb (*Dazap1-L*) and 1.8-kb (*Dazap1-S*). Translation of the two *Dazap1* transcripts are differentially regulated, with *Dazap1-S* exhibiting translational repression associated with inactive messenger ribonucleoprotein particles and a longer poly(A)-tail in comparison to *Dazap-L* (Yang & Yen, 2013).

The 3' UTR contributes to gene regulation in male germ cells. In transgenic mice, the 3'UTR of normal transition protein 2 (*Tnp2*) was replaced with the 3'UTR from the human growth hormone gene (hGH). Disruption of the normal 3'UTR resulted in premature translation of *Tnp2* and infertile males with defective spermatid

morphology and failure of spermatozoa formation (Tseden et al., 2007). Similarly, generated transgenic mice with *Tnp2* 3'UTR replaced with the 3'UTR of hGH resulted in abnormal sperm head morphology, reduced sperm motility, and male infertility (Tseden et al., 2007).

Evidence of further germ cell-specific characteristics can be observed in the polyadenylation process. Mammalian germ cells have been found to express a cell-specific variant of *Cstf2*, a subunit of the CstF complex, known as Cstf2t (Dass et al., 2007). The variant is primarily expressed in meiotic pachytene spermatocytes and post-meiotic spermatids. A germ cell-specific variant of cytoplasmic poly(A) polymerase has been similarly been identified (Liu et al., 2007). These germ cell-specific factors suggest that polyadenylation and 3'-processing may differ in male germ cells, and contribute to specialized germ cell function.

While it is evident that the 3'UTR impacts post-transcriptional modification, gene expression, and its truncation in alternative polyadenylation appears to be characteristic of male germ cells, the importance in the use of a different polyA-site and PAS within the 3'UTR during spermatogenesis is not known. To elucidate the functional purpose of alternative polyadenylation in male germ cell development, a global assessment of all transcripts is necessary.

Global mapping of alternative polyadenylation (APA) transcripts

Several hundred APA transcripts from a variety of tissue types have been identified and studied by Northern blotting. These RNA gels typically require lengthy incubations, and radioactive RNA probes for hybridization but can identify different mRNA isoforms for only one gene at a time. Analysis of EST databases was a first

attempt to identify more APA genes in humans and mice. With EST libraries, polyadenylation sites were computationally derived, but high cost and limited data did not make this a viable approach for global analysis (Shi, 2012; Tian et al., 2005; Elkon et al., 2013). Microarrays have been instrumental in APA gene profiling with the ability to measure quantitative gene expression. The ratio of RNA probe fluorescent intensities are used to detect widespread APA changes. However, microarrays, can only capture verified transcripts and limit data output only to what is already known about the transcriptome. Thus, they cannot directly map, accurately quantify, or detect novel polyadenylation signals (Shi, 2012).

High-throughput sequencing plays a critical role in determining the overall usage and function of APA in specific tissues. RNA 3'UTR and PAS-use profiles on various tissues have been characterized in mammals, yeast, fruit fly (*Drosophila*), nematode (*Caenorhabdits elegans*), and thale cress (*Arabidopsis*) using high-throughput sequencing. This approach has allowed for more insight into tissue-specific use of APA, discoveries of new APA regulators, and changes in APA linked to human diseases (Shi, 2012). While several different sequencing approaches have been developed—poly (A) capture, sequencing alternative polyA-sites (SAPAS), direct RNA sequencing (DRS), 3'READS (Hogue et al., 2014), MAPS (Zhou et al., 2014), RNA-Seq direct sequencing (Ozsolak et al., 2010), PA-Seq (Ni et al., 2013), PAS-Seq (Shepard et al., 2011), and multiplex RNA-seq (Fox-Walsh et al., 2011)—the focus of this research is using the Poly(A)-Sequencing (Sun et al., 2012, Derti et al., 2012).

PolyA-seq is a strand-specific, quantitative method for sequencing 3' polyadenylated ends (Derti et al., 2012). PolyA-seq is a modification of the deep

sequencing-based method, Poly(A) Site Sequencing (PAS-Seq) that can quantitate polyadenylation in RNA. PolyA-seq uses a modified RNA-Seq cDNA library protocol to immediately capture transcript sequence upstream of the polyA-site, preserve strand specificity, represent transcriptional abundance, and quantitate the amount of use in any given polyA-site. The method uses random priming in second-strand synthesis to rapidly generate a library. Internal priming events, which may occur at internal transcript A-rich sequences and account for false discovery rates, are eliminated through calibrated and rigorous filtering. Aligned 5' ends with unaligned adenine stretches on the 3'-end are considered as authentic polyA-sites. Annotated 3'UTRs are clustered and known and novel polyA-sites mapped (Derti et al., 2012). The process involves the conversion of RNA into a cDNA library (typically fragments of 30-400 base pairs), followed by the attachment of sequence adaptors to each fragment (either to one or both ends), the binding of fragments to flow cells, solid-phase bridge amplification via PCR, and the repeated base-by-base sequencing of fragments using fluorescence emission. The resulting reads are then aligned to a reference genome or assembled *de novo*, to produce a genome-scale map. Identical reads are clustered and quantified to give an accurate output on expression levels, relative abundance, and transcriptional structure (Wang et al., 2009).

Presently, PolyA-seq has been used to globally map polyA-sites in twenty-four tissues types in dog, human, mouse, and rat, and has discovered almost 300,000 novel sites (Derti et al., 2012). While this study sequenced the polyA-sites in testis, the presence of somatic cells obscures the APA events for the developing male germ cells. A genome-wide profile on male germ cell polyadenylation signal (PAS) and polyA-site

use has yet to be established and presents a knowledge gap that can potentially lead to greater understanding of APA in spermatogenesis. Characterization of male germ cell functional PAS and polyA-sites involved in APA using PolyA-seq has the potential to increase understanding in the molecular regulation of sperm production and identify precise causes of male infertility.

The aims of this project: 1) identify transcripts in testicular tissue with short isoforms from existing literature, 2) identify transcripts that use a different PAS in testis compared to the liver at the same polyA-site location as determined by its genomic coordinate, 3) develop a novel germ cell isolation method, and 4) generate PolyA-seq libraries from isolated male germ cells. Identified candidate transcripts will then be used for further functional studies of the role of the 3'UTR. This study hypothesizes that transcripts will be identified in male germs cells that have shorter 3'UTR transcripts in comparison to somatic tissues.

CHAPTER 2: BIOINFORMATICS ANALYSIS OF DIFFERENTIAL ALTERNATIVE POLYADENYLATION SITE USE IN TESTIS AND LIVER

INTRODUCTION

Alternative polyadenylation (APA) occurring in the 3'UTR generates transcripts with a different length 3' untranslated region (3'UTR). APA is characteristic of highly proliferative cells such as cancer cells, neurons, and is emerging as a defining trait of spermatogenic cells (Liu et al., 2007; McMahon et al., 2006; Sartini et al., 2008). Male germ cells have a prevalence for transcripts with a short 3'UTR compared to somatic testicular cells and among different male germ cell types as found by analyzing EST libraries (Liu et al., 2007). The truncated 3'UTR in male germ cell transcript suggest that specific regulation of mRNA translation or metabolism is unique in developing germ cells. Characterization of male germ cell functional PAS, the signal for cleavage of a 3'UTR therefore determining its length, has the potential to increase understanding in the molecular regulation of sperm production and identify precise causes of male infertility. However, the role of APA in male germ cells is not completely understood.

PolyA-seq is a novel high-throughput, strand-specific quantitative method for sequencing 3' polyadenylated ends (Derti et al., 2012). PolyA-seq is a modification of the high-throughput RNA Sequencing (RNA-Seq) method, which can sequence total or fractionated RNA; and the PAS-seq method, which can quantitate polyadenylation in RNA transcripts. It is not limited to characterized transcripts and is powerful for the detection of mRNA isoform expression and quantification. PolyA-seq has been used to analyze polyA-site use in several tissues and has found that over 69% of known human genes have multiple polyA-sites in the 3'UTR (Derti et al., 2012). In addition, polyA-

site usage appears to be similar across same types of tissues in different species rather than within the same species, indicating that there may be conserved tissue-specific regulation (Derti et al., 2012).

PolyA-seq data for mouse testis and liver was analyzed to compare reported PAS use with previously published transcripts to determine if PolyA-seq is a reliable predictor of tissue-specific PAS selection. Bioinformatics analysis of available human testis and liver PolyA-seq data was conducted to determine if there is a difference in PAS use in testis in comparison to liver somatic tissue at the same polyA-site (Figure 5). To test this hypothesis, PAS and number of reads for testis and liver from human and mouse from previously published PolyA-seq data (Derti et al., 2012) were compared to discern differential expression. Transcripts with use of the same polyA-site between testis and liver were annotated and characterized using gene ontology. Identification of unique transcripts involved in spermatogenesis has the potential to elucidate targets that may provide further insights in the diagnosis and treatment of male infertility.

METHODS

Obtaining PolyA-seq raw data

Since experimental design uses tissue samples from mouse model organisms with the ultimate end goal of scientific discovery with beneficial implications in human male reproductive biology, sequencing data was collected from species *Homo sapiens*, sequenced with biological replicates, and *Mus musculus*. PolyA-seq text files were downloaded from ArrayExpress Archive of Functional Genomics Data (http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-30198/samples/) for testis and liver tissue for both respective organisms (Derti et al., 2012). PolyA-Seq text files

contained information for polyA-site location, PAS, number of reads, and polyA score, a log-likelihood score reflecting the probability that the observed site is genuine and not an internal priming event.

Manual searching for known testis APA transcripts

Known testis-specific transcripts involved in alternative polyadenylation were manually identified in the data set by searching for genomic coordinates for the transcript of interest (mouse NCBI37/mm9 July 2007) with the predicted PAS in the PolyA-Seq for testis and liver. Transcripts were further validated by comparing polyadenylation signal(s) in the data set with reported polyadenylation signal(s) in published literature.

Data parsing and gene annotation of testis and liver PolyA-seq

A program was written in Perl to open the sequencing text files for testis and liver, match transcripts according to their polyA-site location with a margin of 100bp, and create an output file containing associated information for tissue-specific comparison of PAS usage. For the subsequent processing steps, codes were written in visual basic for applications (VBA) to develop macros within Excel to compare testis and liver PolyA-seq for PAS usage at the genomic coordinate, which in this case represents the location of the polyA-site used. Current sequencing and annotation data from the National Center of Biotechnology Information (NCBI) was used as reference files. Chromosome assembly data from Genome Reference Consortium Human Build 38 patch release 2 (ftp://ftp.ncbi.nih.gov/genomes/Homo_sapiens) and Genome Reference Consortium Mouse Build 38 patch release 3 (ftp://ftp.ncbi.nih.gov/genomes/M_musculus/), were extracted for genomic coordinate

ranges, gene short names, full gene names, and accession numbers for all contained mRNA and miscellaneous RNA identifiers. This collection is an amalgamation of computational sequencing and annotation, manual curation, and predictive annotation. The Genome Reference Consortium Mouse Build 38 patch release 3, part of the mouse RefSeq project assembly on the C57BL/6J strain, at the time of its release date in February 2015, represents 71% of total protein-encoded transcripts annotated and 38% of which with known RefSeq accessions (McGarvey et al., 2015). The data then became the processed reference files used to match the output file for each respective organism.

Comparison of testis and liver polyA-signal use at the same polyA-site

For each chromosome, transcript polyA-site genomic coordinates were matched with coding region genomic coordinates as reported by NCBI with a margin of 100bp. For each successfully referenced match, output data printed information on reported gene name, gene short name, and accession number if available to yield a list of annotated transcripts. From the full list of transcripts, a subset, termed “honed”, was created with transcripts where reads in testis were greater than those in liver and both tissues utilized a different PAS at the same polyA-site. This function was performed by first determining if the PAS in testis matched with the PAS in liver. Transcripts with the same PAS were then deleted, leaving only transcripts with varying polyadenylation signals for the same chromosome and polyA-site within 100bp. Transcripts were then sorted in ascending order by difference between testis and liver number of reads and anything equal to or less than 0 were deleted, leaving only transcripts where testis reads are greater than liver reads by at least one.

Complete polyadenylation signals and respective transcript frequencies were recorded for both groups—comprehensive and honed transcripts—and both tissue types. Student's unpaired T-test was performed on human data, where replicates were available, to determine statistical significance in PAS usage. Due to limitations of the reference annotation and sequencing data, annotations failed to exceed a certain genomic coordinate threshold. For specific known transcripts of interest, manual annotations were completed. Gene ontology analysis was conducted using the Protein ANalysis THrough Evolutionary Relationships (PANTHER) Classification System (pantherdb.org) to classify biological processes.

Individual tissue annotation and analysis using comprehensive PolyA-seq data

Raw PolyA-seq mouse data for testicular and liver tissue were individually annotated to identify potential transcript isoforms resulting from APA. PolyA-site genomic coordinates from were matched with reference files with a 100bp margin to produce an output annotation file containing chromosome number, polyA-site, polyA-score, PAS, number of reads, strand sense, accession number, short gene name, and long gene name. From this, transcript isoforms were identified from annotations using a different polyA-site and an identifiable PAS.

RESULTS

PolyA-seq data for mouse testis and liver was manually searched to compare reported PAS use with previously published transcripts to determine if PolyA-seq is a reliable predictor of tissue-specific PAS selection. Manual mining of the testis and liver PolyA-seq for known APA testis transcripts identified 10 of 14 known transcripts. PAS for five known APA transcripts (*NUDT21, ARF, eIF-2a, ERCC6 and Klf4*) were not

identified in either the testis or liver PolyA-seq data. The PAS identifying the 3'UTR end for 6 of these transcripts, identified by Northern blots and, in some cases, 3'RACE, were confirmed in the PolyA-seq data. For three other transcripts: *CPSF6, DAZAP1, DlnB/POLK and SP1,* additional PAS were identified.

In human samples, from the 19,906 and 10,179 individual transcripts identified in the testis and liver PolyA-seq, 0.718-0.844% of the comprehensive transcripts were annotated (Table 2). In order to identify ubiquitous transcripts upregulated in testicular tissue as a potential byproduct of APA, a data subset—so called "honed"--was created. From the comprehensive transcripts, transcripts were parsed where the number of reads in testis were greater than those in liver by at least one and PAS usage differed at the same polyA-site--irrespective of available annotation. The analysis aimed to identify novel transcripts where isoforms may exist due to 3'UTR post-transcriptional modification. From the comprehensive transcripts in human, 5.435% and 5.363% were extracted to form the honed subset (Table 2). Where canonical PAS was used in liver, testis transcripts had an overwhelming preference for PAS: ATTAAA. This preference was similarly observed vice versa where canonical PAS expression in testis showed a dramatic expression of PAS: ATTAAA in liver transcripts. In general, PAS usage at the same polyA-site was not different for both tissue types from human PolyA-seq analysis (Table 3). For the honed subset in humans, both data files showed a predominant variant use of ATTAAA in testis when the canonical PAS in liver is used—and vice versa (Table 4; Table 5).

Of the mouse sample, 34 transcripts out of the comprehensive 6341 transcripts were annotated (Table 6). Comparison of polyA-site use between tissues for the same

transcripts revealed 251 transcripts that use a different PAS in testis versus liver. Overall, transcripts in liver were more likely to use the canonical PAS (AATAAA) than testicular transcripts. In comparing alternative polyadenylation in both tissue types, transcripts in liver using the canonical PAS were most likely to use the alternative PAS: ATTAAA, with signals AGTAAA and TATAAA following up second (Table 7; Table 8).

Literature-mined transcripts were manually identified in the PolyA-seq data based on NCBI sequencing data (2007) and confirmed some transcript isoforms (Table 9). Comprehensive transcripts from mouse testis and liver at the same polyA-site were annotated (Table 10). Table 11 depicts the honed subset generated and sorted to reflect descending gene expression. From the individually annotated tissue transcripts in mouse, 61 transcripts were identified from testis containing at least 2 isoforms—17 of which contained 3 or more transcript isoforms. From mouse liver tissue sample PolyA-seq data, 83 transcripts were identified contained at least 2 transcript isoforms—19 of which contained 3 or more transcript isoforms as signified by variant polyA-site usage (Table 12). These transcripts isoforms were determined through a separate annotation (Table 13; Table 14).

Gene ontology (GO) analysis was performed on the collective PolyA-seq data in mouse where transcripts were matched at the same genomic coordinate between the two tissue types (Table 15). Of the 6,341 transcripts matched in total, 34 were successfully annotated using the NCBI references files. GO terms associated with cell communication, cell cycle, and primary metabolic processes were overrepresented. Functional annotation included biological processes in ranked order: cellular process

(16.6%), of which include functions relating to cell communication and cell cycle; metabolic process (16.6%), including primary metabolic processes; localization (14.2%), dealing with RNA localization and transport; biological regulation (10.7%), developmental process (8.3%), relating to cell differentiation; response to stimulus (8.3%), cellular component organization or biogenesis (7.1%), immune system process (7.1%), multicellular organismal process (7.1%), apoptotic process (1.1%), biological adhesion (1.1%), and reproduction (1.1%), which encompasses gamete production.

DISCUSSION

In this study, available PolyA-seq data in testis and liver was mined and analyzed to determine if testis-specific APA transcripts or patterns of APA use could be identified. PolyA-seq data for mouse testis and liver was manually searched to compare reported PAS use with previously published transcripts to determine if PolyA-seq is a reliable predictor of tissue-specific PAS selection. For 50% of transcripts, PolyA-seq PAS, polyA-sites, and expression levels matched published 3'UTR isoforms identified from Northern blots (Table 9). For two transcripts (CPSF6 & DAZAP1), the PolyA-seq identified more PAS than had previously been found while for two other transcripts (DinB & TP1), where multiple isoforms have been validated, only one PAS was found (Table 9). 35% of the transcripts were not found in the PolyA-seq data, potentially identifying limitations to the poly(A)+ amplification method used to generate the cDNA libraries.

Analysis of PAS selection at the same polyA-site between testis and liver did not reveal any differences in use; however, some transcripts use a different PAS at the same site. This would be expected due to genetic variation but the function of this

different PAS selection needs to be investigated. Findings from this analysis are complementary to previous reports of differential PAS use within the 3'UTR in individual male germ cells compared to somatic testicular Sertoli cells (Liu et al., 1995).

For human data, frequency of PAS use was not different between testis and liver at the same polyA-site. Transcripts in human data were identified that are expressed in both tissues but utilized a different PAS. Some of these transcripts were known to have a short 3'UTR in testis (e.g. CPSF6) while other transcripts are novel. GO analysis revealed that the majority of transcripts are involved in cell cycle regulation and cell communication, although it is difficult to draw direct implications from these results as the population annotated and processed through GO are less than 1% of the total transcripts.

In total, 32 of the same transcripts were identified with isoforms in both testis and liver from mouse. These transcripts—which exist in both tissues—are involved in cell cytokinesis, cell signaling, and translation, suggesting that a fundamental core biological processes require ubiquitous gene expression.

One limitation to this study is the incomplete reference files that were used to create the annotations. To date approximately 70% of the protein-encoding transcripts have been annotated—38% of which contain known RefSeq accession numbers. Furthermore, the methods of this analysis structured the available PolyA-seq data to the same polyA-site used within a margin for PAS comparisons. Future steps may include analyzing more completely annotated transcripts between and across tissue types for PAS usage. Further analysis can be conducted to: compare testis and liver for different

APA transcripts, more than one 3'UTR isoform in testis, and different 3'UTR isoforms in testis versus liver. Analyses may also consider exploring other tissue types to potentially identify tissue-specific APA characteristics.

Potential steps to improve annotation endeavors include modifying the annotation VBA code to match PolyA-seq data to reference transcripts with only an upstream margin since any hits would only reflect the coding region product. Additionally, 3'UTRs have an average length of 800 nt in humans so extending the current 100bp margin may help increase the percent of annotated transcripts.

CHAPTER 3: METHOD REFINEMENT OF GERM CELL ISOLATION BY FLOW CYTOMETRY

INTRODUCTION

Successful development of male germ cells requires progression through different maturation stages (from spermatogonia, spermatocyte, spermatid and spermatozoa) and the support of testicular somatic cells including Sertoli and Leydig cells. Isolation of male germ cells from all developmental stages is essential for further investigation of cell-specific gene expression and regulation requires separation of the germ cell populations with minimal contamination from other germ cells or testicular somatic cells. Procedures to obtain relatively pure populations of spermatogonia, spermatocytes and spermatids have been developed but most are labor intensive and contamination from somatic cells remains. The use of developmental stages of the testis during the time of the emergence of the specific germ cell maturation stage is problematic, although widely used, due to the presence of other germ cell and somatic cell types (Bellve et al., 1977). Methods for obtaining only one cell type at the exclusion of others have focused primarily on spermatogonia stem cells by flow cytometry FACS separation detecting a stem cells specific antibody (Sun et al., 2011). STAPUT density gradients can be utilized to obtain isolated cells of different types from the same animal but this method yield relatively enriched populations but with cross-contamination from other cells and are highly sensitive to mechanical disruption (Liu et al., 2015; Bryant et al., 2013). A recent report of a centrifugation gradient that isolating relatively pure mouse germ cells requires access to a specialized rotor and can be prohibitive (Chang et al., 2011).

More recent flow cytometry protocols have facilitated the separation of relatively pure multiple germ cell populations from the same animal by utilizing the differences in DNA amount in different germ cell types. Spermatogonia are the only diploid germ cell population while meiotic spermatocytes contain double the DNA content (4N) and spermatids are haploid (1N). Staining a suspension of germ cells with a DNA stain has been relatively successful in obtaining populations of spermatocytes and mixed stages of spermatids in the mouse (Bastos et al., 2005). The current protocol requires fixation of the cells to allow for permeability of the Hoechst 33342 DNA stain to enter the cell membrane and bind to the DNA and is therefore prohibitive for downstream DNA and RNA isolation for further gene expression studies. Separation of live germ cell populations has been achieved with Hoechst 33342 but access to a FACS facility with a UV laser (needed to excite the dye) is not available in most flow cytometers. The goal of this research was to develop a method for separation of live germ cell populations using the DNA stain DRAQ5 that does not require a UV laser for excitation but more common air-cooled 488 nm laser.

In this study, a mechanical and enzymatic digestion of testis was coupled with dual nuclear cell staining to isolate two populations of male germ cells, round spermatids and spermatocytes, using FACS. An advantage to this method is high yield of live cells that can be used in downstream RNA extraction and for further gene expression studies. Identity of the isolated germ cell populations was validated by germ cell expression by RT-PCR. Additionally, the ability to isolate multiple cell types from one mouse is both cost-effective and allows for decreased genetic variability for further gene expression studies.

METHODS

Male Germ Cell Isolation

Male germ cell isolations from 6-8 week-old CD-1 strain mice were performed according to adapted published protocols (Getun, 2011). All male mice used in this study were purchased from Harlan Laboratories (Indianapolis, IN, USA) and protocols were approved by the URI Institutional Animal Care and Use Committee (protocol # AN07-05-029). Testes were collected and the tunica albuginea was removed. Decapsulated testes were placed into 15mL conical tubes and digested in Collagenase from 1 mg/mL *Clostridium histolyticum* (Sigma-Aldrich, St. Louis, MO, USA, cat# C5081), Gey's Balanced Salt Solution (GBSS; Sigma-Aldrich, St. Louis, MO, cat# G9779) and 1 mg/mL DNase I Solution (Stem Cell Technologies, Vancouver, Canada, cat# 07900). At this point seminiferous tubules were visible in a cloudy solution. Conical tubes were manually agitated until testicular tubules started to dissociate. The tubes were then agitated horizontally at 120rpm for 15 minutes at room temperature on a Platform Vari-mix test tube rocker (Thermolyne; Thermo Scientific, Ashville, NC, USA). Tubes were left at room temperature for 15 minutes for phase separation. The supernatant was decanted into a new 15-mL conical tube. Here, spermatozoa along with populations of male germ cells, supporting cells, and debris were visible under light microscope. Digestion and agitation steps were repeated for both the pellet and supernatant with reagents only added to the pellet. Digestion steps were repeated again with the addition of 1mg/mL trypsin (Fisher Scientific, Pittsburg, PA, USA, cat# 27250018) in 1mM HCl. The tubes were inverted to mix and then agitated horizontally at 120rpm for 15 minutes at room temperature. Any visible clumps were dissociated

using a plastic Pasteur pipette to mix for 3 minutes. Immediately following manual dissociation, fetal calf bovine serum (Sigma-Aldrich) was added to neutralize the trypsin. DNase I was added, tubes inverted to mix, and another round of digestion and agitation was performed using standard conditions. Samples were filtered through GBSS pre-wetted 20μM nylon net vacuum-driven filtration systems (Millipore Corporation, Billerica, MA, USA, cat# SCNY00020) and stored on ice. Following filtration, the final pellet solution shows little to no debris and cells with round morphology.

Cell Purification by Flow Cytometry

Cells from pellet fractions were counted on the hemocytometer to determine the number of cells obtained, and stained with 5mM DRAQ5 (abcam, Cambridge, MA, USA, cat# ab108410) and counter-stained with propidium iodide solution (PI; Sigma-Aldrich, cat# P4864) according to manufacturer's recommendations based on cell density. Target cells were then separated and quantified based on emission of DRAQ5 (excitation wavelength: 647nm) and dead cell exclusion applied using propidium iodide solution (excitation wavelength: 535nm-617nm) on a BD Influx Cell Sorter (BD Sciences, Brown University, Providence, RI, USA). Live cell populations were gated into hypothesized meiotic spermatocytes and round spermatids and collected in 10% bovine serum albumin (BSA; Sigma-Aldrich)-covered 15mL conical tubes filled with TRIzol LS reagent (Life Technologies, Grand Island, NY, USA, cat# 10296028) to separate samples into RNA, DNA, and protein fractions. Collected samples suspended in TRIzol LS were incubated on ice in preparation for RNA extraction.

RNA Extraction and Isolation

RNA was extracted from the germ cell populations according to Life Technologies manufacturing protocol for TRIzol LS reagent (Chomczynski, 1993). RNA concentrations were determined and purity was verified for ideal 260/280 absorbance ratio >1.8 by NanoDrop 8000 spectrophotometer (GSC Sequencing Center, University of Rhode Island, Kingston, RI, USA). Unused RNA were stored at -80°C.

RT-PCR

RNA was reversed transcribed (RT) into cDNA using reagents and protocol from QuantiTect Reverse Transcription Kit (Qiagen, Hilden, Germany, cat# 205310) for use in polymerase chain reaction (PCR) amplification. Genomic DNA was eliminated in the RT reaction with gDNA Wipeout Buffer, template RNA, and RNase-free water to volume. Samples were incubated for 2 minutes at 42°C and then reverse transcribed with Quantiscript Reverse Transcriptase, Quantiscript Buffer, and RT Primer Mix with incubation for 15 minutes at 42°C and 3 minutes at 95°C to inactivate Quantiscript Reverse Transcriptase. PCRs were performed under standard temperature condition: 94°C for 30 seconds, 35 cycles of 94°C for 30 seconds, primer-dependent annealing temperature for 30 seconds, 72°C for 2 minutes, and final extension for 10 minutes at 72°C with 1X reaction buffer, 1.5mM $MgCl_2$, 10mM dNTPs, 2.5μM gene-specific forward and reverse primers, and 2.5U Taq polymerase (NEB, Ipswich, MA, M0273S). All PCRs were performed with a positive control with mouse testis cDNA, a negative control without template cDNA, and a negative control without enzyme. Amplified PCR products were run on a 2% agarose electrophoresis gel, purified and extracted (Qiagen QIAquick Gel Extraction Kit, cat# 28704), and submitted for

sequencing (URI Genome Sequencing Center). Sequenced amplicons were identified and verified using NCBI BLAST. cDNA was stored at -20°C.

PCR Reactions

PCR reagents and manufacturer protocol from New England BioLabs (Ipswitch, MA) were used for gene amplification. Isolated populations of spermatocytes, round spermatids, and Sertoli cells and were verified for identity and purity using designed transcript-specific primers, including proto-oncogene (c-Kit, spermatogonia), synaptonemal complex protein 3 (SCYP3, spermatocytes), protamine 2 (PRM2, round spermatids), follicle-stimulating hormone (FSHR, Sertoli cells), and cholesterol side-chain cleavage enzyme (P450scc, Leydig cells). Sample populations were tested to ensure absence of contamination. PCR products were separated by a 2% agarose gel electrophoresis, gene amplified band were captured using UV light to fluoresce ethidium bromide, bands were extracted for sequencing (URI Genomics and Sequencing Center), and verified with National Center for Biotechnology Information's Basic Local Alignment Search Tool (NCBI BLAST).

RESULTS

After mechanical and enzymatic disruption of the mouse testes and double staining with PI and DRAQ5, a representative FACS distribution of cells by size is shown in Figure 7 and Figure 8.

Characterization of germ cell populations separated by flow cytometry

The identity of the two germ cell populations, P1 & P2, were investigated by validated using germ cell-specific primers in RT-PCR (Table 18). The absence of testicular somatic Sertoli and Leydig cells were confirmed by the lack of amplification

of *FSHR* and *p450scc* respectively. Diploid spermatogonia were also not detected in these two populations. Positive amplification of *SYCP3* in the P2 population, but lack of amplification in the P1 population, indicated that this population with the highest fluorescence intensity included 4N spermatocytes. Population P1 contained haploid round spermatids as indicated by the amplification of PRM2 and Dbil5.

DISCUSSION

A protocol for FACS isolation of live round spermatid and spermatocytes from adult male mice was developed using the nuclear stain DRAQ5. Purified live cells may be used in further downstream molecular biology applications, including RNA extraction and PCR. In total, this method takes roughly two and a half hours to prepare the cells, an additional three hours to sort cells by FACS, and additional time for post-sort manipulations.

Further modifications of the protocol are necessary to isolate the diploid spermatogonia germ cells in addition to the somatic Sertoli and Leydig cells. The FSC/SSC cell distribution with DRAQ5 did not demonstrate as many distinct cell populations as previously published in fixed cells populations stained with Hoescht 33342. Diploid cells may have been lost in the digestion and filtration steps prior to cell staining but further examination of this stage of the protocol is necessary.

CHAPTER 4: PRODUCTION OF THE POLYA-SEQ LIBRARIES FROM ISOLATED MALE GERM CELL POPULATIONS

INTRODUCTION

PolyA-seq is a robust method of sequencing polyA-site use and determining 3'UTR length in different tissues (Derti et al., 2012). Previous PolyA-seq studies have sequenced polyA+ RNA from whole testicular tissue, including germ cells and somatic cells. The bioinformatics analysis described in Chapter 2 is from testis PolyA-seq and is a mix of germ cells and somatic testicular cells (Sertoli cells, Leydig cells and peritubular cells). The inclusion of somatic tissue may mask the germ cell specific gene expression patterns, making a male germ cell specific pattern of PAS use and identification of transcripts with a short 3'UTR difficult to examine. Based on the method of PolyA-seq library preparation (Derti et al., 2012), sequencing libraries for the isolated germ cell populations separated by DRAQ5 were developed as described in Chapter 3. Briefly, generation of these libraries involved isolating poly(A)$^+$ mRNA from male germ cell total RNA, attaching a d(T) primer to specifically target the PAS and polyA-site, PCR amplification and purification, and gel electrophoresis to visualize the product size.

METHODS

cDNA libraries were generated from two populations of male germ cells (P1 & P2) that were separated by DRAQ5 (Chapter 3). After RNA isolation from Trizol as described in Chapter 3, poly(A)$^+$ RNA was isolated from the total RNA using the Oligotex mRNA Mini Kit (Qiagen, cat# 70022). 150 ng poly(A)$^+$ RNA from each cell

type was reverse transcribed using a $T_{10}VN$ (ten thymidine and non-thymine nucleotide attached to a random base) primer and second strand synthesis completed with random hexamer-PCR anchor primers. U1 and U2 have sequence complementary to Illumina-specific adapters which were added through a nested PCR to maintain strand-specificity and cDNA were fragmented for sequencing. Single-end sequencing in parallel was attempted for the prepared spermatocyte (n=2), round spermatid (n=2) PolyA-seq cDNA libraries at LC Sciences (Houston, TX), but was unsuccessful.

RESULTS

Poly(A)$^+$ mRNA was isolated from spermatocyte (P1) and round spermatid (P2) total RNA for PolyA-seq library preparation, which required 100ng input RNA. Yields from the poly(A)$^+$ enrichment are shown in Table 18. Based on the low poly(A)$^+$ RNA yields, the individual germ cell populations were pooled to make the PolyA-seq libraries that contained only spermatid and spermatocyte male germ cells with biological replicates isolated by FACS on two separate days. Whole testis and liver, used as a somatic tissue control, PolyA Seq libraries were also generated (Figure 8).

DISCUSSION

PolyA-seq libraries were generated from round spermatids, spermatocytes, pooled spermatids and spermatocytes, liver, and whole testis tissues. Only one library each from the spermatids and spermatocytes could be generated due to the low yield of poly(A)$^+$ mRNA. Several more runs of FACS isolation of cells are needed to generate enough poly(A)$^+$ mRNA to potentially be more successful.

Due to the low yield of poly(A)$^+$ mRNA, sorted round spermatid and spermatocyte populations were combined to make a library that contained only male

germ cells. These were more successful due to the larger amount of starting poly(A)$^+$ mRNA. Unfortunately, these libraries would not be sequenced.

It is worth-while to note that about 80% of total RNA from mammalian cells is ribosomal RNA (rRNA). This leaves the remainder comprised of transfer RNA (tRNA) and about 2-3% polyA+ mRNA (Lodish et al., 2000). Furthermore, RNA is susceptible to degradation by RNases. Based on these results and background information, it is recommended to sort for 100,000 events by FACS to yield 2,500µg total RNA, which should yield approximately 1µg poly(A)$^+$ mRNA for library generation. Prior to enrichment of polyA$^+$ mRNA, total RNA should be evaluated for quality and the absence of degradation. Running extracted total RNA on a 1% agarose gel should yield crisp bands for 28S rRNA that is twice the intensity of lower 18S rRNA. Equal rRNA band intensities or mRNA smearing below the rRNA bands indicate poor quality of RNA that should not be used for polyA$^+$ mRNA isolation. Bioanalyzers may also be used to measure rRNA band sizes and will produce an RNA Integrity Number used to quantify RNA quality from RIN 10, representing intact RNA; to RIN 2, indicating degraded RNA.

APPENDIX I

Germ Cell Isolation for Flow Cytometry Protocol with DRAQ5

Supplies Needed—

15mL tube (x6)
Gey's Balance Salt Solution (GBSS)
Collagenase type I
DNase I (x3 50µL aliquots)
Trypsin
Fetal Calf Serum
Aluminum foil
Propidium Iodide
5mL plastic syringe
Plastic Pasteur pipette (x8)
Millipore Conical Tube Filter (x2)
50mL conical tube/Isoflurane
DRAQ5

Solution to prepare—

GBSS + Collagenase Mixture (GBSS/Col):
25mg Collagenase
25mL GBSS

Part I: Testis Preparation

1. Place one decapsulated testis in a 15mL tubes, 2 tubes: 1 for each testis
2. Add to each tube: **3mL GBSS/Col**
   10µL DNase I
3. Shake vigorously by hand until testicular tubules start to dissociate.
4. Agitate horizontally at 120rpm for 15 mins at room temperature.
5. Let sit for 1 minute, then decant supernatant off into separate tube, keep using all tubes (4).
6. Repeat steps 2-4, adding more reagents to the pellet tubes only.
7. Let sit for 15 minutes at room temperature, if pellet forms, decant off the supernatant!
8. Add to each tube: **2.5mL GBSS/Col**
   50µl Trypsin
   10µl DNase I
9. Invert tube several times to mix. Agitate horizontally at 120rpm for 15 mins at 33°C.
10. Pipette up and down gently using plastic Pasteur pipette for 3 min. No clumps should be visible.
11. Add to each tube: **30µl Trypsin**
    10µl DNase I

12. Invert tube several times to mix. Agitate horizontally at 120rpm for 15 mins at 33°C.
13. Add to all tubes: **400µL fetal calf bovine serum (FCS)**, invert tubes to mix.
    10µL DNase I
14. Pass samples through GBSS pre-wetted filters. Wrap all tubes in aluminum foil.
15. **Hemocytometer count.** ***See Part II.***
16. Add to selected tubes: **5µl propidium iodide (PI) solution**
    DRAQ5 (amount based on hemocytometer count)

Part II: Hemocytometer Count

1. Transfer **10µl of sample** into hemocytometer on both sides. (If too much overlap, dilute sample.)
2. Count all 5 squares within triple lines.
3. Take total count in square * 50,000 → number of cells/mL. Dilute solution with GBSS if needed.
4. Add appropriate amount of DRAQ5 according to chart and allow to sit at room temperature for 5-30 minutes before processing through flow cytometer.

Part III: Samples and Reagents to Sort on Flow

1. DRAQ5 + PI Stained
2. DRAQ5 Only
3. PI Only
4. Unstained & extra dye (DRAQ5 & PI)
5. Trizol LS Tubes (1.5mL)

Note: 750µl Trizol LS has been doubled. Aliquot solutions during RNA isolation.

APPENDIX II

RNA Isolation Protocol for Male Germ Cells

Part I: Phase Separation

1. Add **200µL chloroform** per 750µL of TRIzol LS in microcentrifuge tubes and shake vigorously for 15 seconds.
2. Let tubes sit at room temperature for 15 minutes.
3. Centrifuge at 12,000 x *g* for 15 minutes at 4°C.
4. Remove and keep clear aqueous phase (top layer) in new tubes.

Part II: RNA Precipitation

1. Add **2µL glycogen** to each sample.
2. Add **500µL isopropanol** and let samples sit for 15 minutes in -20°C.
3. Centrifuge at 12,000 x *g* for 15 minutes at 4°C.

Part III: RNA Wash and Resuspension

1. Keep samples on ice!
2. Remove supernatant from tube, leaving only RNA pellet.
   (If there is no visible pellet, keep supernatant in conical tubes and leave a layer of liquid where pellet is estimated to be.)
3. Wash the pellet with **1mL 75% ethanol** and leave in -20°C overnight.
4. Centrifuge sample at 7500 x *g* for 10 minutes at 4°C and discard the supernatant.
5. Air dry samples on ice for 10 minutes.
6. Add **25µL RNase-free water** and pipette to mix and dissolve pellet.
7. Keep samples on ice and nanodrop. Store RNA samples in -80°C.

APPENDIX III

Reverse Transcription (RT) – Qiagen QuantiTect Version Protocol

1. Mix together:
   Reagent Volumes per Reaction
   gDNA Wipeout Buffer, 7x → 2µL
   Template RNA → variable (up to 1µL)
   RNase-free water → to total volume of 14µL
2. Incubate for 2 minutes at 42ºC then immediately place on ice.
3. Add:
   Reverse Transcriptase → 1µL
   RT Buffer, 5x → 4µL
   RT Primer Mix → 1µL
4. Incubate for 15 minutes at 42ºC.
5. Incubate for 3 minutes at 95ºC to inactivate Reverse Transcriptase.
6. Hold at 4ºC.
7. Store cDNA at -20ºC.

APPENDIX IV

Polymerase Chain Reaction (PCR) Protocol

Samples

a. cDNA from RT reaction
b. RT negative control: no enzyme
c. PCR negative control: no template RT added

Reagents

Standard Taq Buffer → 5µL
Forward Primer → 4µL
Reverse Primer → 4µL
MgCl2 → 3µL
dNTPs → 1µL
Taq polymerase → 0.5µL
diH2O → 31.5µL
cDNA template → 1µL

Thermocycler conditions

1 cycle: 94ºC for 3 minutes
35 cycles: 94ºC for 30 seconds, ____ºC for 30 seconds, 72ºC for 30 seconds
1 cycle: 72ºC for 10 minutes
Hold at 4ºC

APPENDIX V

NEBNext Oligo d(T)25 Magnetic Beads mRNA Isolation Protocol

Supplies Needed—0.2mL PCR tubes
RNA sample
NEBNext Magnetic Oligo d(T)25 Beads & Kit
Tris Buffer
RNA Binding Buffer
Wash Buffer

Equipment Needed—
Thermocycler
Magnetic rack/plate

Starting Material: 1–5 µg of total RNA.

1. Dilute total RNA with nuclease-free water to a final volume of 50µl in a PCR tube.
2. Add to a new PCR tube: **15µl NEBNext Magnetic Oligo d(T)25 Beads**
3. Wash beads two times with **100µl RNA Binding Buffer**. Pellet with magnet and discard supernatant.
4. Resuspend beads in **50µl RNA Binding Buffer** and add to tube: **50µl total RNA sample** from step 1.
5. Place the tubes in thermocycler for conditions: 65°C for 5 minutes, hold at 4°C to denature the RNA and facilitate binding of the poly-A-RNA to the beads and remove sample.
6. Place the tubes on the bench and incubate at room temperature for 5 minutes to allow the RNA to bind to the beads.
7. Place the tubes on the magnetic rack at room temperature for 2 minutes to separate the poly-A RNA bound to the beads from solution.
8. Remove and discard all supernatant. Do not disturb beads!
9. Remove the plate from the magnetic rack.
10. Wash the beads twice with **200µl Wash Buffer** to remove unbound RNA. Pipette the up and down 6 times to mix.
11. Place the tubes on the magnetic rack at room temperature for 2 minutes.
12. Remove and discard all supernatant from each well of the plate using a multichannel pipette. Do not disturb beads!
13. Remove tubes from the magnetic rack.
14. Add **50µl Tris Buffer** to each well of plate. Gently pipette up and down 6 times to mix.
15. Place the tubes in thermocycler for conditions: 80°C for 2 minutes, hold at 25°C for sample removal.
16. Add **50µl RNA binding buffer** to each sample to allow RNA to bind to beads. Gently pipette up and down 6 times.
17. Incubate the tubes at room temperature for 5 minutes.
18. Place the tubes on the magnetic stand at room temperature for 2 minutes.

19. Remove and discard all of the supernatant from each tube. Do not disturb beads!
20. Remove the tubes from magnetic stand.
21. Wash the beads twice with **200µl Wash Buffer**. Gently pipette up and down 6 times.
22. Place the tubes on magnetic stand at room temperature for 2 minutes.
23. Remove and discard all supernatant from each tube. Do not disturb beads!
24. Remove tubes from magnetic stand.
25. Elute the mRNA from the beads by adding **17µl Tris Buffer** and incubating sample at 80°C for 2 minutes. Immediately place tubes on magnetic rack.
26. Collect purified mRNA by transferring the supernatant to a clean PCR tube.
27. Place tube on ice and nanodrop.

Thermocycler Conditions

65°C: 5 minutes
Hold: 4°C
80°C: 2 minutes
Hold: 25°C
80°C: 2 minutes
Hold: 4°C

APPENDIX VI

PolyA-seq Protocol

Supplies Needed—Primers (TB1003, TB-rev (GA), TB1007, TB1002, TB1005)
Elution water
5X buffer
dNTP (25mM)
DTT
RNase OUT
Superscript III Reverse Transcriptase
RNase H
10X NE Buffer 2
Klenow Enzyme
$MgCl_2$
Roche HF Enzyme Mix
AMPure XP Beads
DMSO

Preparation

1. Heat microcentrifuge tube with nuclease-free elution water to 55°C for part II.
2. Reconstitute primers to 50µM and 10µM. TB1003 primer will be reconstituted from 50µM to 0.1µM.

Part I: First Strand Synthesis Reaction (Invitrogen)

1. Add and mix following in a PCR strip tube:
   x µl polyA + RNA (100ng)
   2µl TB1003 primer (0.1µM) **
   *****[Add 656µl of primer water to blue top to reach 50µM. Then 1µl of 50µM + 499µl primer water.]***
2. Bring volume to 11.4µl with water.
3. Heat to 65°C for 10 minutes and hold at 4°C.
4. Add and mix following in PCR strip tube:
   4µl 5X buffer (Invitrogen)
   1.6µl dNTPs (25mM)
   1µl DTT
   1µl RNase OUT (Invitrogen)
   1µl Superscript III Reverse Transcriptase (Invitrogen)
5. Incubate at 40°C for 90 minutes, 70°C for 15 minutes, and cool to 4°C.
6. Add to tube and mix: **1µl RNase H (Invitrogen)**
7. Incubate at 37°C for 20 minutes, 75°C for 15 minutes, and cool to 4°C.

Part II: First Qiagen Purification

Add and mix **79µl water**.
Add **500µl buffer PB** and pass over qiaquick at 13,000*g* for 1 minute.

Wash with **750µl wash buffer** and spin at 13,000*g* for 2 minutes.
In new microcentrifuge tube, elute twice with **30µl elution water (55°C from preparation step).**
Let sit for 1 minute before spinning at 13,000*g* for 2 minutes.

Part III: Second Strand Synthesis Reaction

1. Add **40µl of Second Strand mix (below)** to 1st strand product (60µl from above) in a PCR tube:
   11.7µl water
   10µl 10XNE Buffer 2
   10µl TB-rev (GA) - 10µM
   5µl 10mM dNTPs
   3.3µl Klenow enzyme (5U/µl of NEB exo' Klenow #M0212L)
2. Incubate at 37°C for 30 minutes in thermocycler.

Part IV: Ampure XP-Purification

1. Add **180µl Agencourt AMPure XP Beads** and let sit for 5 minutes at room temperature.
2. Separate supernatant using magnetic rack and discard supernatant.
3. Wash twice with **70% EtOH**. Remove as much as possible on magnetic stand and dry off residual EtOH on ice.
4. Elute with **50µl elution buffer**.
5. Nanodrop to check for DNA concentration.

Part VI: PCR

1. In new PCR tube, add **31µl of purified Second Strand Synthesis product** to **18µl of PCR mix:**
   10µl 10X Buffer (Roche Reagent #2)
   2.5µl 99.9% DMSO
   1µl 10mM dNTP
   2µl 10µM TB1007 primer
   2µl 10µM TB 1002 primer
   1µL 25mM $MgCl_2$ (Roche Reagent #4)
2. Add to each tube **0.5µl Roche HF Enzyme Mix** at 4°C and mix by pipette.
2. Amplify using conditions:
   2 Cycles—
   94°C for 10 seconds
   40°C for 2 minutes
   72°C for 1 minute

   8 Cycles—
   94°C for 10 seconds
   60°C for 30 seconds
   72°C for 1 minute

20 Cycles—
94°C for 15 seconds
60°C for 30 seconds
72°C for 1 minute + 10 seconds

72°C for 5 minutes
4°C hold

Part VII: Second Ampure Purifucation

1. Add **90µl Agencourt AMPure XP Beads** and let sit for 5 minutes at room temperature.
2. Wash twice with **70% EtOH**.
3. Elute with **50µl elution buffer.**

Part VIII: Post PCR Analysis

1. Nanodrop.
2. Run PCR reaction out on gel. Should see band ranging from ~80-250nt.

Part VIV: Sequencing

1. Standard Illumina set-up using TB1005 as sequencing primer.

APPENDIX VII

Step 1 VBA: Delete Rows with Same PAS

**Note: This code is written in Visual Basic for Applications (VBA) for Excel macro development. Any line that starts with an apostrophe indicates a comment.*

```
Sub remove()
  Dim sr1 As Long
  'This code works with the assumption that you have the first row as a header.
  'If there is no header, just add in a blank row.
  'Active sheet for this was written as "Sheet4".
  'This sheet name can be changed or you can re-name your sheet to match the code.

  'MISC. NOTES & DIRECTIONS
  'For easier visualization check purposes, I choose to perform is TRUE/FALSE
function before running this.
  'In Excel, go to the "Formula" tab --> "Text" --> "EXACT".
  'Select your two cells of polyadenylation signals (PAS) to compare.
  'Carry out this function to the reminder of the column(s).
  'If your PAS match, in a new column "TRUE" will print. If they differ, "FALSE" will
print.
  'Run code. This will delete all rows where PAS are the same, leaving you with a
bunch of "FALSE"s.
  'Check the last two rows once the code has finished.
  'This isn't perfect so you might have to manually delete the last two if they're
matching.

  For sr1 = 2 To Sheets("Sheet4").Range("D3").CurrentRegion.Rows.Count - 1
        If Sheets("Sheet4").Cells(sr1, 4).Value = Sheets("Sheet4").Cells(sr1, 5).Value
Then
           Sheets("Sheet4").Rows(sr1).EntireRow.Delete
           Exit For
        End If
  Next sr1
End Sub
```

APPENDIX VIII

Step 2 VBA: Extract Data from Reference Files

```
Sub Step2ExtractDataFromRawFiles()

Dim CurrentRow, EndsAt, Progress, NumOfRows As Long
Dim DataSheet, CurrentChrNum, CurrentGeneLong, CurrentID, CurrentGeneShort As
String

CurrentRow = 1
EndsAt = 0

'COUNTS ROWS
Range("A1").CurrentRegion.Select
NumOfRows = Selection.Rows.Count
Cells(1, 1).Select
DataSheet = ActiveSheet.Name

'PROGRESSBAR INITIALIZATION
MakingProgress.Show vbModeless
Progress = 0
OnePercent = 100 * Round(NumOfRows / 10000)
HowManyKs = Left(NumOfRows, Len(Str(NumOfRows)) - 4)

'FIND THE CHROMOSOME #
For i = 1 To 100
   If InStr(Cells(i, 1).Value2, "DEFINITION") > 0 Then

     If InStr(Cells(i, 1).Value2, "unplaced") > 0 Then
          CurrentChrNum = "ChrUn"
        Exit For
     End If
     '                                37 for humans   VV   52 for mice
     CurrentChrNum = "chr" & Replace(Mid(Cells(i, 1).Value2, 52, 2), " ", "")
     Exit For
   End If
Next i

MsgBox ("Processing data for " & CurrentChrNum & ". It has " & NumOfRows & "
rows. If either of those are wrong, do a manual fix in the code.")
'NumOfRows = if it's not correct, then write the actual number here
'CurrentChrNum = if it's not correct, then write the actual chr# here. Format, for
example: chr1, chr2, chr10, chr18, chrX, chrUn

'CREATE NEW WORKSHEET CALLED CHR#
```

```
Sheets.Add.Name = CurrentChrNum

'CHANGE THE ACTIVE SHEET BACK TO THE ONE WITH DATA
Sheets(DataSheet).Activate

'START EXTRACTING DATA
For i = 1 To NumOfRows
   'This looks for mRNA, or misc_RNA, and also for numbers because those
keywords appear randomly as well.
   If (InStr(Cells(i, 1).Value2, "misc_RNA") + InStr(Cells(i, 1).Value2, "mRNA") >
0) Then
      If (HaveNumbers(Cells(i, 1)) = True And InStr(Cells(i, 1).Value2, "..") > 0) Then

      Cells(i, 2).Value2 = "X" 'for debugging only, can be deleted

   'ID
   For iplus = 1 To 150
   Where = InStr(Cells(i + iplus, 1), "/transcript_id=")
   If Where > 0 Then
    CurrentID = Mid(Cells(i + iplus, 1), Where + 16, Len(Cells(i + iplus, 1)) - Where -
16)
     Cells(i, 3).Value = CurrentID 'for debugging only, can be deleted
     Exit For
   End If
   Next iplus
   iplus = 0
   j = 0

   'SHORTNAME
   For iplus = 1 To 150
   Where = InStr(Cells(i + iplus, 1), "/gene=")
   If Where > 0 Then
    CurrentGeneShort = Mid(Cells(i + iplus, 1), Where + 7, Len(Cells(i + iplus, 1)) -
Where - 7)
     Cells(i, 4).Value = CurrentGeneShort 'for debugging only, can be deleted
     Exit For
   End If
   Next iplus
   iplus = 0
   j = 0


   'LONGNAME
   For iplus = 1 To 150
   Where = InStr(Cells(i + iplus, 1), "/product=")
   If Where > 0 Then
```

```
    CurrentGeneLong = Mid(Cells(i + iplus, 1), Where + 10, Len(Cells(i + iplus, 1)) -
Where)

    For j = 1 To 20
     If InStr(Cells(i + iplus + j, 1), "         /") > 0 Then
        Do Until InStr(CurrentGeneLong, "  ") = 0
          CurrentGeneLong = Replace(CurrentGeneLong, "  ", " ")
          If Right(CurrentGeneLong, 1) = Chr(34) Then CurrentGeneLong =
Replace(CurrentGeneLong, Chr(34), "")
        Loop
        Exit For
     End If
     CurrentGeneLong = CurrentGeneLong & Cells(i + iplus + j, 1)
    Next j
    If Right(CurrentGeneLong, 1) = Chr(34) Then CurrentGeneLong =
Replace(CurrentGeneLong, Chr(34), "")
    Cells(i, 5).Value = CurrentGeneLong 'for debugging only, can be deleted
    Exit For
   End If
   Next iplus
   iplus = 0
   j = 0


   'INTERVALS
   'If it finds them, looks for the intervals.
    Do
     'Some of them are in multiple cells, so this'll keep checking until it finds a )
     For j = 1 To Len(Cells(i + iplus, 1))
      Select Case Mid(Cells(i + iplus, 1), j, 1)
           Case 0 To 9
             Sheets(CurrentChrNum).Cells(CurrentRow, 2 + EndsAt).Value =
Sheets(CurrentChrNum).Cells(CurrentRow, 2 + EndsAt).Value & Mid(Cells(i + iplus,
1), j, 1)
           Case "."
             EndsAt = 1
           Case ","
             EndsAt = 0

             Sheets(CurrentChrNum).Cells(CurrentRow, 1).Value = CurrentChrNum
             Sheets(CurrentChrNum).Cells(CurrentRow, 4).Value = CurrentID
             Sheets(CurrentChrNum).Cells(CurrentRow, 5).Value = CurrentGeneShort
             Sheets(CurrentChrNum).Cells(CurrentRow, 6).Value = CurrentGeneLong
             CurrentRow = CurrentRow + 1
      End Select
```

```
      Next j

      iplus = iplus + 1
    Loop Until InStr(Cells(i + iplus, 1).Value2, "/") > 0
    Sheets(CurrentChrNum).Cells(CurrentRow, 1).Value = CurrentChrNum
    Sheets(CurrentChrNum).Cells(CurrentRow, 4).Value = CurrentID
    Sheets(CurrentChrNum).Cells(CurrentRow, 5).Value = CurrentGeneShort
    Sheets(CurrentChrNum).Cells(CurrentRow, 6).Value = CurrentGeneLong
    CurrentRow = CurrentRow + 1
    EndsAt = 0
  iplus = 0
  j = 0

  End If
  End If

'TO CHECK LATER IF THE CODE FAILED TO FIND ANYTHING
CurrentID = "MISSINGDATA"
CurrentGeneShort = "MISSINGDATA"
CurrentGeneLong = "MISSINGDATA"
'PROGRESSBAR UPDATE
Progress = Progress + 1
If Right(Progress, 2) = "00" Then
  Progresshundreds = Progresshundreds + 1
  If Progress = OnePercent Then
    MakingProgress.ProgressBar.Width = Round(200 * i / NumOfRows)
    MakingProgress.ProgressText.Caption = Round(100 * i / NumOfRows) & "%"
    Progress = 0
  End If
  MakingProgress.Caption = Format(Progresshundreds / 10, "#.0") & "k/" &
HowManyKs & "k lines processed."
  DoEvents
  MakingProgress.Repaint
End If

Next i

Sheets(CurrentChrNum).Columns("A:F").AutoFit

'TAKES THE DATA, AND COPIES IT INTO A NEW FILE FOR LATER
PROCESSING
Application.DisplayAlerts = False
ThisWorkbook.Sheets(DataSheet).Delete
Set ExportTo = Workbooks.Add
ThisWorkbook.Sheets(CurrentChrNum).Move Before:=ExportTo.Sheets(1)
```

```
If Dir(ThisWorkbook.Path & "\processed", vbDirectory) = "" Then MkDir
ThisWorkbook.Path & "\processed"
ExportTo.Sheets(2).Delete
Application.DisplayAlerts = True
Application.ScreenUpdating = False
ExportTo.Sheets(1).Columns("A:F").Sort key1:=Range("B1"), _
    order1:=xlAscending
Application.ScreenUpdating = True
ExportTo.SaveAs Filename:=ThisWorkbook.Path & "\processed\" & CurrentChrNum,
FileFormat:=50
ExportTo.Close
MsgBox (CurrentChrNum & " processed, and saved!")
Unload MakingProgress

End Sub

Function HaveNumbers(oRng As Range) As Boolean
'Source: http://stackoverflow.com/questions/18906581/how-can-i-check-if-a-cell-in-
excel-spreadsheet-contains-number
    Dim bHaveNumbers As Boolean, k As Long
    bHaveNumbers = False
    For k = 1 To Len(oRng.Text)
       If IsNumeric(Mid(oRng.Text, k, 1)) Then
          bHaveNumbers = True
          Exit For
       End If
    Next
    HaveNumbers = bHaveNumbers
End Function
```

APPENDIX IX

Step 3 VBA: Annotate Data

```
Sub Step3AnnotateData()
Dim chrNumOfRows, NumOfRows, StartAt, NumOfMatches As Long
Dim Progress As Double
Dim CurrentlyOpen, DataSheet, DataWkbk As String
Dim WhichColor, Previous As Integer

'COLOR SETUP
'Set up some arrays for the colors. These are used later visually separate blocks of
annotated transcripts.
R = Array(255, 181, 154, 193, 134, 251)
G = Array(250, 225, 206, 179, 207, 182)
B = Array(129, 174, 223, 215, 190, 209)
WhichColor = 0

'COUNTS ROWS
'NumOfRows: Length of DataToMatch. Counting the rows will tell the program how
to loop.
NumOfRows = ActiveSheet.Range("A1",
ActiveSheet.Range("A1").End(xlDown)).Rows.Count

'PROGRESSBAR INITIALIZATION
'Shows the ProgressBar2 UserForm.
MakingProgress2.Show vbModeless
Progress = 0

'OTHER PREPARATIONS
'Sorts the chromosome in ascending order to streamline annotating.
Application.ScreenUpdating = False
ThisWorkbook.ActiveSheet.Columns("A:H").Sort key1:=Range("a1"), _
      order1:=xlAscending, key2:=Range("b1"), order2:=xlAscending
Application.ScreenUpdating = True

'SAVE THE NAMES
'We'll be switching back and forth between the data to match, and the chr# files, so
again, just to be safe, everything is stored in strings.
DataSheet = ActiveSheet.Name
DataWkbk = ThisWorkbook.Name

NumOfMatches = 0
i = 1
```

```
'CREATE NEW SHEET
'Creates a new sheet to save the annotations. Default named: "Matches". To avoid errors, make sure there isn't already an existing sheet named "Matches."
Workbooks(DataWkbk).Sheets.Add.Name = "Matches"

'DEFINING VARIABLES
'To define some variables: Workbooks(DataWkbk).Sheets(DataSheet) = this refers to the mouse/human data file that we're looping through.
'In this file: Cells(i, 1) = chr1, chr2, chr3, and so on where "i" refers to the row index and the number refers to the column index.
'Important: the variable CurrentlyOpen will always get this value, because that's the file it needs to open.
'Cells(i, 2) = Testis number | Cells(i, 3) = Liver number | Cells(i, 4-5-6) = other data
'Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen) = is the chr# file we currently have open to go through looking for matches.
'The structure of the other file is basically the same: 1st column: chr#, 2nd: testis data, 3rd: liver data, 4-5-6th: misc data.

'MAIN LOOP STARTS
'It runs until we get to the end of the file (meaning i = the number of rows it needs to go through).
'You'll see this condition at the very end of the code in a row starting with "Loop Until". Watch the #####s!
Do

CurrentlyOpen = Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 1).Value2
'MsgBox (i & "  Opening " & Workbooks(DataWkbk).Path & "\processed\" & CurrentlyOpen & ".xlsb")
Application.ScreenUpdating = False

'OPEN FILE
'This (because of the nested loop) is only called again, if the chr# changes.
'The Dir() command checks if the specified file path exists. If it equals "", then it does NOT.
If Dir(Workbooks(DataWkbk).Path & "\processed\" & CurrentlyOpen & ".xlsb") = "" Then
   Workbooks.Open (Workbooks(DataWkbk).Path & "\processed\" & "chrM.xlsb")
   CurrentlyOpen = "chrM"
   Else: Workbooks.Open (Workbooks(DataWkbk).Path & "\processed\" & CurrentlyOpen & ".xlsb")
End If
Application.ScreenUpdating = True

'Another file length count, but this time it's the chr# file.
```

```
chrNumOfRows = Workbooks(CurrentlyOpen &
".xlsb").Sheets(CurrentlyOpen).Range("A1",
Sheets(CurrentlyOpen).Range("A1").End(xlDown)).Rows.Count

'***** 1ST NESTED LOOP ***** starts
'A new file will only need to be opened if the next row's chr# is different from the
current one.
'Again, you can see this later when this loop ends in a row starting with "Loop Until".
Watch the *****s!

Do

'This code will check the values at every 2% of the file, and start the loop when we're
close enough.

'FIND WHERE TO START LOOKING
For j = 1 To 50

'IF[              current liver data value                ] >= [                        chr# file, 2-4-
6%th row liver data value                                  ]
If Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 2).Value2 >=
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(1 + Round((j - 1)
* chrNumOfRows / 50), 2) Then
   'This might produce a false negative if the matches are in the last 2%, so just in
case:
   If j = 50 Then StartAt = Round(chrNumOfRows * 0.98) - 100
   '-100 in case the 2%-th cell is in the middle of identical matches (variants, for
example)
  'IF [              current liver data value                ] >= [                        chr# file, 4-6-
8%th row liver data value                                 ]
   If Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 2).Value2 <=
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(1 + Round((j) *
chrNumOfRows / 50), 2) Then

      'We only get here if we found the interval of our matches.
      StartAt = Round(chrNumOfRows / 50 * (j - 1)) - 100

      Exit For
   End If
End If
Next j

If StartAt < 1 Then StartAt = 1
' FOUND WHERE TO START LOOKING

Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 9).Value2 = "No Matches"
```

```
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 10).Value2 = StartAt
'Now we know where we should start looking.

j = 0

' MATCH FINDING starts
'This starts looking for matches where it seemed optimal to do so, and will keep looking for them until the supposedly matching liver data have a difference larger than 100bp between them. (Note: if you want to change the bp margin, just change this number.) You'll see this condition, again, lower down when the row says "Loop Until". Look for the &&&&&s!

Do
If Abs(Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 2).Value2 - Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j, 2).Value2) <= 100 Then
   If Abs(Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 3).Value2 - Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j, 3).Value2) <= 100 Then
   'Abs() gives you the absolute value of something. Basically what we're calculating here:
   'If Abs(liver data - chr file liver data) is <= 100 then If Abs(testis data - chr file testis data) is <= 100 then we have a match!
   'We start the search at the StartAt row, we use j to loop until the difference is greater than 100. If either of the Ifs fail, we jump straight to the next j.

      'We have a match!
   NumOfMatches = NumOfMatches + 1
   Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 9).Value2 = "Match!"

   'The coloring just serves as a way to separate different genes from the next.
   If (NumOfMatches > 1 And Previous <> i) Then
      If Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 1).Value2 <> Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches - 1, 1).Value2 Then
      WhichColor = WhichColor + 1
      Workbooks(DataWkbk).Activate
      If WhichColor = 5 Then WhichColor = 0
   End If
   End If
      'We found a match. This takes all the data attached to it and put it in a new row in our Matches sheet.
   Workbooks(DataWkbk).Sheets("Matches").Rows(NumOfMatches).Interior.Color = RGB(R(WhichColor), G(WhichColor), B(WhichColor))
```

```
    Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 1) =
CurrentlyOpen
    Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 2) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 2)
    Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 3) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 3)
    Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 4) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 4)
    Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 5) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 5)
    Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 6) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 6)
    Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 7) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 7)
    Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 8) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 8)
    Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 9) =
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j, 4)
    Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 10) =
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j, 5)
    Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 11) =
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j, 6)

    Previous = i

    End If
End If
j = j + 1
Loop Until (Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 2).Value2 + 100 <
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j,
2).Value2) Or (StartAt + j = chrNumOfRows)
'&&&&& MATCH FINDING &&&&& ends
'We found all the possible matches for the current, i-th row. The difference between
liver data we're supposed to match is larger than 100. There's also another check here.

'Next row!
i = i + 1

If Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 2).Value2 - 100 >
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(chrNumOfRows,
2) Then
Do Until CurrentlyOpen <> Workbooks(DataWkbk).Sheets(DataSheet).Cells(i,
1).Value2
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 9).Value2 = "No Matches"
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 10).Value2 = "-"
i = i + 1
```

```
Loop
End If

'PROGRESSBAR2
Progress = i / NumOfRows
MakingProgress2.ProgressBar.Width = Round(Progress * 200)
MakingProgress2.ProgressText.Caption = Format(Progress * 100, "0.00") & "%"
MakingProgress2.Caption = i & "/" & NumOfRows
MakingProgress2.InfoBox.Caption = NumOfMatches & " matches so far." & "
Currently looking in " & CurrentlyOpen & ". "
DoEvents
MakingProgress2.Repaint

Loop Until CurrentlyOpen <> Workbooks(DataWkbk).Sheets(DataSheet).Cells(i,
1).Value2
'***** 1ST NESTED LOOP ***** ends
'The next row's chr# is different than the current one. We have to open a new file.
'Close the old one first.
Workbooks(CurrentlyOpen & ".xlsb").Close

If i >= NumOfRows Then Exit Do
Loop Until i = NumOfRows
'##### MAIN LOOP ##### ends

'Unload the Progressbar window, and give the columns automatic width.
Unload MakingProgress2
Workbooks(DataWkbk).Sheets("Matches").Columns("A:K").AutoFit

'The End
MsgBox ("Annotation successfully complete.")

End Sub
```

APPENDIX X

Step 3.5 VBA: Annotate Individual Transcripts

```
Sub step3andahalf()
Dim chrNumOfRows, NumOfRows, StartAt, NumOfMatches As Long
Dim Progress As Double
Dim CurrentlyOpen, DataSheet, DataWkbk As String
Dim WhichColor, Previous As Integer

'COLOR SETUP
'Set up some arrays for the colors. These are used later to "block" the matches that
from the same data row.
R = Array(255, 181, 154, 193, 134, 251)
G = Array(250, 225, 206, 179, 207, 182)
B = Array(129, 174, 223, 215, 190, 209)
WhichColor = 0

'COUNTS ROWS
'NumOfRows: length of DataToMatch
NumOfRows = ActiveSheet.Range("A1",
ActiveSheet.Range("A1").End(xlDown)).Rows.Count

'PROGRESSBAR INITIALIZATION
'Show the ProgressBar2 UserForm.
MakingProgress2.Show vbModeless
Progress = 0

'OTHER PREPARATIONS
'Sort data.
Application.ScreenUpdating = False
ThisWorkbook.ActiveSheet.Columns("A:H").Sort key1:=Range("a1"), _
      order1:=xlAscending, key2:=Range("b1"), order2:=xlAscending
Application.ScreenUpdating = True

'SAVE THE NAMES
DataSheet = ActiveSheet.Name
DataWkbk = ThisWorkbook.Name

NumOfMatches = 0
i = 1

'CREATE NEW SHEET
Workbooks(DataWkbk).Sheets.Add.Name = "Matches"

'##### MAIN LOOP ##### starts
```

```
'It runs until we get to the end of the file (meaning i = the number of rows it needs to
go through).
'You'll see this condition at the very end of the code in a row starting with "Loop
Until". Watch the #####s!
Do

CurrentlyOpen = Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 1).Value2
'MsgBox (i & "  Opening " & Workbooks(DataWkbk).Path & "\processed\" &
CurrentlyOpen & ".xlsb")
Application.ScreenUpdating = False

'OPEN FILE
If Dir(Workbooks(DataWkbk).Path & "\processed\" & CurrentlyOpen & ".xlsb") = ""
Then
   Workbooks.Open (Workbooks(DataWkbk).Path & "\processed\" & "chrM.xlsb")
   CurrentlyOpen = "chrM"
   Else: Workbooks.Open (Workbooks(DataWkbk).Path & "\processed\" &
CurrentlyOpen & ".xlsb")
End If
Application.ScreenUpdating = True

'Another file length count, but this time it's the chr# file.
chrNumOfRows = Workbooks(CurrentlyOpen &
".xlsb").Sheets(CurrentlyOpen).Range("A1",
Sheets(CurrentlyOpen).Range("A1").End(xlDown)).Rows.Count

'***** 1ST NESTED LOOP ***** starts
'A new file will only need to be opened if the next row's chr# is different from the
current one.
'Again, you can see this later when this loop ends in a row starting with "Loop Until".
Watch the *****s!

Do

'@@@@@ FIND WHERE TO START LOOKING @@@@@
For j = 1 To 50

'IF[              current liver data value                ] >= [                        chr# file, 2-4-
6%th row liver data value                                  ]
If Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 2).Value2 >=
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(1 + Round((j - 1)
* chrNumOfRows / 50), 2) Then
   If j = 50 Then StartAt = Round(chrNumOfRows * 0.98) - 200
   '-100 in case the 2%-th cell is in the middle of identical matches (variants, for
example)
```

```
  'IF [             current liver data value              ] >= [                        chr# file, 4-6-
8%th row liver data value                          ]
   If Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 2).Value2 <=
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(1 + Round((j) *
chrNumOfRows / 50), 2) Then

      'We only get here if we found the interval our matches are in.
      StartAt = Round(chrNumOfRows / 50 * (j - 1)) - 200

      Exit For
   End If
End If
Next j

If StartAt < 1 Then StartAt = 1
'@@@@@ FOUND WHERE TO START LOOKING @@@@@

'debugging only
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 9).Value2 = "No Matches"
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 10).Value2 = StartAt
'Now we know where we should start looking.

j = 0

'&&&&& MATCH FINDING &&&&& starts

Do
If ((Abs(Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 2).Value2 -
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j,
2).Value2) <= 100) Or (Abs(Workbooks(DataWkbk).Sheets(DataSheet).Cells(i,
3).Value2 - Workbooks(CurrentlyOpen &
".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j, 3).Value2) <= 100)) Then
   'Abs() gives you the absolute value of something. Basically what we're calculating
here:
   'If Abs(liver data - chr file liver data) is<= 100 then If Abs(testis data - chr file testis
data) is <= 100 then we have a match!
   'We start the search at the StartAt row, we use j to loop until the difference is
greater than 100. If either of the Ifs fail, we jump straight to the next j.

   'We have a match!
   NumOfMatches = NumOfMatches + 1
   Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 9).Value2 = "Match!" 'debug
only, this is in the original data file

   If (NumOfMatches > 1 And Previous <> i) Then
```

```
      If Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 1).Value2
<> Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches - 1, 1).Value2
Then
      WhichColor = WhichColor + 1
      Workbooks(DataWkbk).Activate
      If WhichColor = 5 Then WhichColor = 0
   End If
   End If


   'Now we take the match and its associated data and put it in a new row in our
Matches sheet.
   Workbooks(DataWkbk).Sheets("Matches").Rows(NumOfMatches).Interior.Color =
RGB(R(WhichColor), G(WhichColor), B(WhichColor))

   Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 1) =
CurrentlyOpen
   Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 2) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 2)
   Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 3) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 3)
   Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 4) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 4)
   Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 5) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 5)
   Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 6) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 6)
   Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 7) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 7)
   Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 8) =
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 8)
   Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 9) =
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j, 4)
   Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 10) =
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j, 5)
   Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 11) =
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j, 6)

   Previous = i

   If (Abs(Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 2).Value2 -
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j,
2).Value2) <= 100) Then
Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 2).Font.Bold =
True
```

```
    If (Abs(Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 3).Value2 -
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j,
3).Value2) <= 100) Then
Workbooks(DataWkbk).Sheets("Matches").Cells(NumOfMatches, 3).Font.Bold =
True

        End If

j = j + 1
Loop Until ((Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 2).Value2 + 100 <
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j,
2).Value2) And (Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 3).Value2 + 100
< Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(StartAt + j,
3).Value2)) Or (StartAt + j = chrNumOfRows)
'&&&&& MATCH FINDING &&&&& ends

'Next row!
i = i + 1


'This is here to speed things up. It tends to happen that the to-be-matched liver data
goes all the way up to 2 million, while the related chr# file's liver data ends at 1
million.
'In that case it'll loop through the data until it finds the next chr# (from chr11 to chr12,
for example).
If Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 2).Value2 - 100 >
Workbooks(CurrentlyOpen & ".xlsb").Sheets(CurrentlyOpen).Cells(chrNumOfRows,
2) Then
Do Until CurrentlyOpen <> Workbooks(DataWkbk).Sheets(DataSheet).Cells(i,
1).Value2
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 9).Value2 = "No Matches"
Workbooks(DataWkbk).Sheets(DataSheet).Cells(i, 10).Value2 = "-"
i = i + 1
Loop
End If



'PROGRESSBAR2
Progress = i / NumOfRows
MakingProgress2.ProgressBar.Width = Round(Progress * 200)
MakingProgress2.ProgressText.Caption = Format(Progress * 100, "0.00") & "%"
MakingProgress2.Caption = i & "/" & NumOfRows
MakingProgress2.InfoBox.Caption = NumOfMatches & " matches so far." & "
Currently looking in " & CurrentlyOpen & ". "
DoEvents
```

```
MakingProgress2.Repaint

Loop Until CurrentlyOpen <> Workbooks(DataWkbk).Sheets(DataSheet).Cells(i,
1).Value2
'***** 1ST NESTED LOOP ***** ends
'The next row's chr# is different than the current one. We have to open a new file.

'Close the old file first.
Workbooks(CurrentlyOpen & ".xlsb").Close

'This is when we basically go all the way back to the beginning of the MAIN LOOP,
open a new file, and start looking for the matches there. Unless, of course,
'we're already out of data to match. In that case:

If i >= NumOfRows Then Exit Do
Loop Until i = NumOfRows
'##### MAIN LOOP ##### ends

Unload MakingProgress2
Workbooks(DataWkbk).Sheets("Matches").Columns("A:K").AutoFit

'The End
MsgBox ("Annotation successfully complete.")

End Sub
```

FIGURES

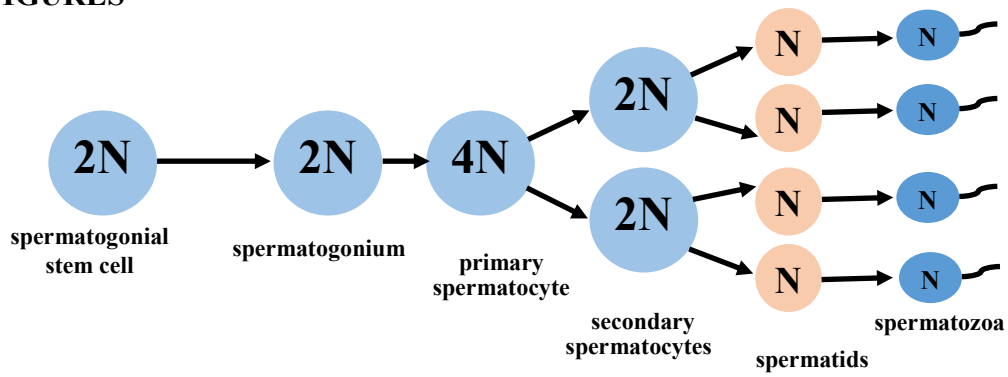

Figure 1. Spermatogenesis

Spermatogonial stem cells undergo a series of mitotic and meiotic divisions to form differentiated and mature spermatozoa during spermatogenesis.

"N" = chromatid quantity

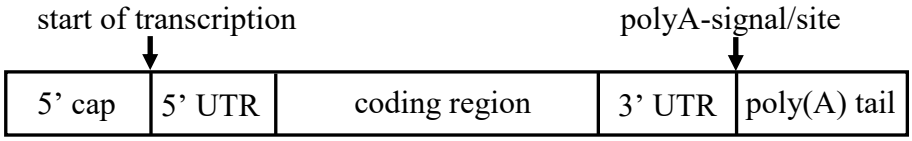

Figure 2. Eukaryotic messenger RNA

Mature messenger RNA (mRNA) structurally consists of a 5'cap, an non-coding untranslated region (UTR) flanking both 3' and 5'-ends, poly(A)-signal marking for cleavage at downstream polyA-site, and a string of adenine residues attached to the post-transcriptionally modified transcript following polyadenylation.

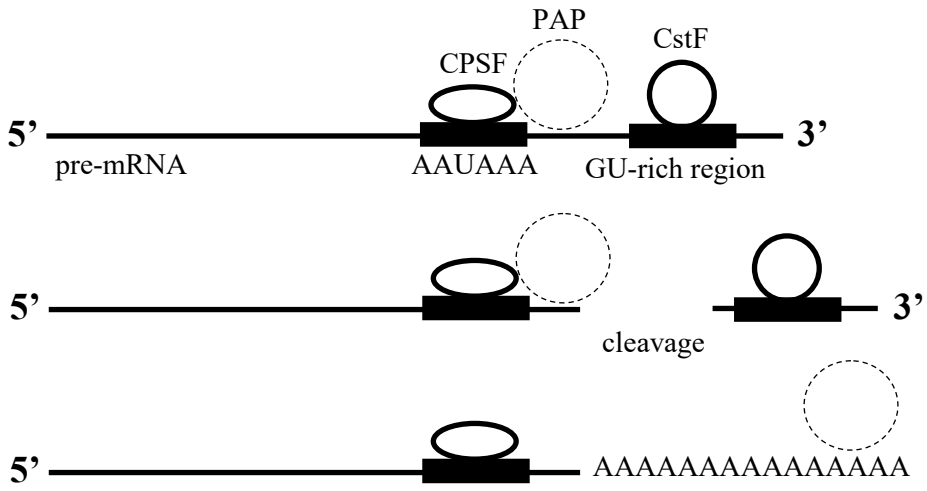

Figure 3. Polyadenylation

Polyadenylation in eukaryotic cells occurs before transcription is completed in the nascent mRNA. It involves the binding of Cleavage and Polyadenylation Specificity Factor (CPSF) to a polyadenylation signal (PAS) and recruitment of Poly(A) Polymerase. Downstream at a GU-rich region, Cleavage Stimulation Factor (CstF) is recruited by CPSF and cleaves the transcript at the polyA-site. PAP initiates the addition of adenine residues to make up the poly(A)-tail.

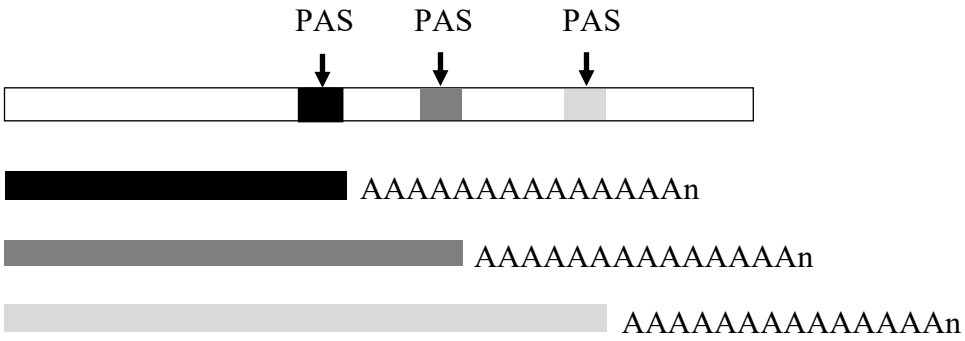

Figure 4. Alternative polyadenylation in the 3'UTR creates transcript isoforms

The white bar represents an mRNA 3'UTR. Usage of variant PAS and cleavage at different polyA-sites involved with alternative polyadenylation results in truncation of the 3'UTR and transcript isoforms that may have altered downstream metabolism due to escape from regulatory sequence elements.

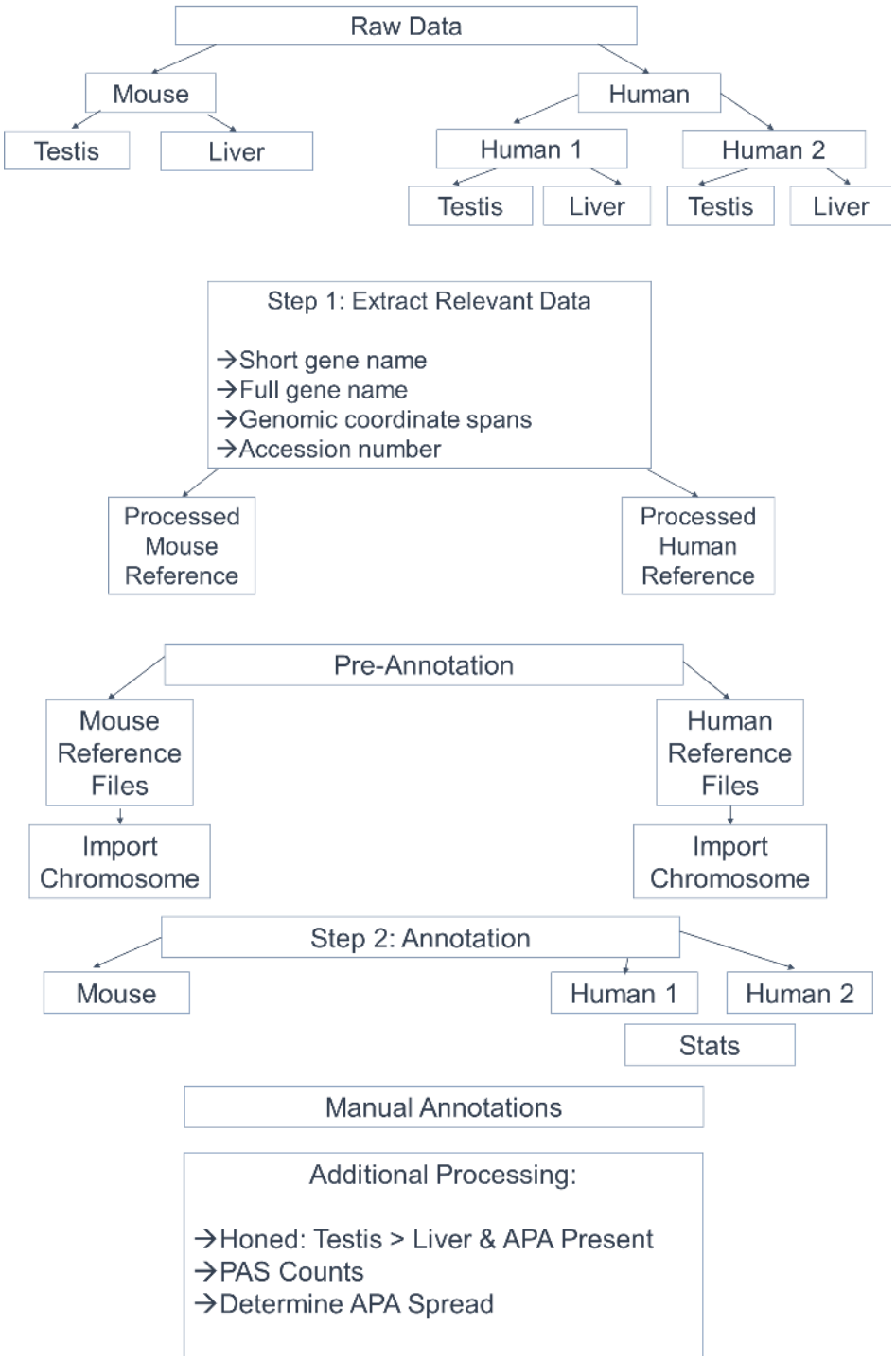

Figure 5. Bioinformatics analysis workflow

Data processing and workflow for PolyA-seq analysis and transcript annotations.

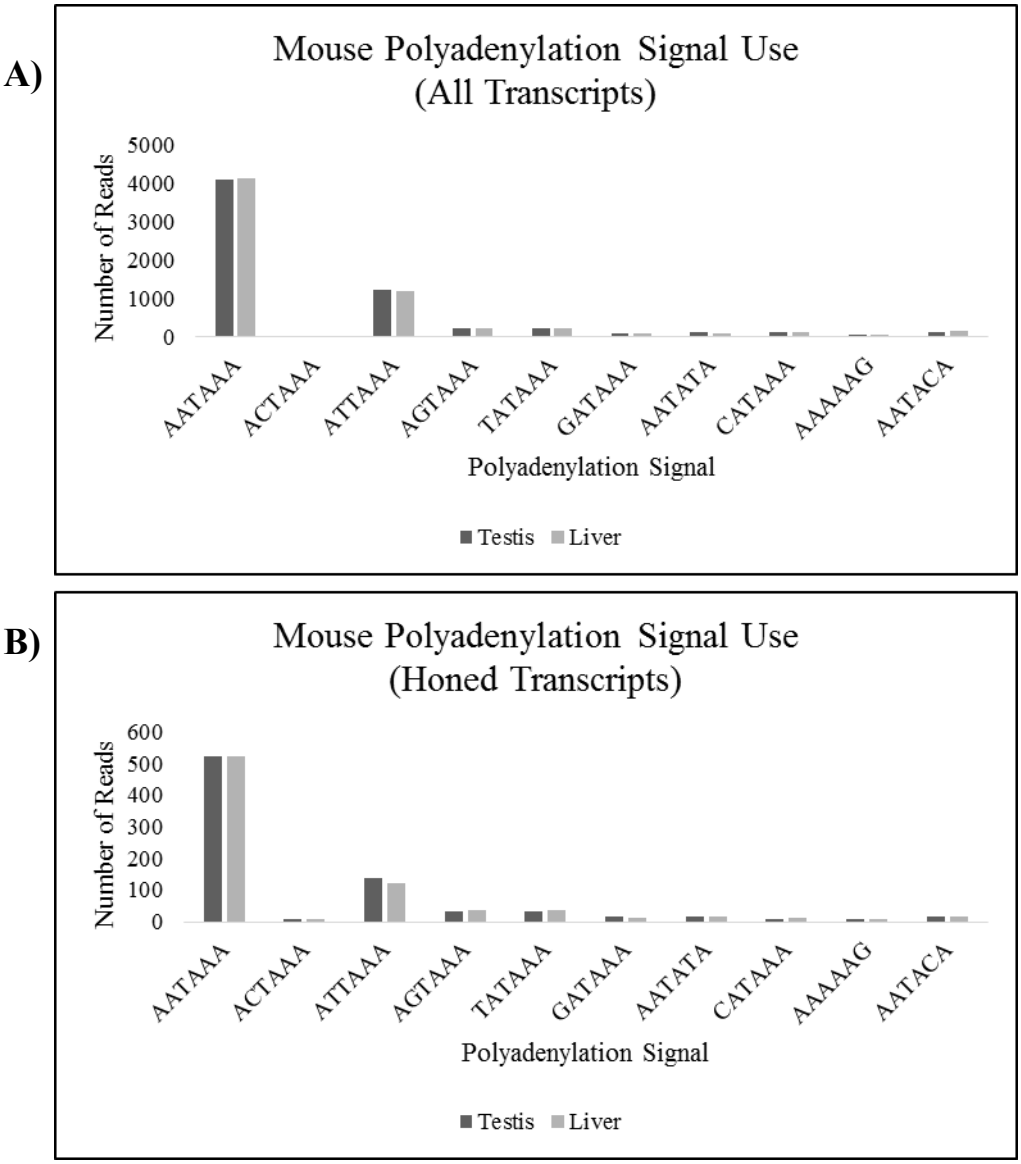

Figure 6. Graphical representation of PAS use in mouse

A) Depicts PAS use in all comprehensive transcripts in testis (dark gray bars) and liver (light gray bars) at the same polyA-site within a 100bp margin.

B) Depicts PAS use in "honed" transcripts, where testis reads are greater than liver by at least one and PAS differ between matched transcripts at the same polyA-site within a 100bp margin, for testis and liver.

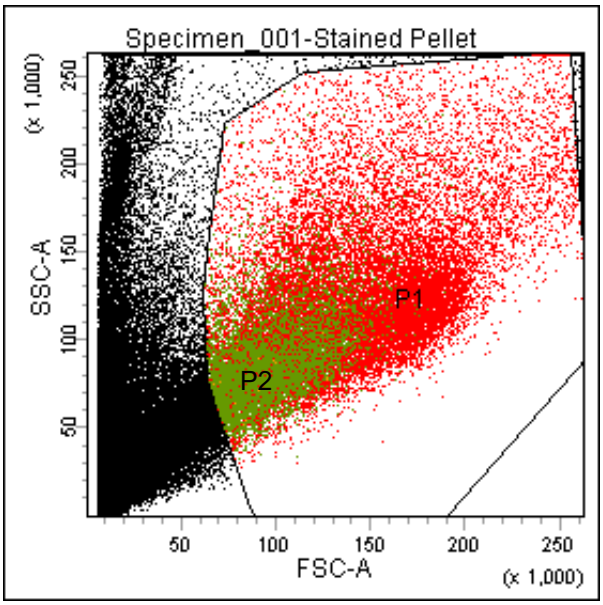

Figure 7. Size distribution of male germ cells and gating for live cells

Live cells were gated into populations based on size. The P1 population was identified as the live cell population after backgating with PI stain.

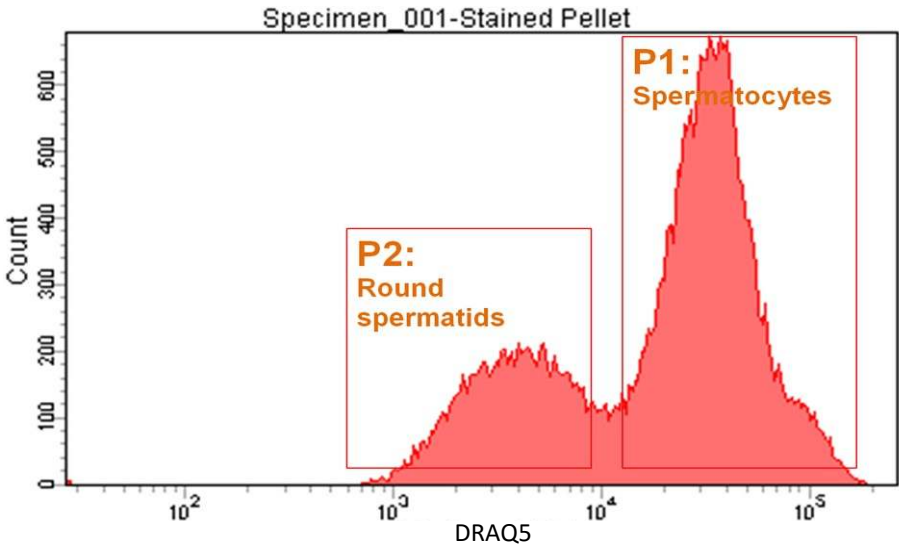

Figure 8. Two distinct cell populations were identified with DRAQ5 nuclear stain

FACS histogram depicting two distinct populations based on fluorescence by DRAQ5 emission. spermatocyte and round spermatid populations are labeled Based on cell-specific PCR markers.

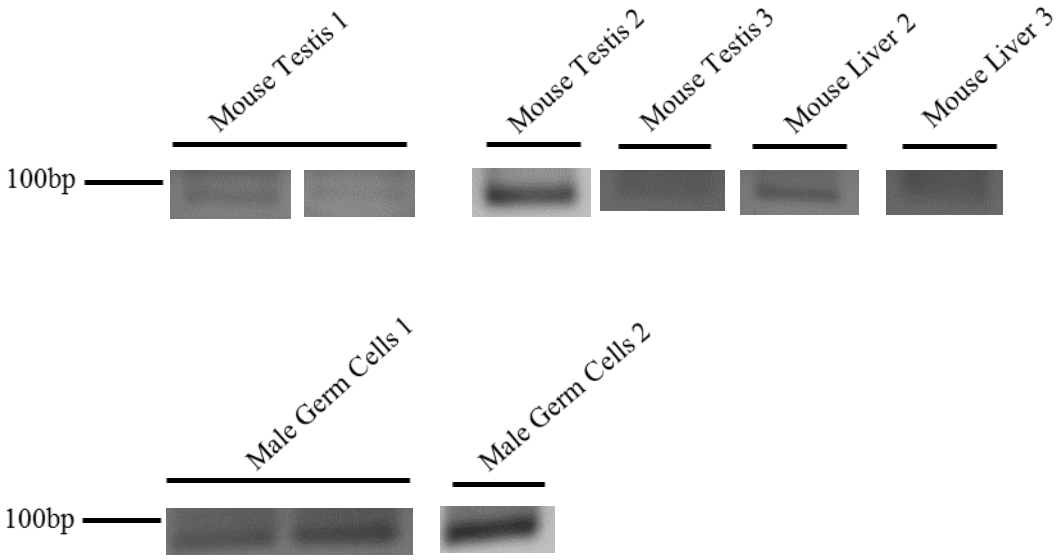

Figure 9. PolyA-seq libraries from mouse

PolyA-seq libraries from mouse testicular tissue, liver, and pooled male germ cells (round spermatids and spermatocytes). Male Germ Cells 1 and Male Germ Cells 2 indicate prepared libraries from two separate sorts and biological replicates.

TABLES

Gene	Somatic Isoforms (kB)	Testis Isoforms (kB)	Validation	Reference
ADP ribosylation factor (ARF)	3.0	1.8, 1.3, 1.1	In situ hybridization, RACE-PCR	Mishima et al., 1992
Angiotensin converting enzyme (ACE)	4.5	4.5	Northern blot	Thekkumkara et al., 1992
Basic leucine zipper and W2 domains 1 (BZW1)	2.9	1.8	Northern blot, 5'- and 3'RACE, Western blot	Yu et al., 2006
Cleavage polyadenylation specificity factor 6 (CPSF6)	4.0	2.1	Northern blot, 3'RACE	Sartini et al., 2007
Cleavage and polyadenylation specificity factor subunit 5 (NUDT21)	4.5	1.1	Northern blot, 3'RACE	Sartini et al., 2007
Deleted in azoospermia associated protein 1 (DAZAP1)	2.4	1.8	Northern blot, 3'RACE	Yang & Yen, 2013
DNA polymerase IV (DinB)	5	4.4, 3.2	Northern blot, RT-PCR	Gerlach et al., 1999
Excision repair gene (ERCC6)	7	5, 3.5	Northern blot	Troelstra et al., 1993
Kruppel-like factor 4 (KLF4)	0.67	1.1, 1.0, 0.66, 0.38	Northern blot, RT-PCR	Goodman et al., 2005
RanGAP1	3.5	3.5, 2.8	Northern blot	Krebber & Ponstingl, 1996
Ring finger protein 4 (RNF4/SNURF)	3.0	3.0, 1.6	Northern blot	Yan et al., 2002
Sp1 transcription factor (Sp1)	8.8	4.1, 3.7, 3.2, 1.4	Northern blot	Thomas et al., 2005
Translation initiation factor 2α (eIF-2α)	4.2	1.6, 1.7	Northern blot	Miyamoto et al., 1996
Triose phophate isomerase (TPI)	2.6	1.5, 1.4	Northern blot	Russell & Kim, 1996

Table 1. Testis-specific genes with alternative poly(A) signals in the 3'-UTR

Literature-mined testis-specific transcripts with APA-mediated 3'UTR isoforms as validated by molecular biology experiments.

	Total Number of	Number of Transcripts	Percent of Transcripts Annotated
Human 1	19906	143	0.718%
Human 2	10179	86	0.844%

	Number of Transcripts Honed	Percent of Transcripts Honed
Human 1	1082	5.435%
Human 2	546	5.363%

Table 2. Human transcript numerical data

Numerical data for PolyA-seq data following annotation and creation of "honed" subset from Human 1 and Human 2 data files.

	Testis Average	Testis Std Dev	Liver Average	Liver Std Dev	P-Value
AATAAA	9141	4140.817	9192.5	4232.034	0.991
ACTAAA	143.5	67.175	138.5	60.104	0.944
ATTAAA	2970.5	1354.109	2926.5	1318.754	0.976
AGTAAA	679	298.399	659.5	283.549	0.952
TATAAA	679.5	316.076	681.5	297.691	0.995
GATAAA	276	128.693	262.5	127.986	0.925
AATATA	298	144.249	311	135.764	0.934
CATAAA	353.5	170.412	360.5	174.655	0.971
AAAAAG	159	79.195	159.5	88.388	0.995
AATACA	344	181.019	352	161.220	0.967

Table 3. Human PAS statistics

Counts for PAS usage in both human PolyA-seq data files for testis and liver were compared for differential usage. Standard deviation and P-values calculated using Student's unpaired t-test indicate that there is no significance.

		Liver PAS									
		AAAAAG	**AATAAA**	**AATACA**	**AATATA**	**ACTAAA**	**AGTAAA**	**ATTAAA**	**CATAAA**	**GATAAA**	**TATAAA**
Testis PAS	**AAAAAG**		7	1	0	0	0	3	0	1	5
	AATAAA	15		34	27	12	53	186	32	14	81
	AATACA	2	18		2	2	1	3	4	2	5
	AATATA	0	13	2		2	2	13	1	2	5
	ACTAAA	5	9	1	0		1	4	1	1	0
	AGTAAA	2	35	3	2	2		17	7	5	10
	ATTAAA	8	130	11	28	10	22		16	13	37
	CATAAA	2	14	4	3	3	5	7		5	3
	GATAAA	2	12	1	1	2	2	8	4		1
	TATAAA	6	26	3	7	1	7	14	6	3	

Table 4. Human 1 Honed PAS spread

		Liver PAS									
		AAAAAG	**AATAAA**	**AATACA**	**AATATA**	**ACTAAA**	**AGTAAA**	**ATTAAA**	**CATAAA**	**GATAAA**	**TATAAA**
Testis PAS	**AAAAAG**		6	1	0	0	1	4	0	0	3
	AATAAA	7		34	18	5	26	95	21	5	39
	AATACA	0	5		0	1	2	2	2	2	1
	AATATA	0	5	0		0	0	6	0	0	2
	ACTAAA	2	3	1	0		0	0	2	0	1
	AGTAAA	1	19	3	0	4		7	1	3	7
	ATTAAA	4	56	2	13	4	8		10	5	19
	CATAAA	2	9	2	2	1	1	4		1	1
	GATAAA	0	5	1	0	0	3	3	1		2
	TATAAA	2	11	3	4	2	2	10	5	1	

Table 5. Human 2 Honed PAS spread

	Total Number of Transcripts	Number of Transcripts Annotated	Percent of Transcripts Annotated	Number of Transcripts Honed	Percent of Transcripts Honed
Mouse	6341	34	0.536%	251	3.958%

Table 6. Mouse comprehensive numerical data

Testis PAS	Liver PAS: AAAAAG	AATAAA	AATACA	AATATA	ACTAAA	AGTAAA	ATTAAA	CATAAA	GATAAA	TATAAA
AAAAAG		9	0	1	1	2	2	0	0	1
AATAAA	5		17	11	5	22	65	13	5	26
AATACA	0	11		1	0	2	3	0	3	0
AATATA	1	12	1		0	1	8	2	2	8
ACTAAA	0	6	0	0		1	1	0	0	2
AGTAAA	1	27	0	1	0		9	3	2	3
ATTAAA	5	92	6	8	3	6		6	5	8
CATAAA	0	11	2	0	1	0	7		0	3
GATAAA	0	7	1	0	0	3	3	0		0
TATAAA	0	27	4	2	0	1	7	2	0	

Table 7. Mouse PAS spread for all chromosomes

Testis PAS	Liver PAS: AAAAAG	AATAAA	AATACA	AATATA	ACTAAA	AGTAAA	ATTAAA	CATAAA	GATAAA	TATAAA
AAAAAG		3	0	0	0	0	0	0	0	1
AATAAA	0		12	8	4	13	39	11	1	14
AATACA	0	3		1	0	1	0	0	3	0
AATATA	1	3	1		0	0	1	1	2	5
ACTAAA	0	2	0	0		1	0	0	0	0
AGTAAA	1	12	0	0	0		3	2	1	2
ATTAAA	4	34	2	8	2	4		5	2	6
CATAAA	0	4	0	0	1	0	2		0	1
GATAAA	0	3	0	0	0	2	2	0		0
TATAAA	0	12	0	2	0	0	3	0	0	

Table 8. Honed mouse PAS spread

Gene	Chr	DNA Coordinates (2007)		PolyA Coordinates		PAS	Liver Reads	Testis Reads	3'UTR Isoform (kb)	Published Liver Isoforms (kb)	Published Testis Isoforms (kb)
ACE	11	105829261	105851259								
				105850572	105850573	ATTAAA	50	337	2.5	5	2.5
ARF	14	27457683	27476744	NOT FOUND						1.8	1.8, 1.1
BZW1	1	58449980	58463392							2.7	2.7, 1.6
				58462333	58462334	ATTAAA	7	2	1.6		
				58463391	58463392	AATAAA	32	8	2.7		
CPSF6	10	116814029	116781724								
				116785906	116785907	ATTAAA	4	5		4	4.0, 2.1, 1.5
				116786094	116786095	AATAAA	0	8	4		
				116786364	116786365	AATAAA	7	31	1.5		
				116784274	116784275	AGTAAA	4	0			
DAZAP1	10	79727736	79751158								
				79750545	79750546	AATAAA	0	9		2.4	2.4, 1.8
				79750642	79750643	AATAAA	13	61	1.8		
				79750801	79750802	AATATA	2	0			
				79750816	79750817	AATAAA	0	8	?		
DinB/POLK	13	97312440	97250644								
				97250644	97250645	AATAAA	0	6	?		5, 4.4, 3.2
eIF-2a	12	79963059	79987997	NOT FOUND							

Table 9. Literature-mined transcripts in PolyA-seq data

Transcripts from Table 1 manually identified in PolyA-seq mouse data.

Gene	Chr	DNA Coordinates (2007)		PolyA Coordinates	PAS	Liver Reads	Testis Reads	3'UTR Isoform (kb)	Published Liver Isoforms (kb)	Published Testis Isoforms (kb)
ERCC6	14	33326707	33394175	NOT FOUND						
Klf4	4	55540009	55545347	NOT FOUND						
NUDT21	8	96543303	96560939	NOT FOUND						
RanGAP1	15	81560349	81534678							
				81534683	81534684	AATAAA	188	527	3.5	3.5, 2.8	3.5, 2.8
				81535319	81535320	ATTAAA	0	14			
RNF4	5	34679039	34696079							
				34694135	34694136	AAAAAG	4	15	3	3.0, 1.6	3.0, 1.6
				34696082	34696083	AATAAA	97	31			
Sp1	15	102236747	102266835							
				102263249	102263250	AATAAA	0	5			4.1, 3.7, 3.2, 1.4
				102263376	102263377	AATAAA	40	36			
				102266820	102266821	ATTAAA	10	6			
TPI	6	124764314	124760610								2.6, 1.5, 1.4
				124760733	124760734	AATAAA	482	348			

Table 9 (continued)

Chr	Testis PolyA-site PAS	Liver PAS	Testis Reads	Liver Reads	Accession Umber	Short Gene Name	Long Gene Name
chr7	20263443 AATAAA	AATAAA	4	4990	XM_011250945.1	LOC101055630	uncharacterized LOC101055630, transcript variant X7
chr19	3206272 AATAAA	AATAAA	5	1	XM_011248643.1	1700123I01Rik	RIKEN cDNA 1700123I01 gene, transcript variant X1
chr19	3206272 AATAAA	AATAAA	5	1	XM_006531820.2	1700123I01Rik	RIKEN cDNA 1700123I01 gene, transcript variant X4
chr19	3206272 AATAAA	AATAAA	5	1	XM_011248645.1	1700123I01Rik	RIKEN cDNA 1700123I01 gene, transcript variant X3
chr19	3206272 AATAAA	AATAAA	5	1	NM_001165919.1	1700123I01Rik	RIKEN cDNA 1700123I01 gene
chr19	3206272 AATAAA	AATAAA	5	1	XM_011248644.1	1700123I01Rik	RIKEN cDNA 1700123I01 gene, transcript variant X2
chr12	29367036 AATAAA	AATAAA	354	984	NM_019785.2	Actr10	ARP10 actin-related protein 10
chr1	36595984 AATAAA	AATAAA	6	12	NM_175200.4	Als2cr11	amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 11 (human)
chr1	36595984 AATAAA	AATAAA	6	12	XM_011238605.1	Als2cr11	amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 11 (human), transcript variant X1
chr4	40217253 AGTAAA	AGTAAA	202	329	NM_177195.3	Atp8b5	ATPase, class I, type 8B, member 5
chr1	34502591 AGTAAA	AGTAAA	42	9	XM_011247976.1	Cfap221	cilia and flagella associated protein 221, transcript variant X1
chr1	34502591 AGTAAA	AGTAAA	42	9	XM_011247979.1	Cfap221	cilia and flagella associated protein 221, transcript variant X4
chr1	34502591 AGTAAA	AGTAAA	42	9	XM_011247978.1	Cfap221	cilia and flagella associated protein 221, transcript variant X3
chr1	34502591 AGTAAA	AGTAAA	42	9	XM_011247977.1	Cfap221	cilia and flagella associated protein 221, transcript variant X2
chr1	34502591 AGTAAA	AGTAAA	42	9	XM_011247980.1	Cfap221	cilia and flagella associated protein 221, transcript variant X5
chr1	34502591 AGTAAA	AGTAAA	42	9	XM_006529404.2	Cfap221	cilia and flagella associated protein 221, transcript variant X7
chr1	34502591 AGTAAA	AGTAAA	42	9	NM_001115074.1	Cfap221	cilia and flagella associated protein 221
chr1	34502591 AGTAAA	AGTAAA	42	9	XR_878321.1	Cfap221	cilia and flagella associated protein 221, transcript variant X6
chr1	34502591 AGTAAA	AGTAAA	42	9	XM_011247981.1	Cfap221	cilia and flagella associated protein 221, transcript variant X8
chr17	29840303 ATTAAA	ATTAAA	126	34	NM_008585.2	Mep1a	meprin 1 alpha

Table 10. Mouse comprehensive transcript annotation

Table 10 (continued)

Chr	Testis PolyA-site PAS	Liver PAS	Testis Reads	Liver Reads	Accession Umber	Short Gene Name	Long Gene Name
chr11	69477715 AATAAA	AATAAA	302	46	XM_006532123.2	Crhr1	corticotropin releasing hormone receptor 1, transcript variant X1
chr11	69477715 AATAAA	AATAAA	302	46	NM_007762.4	Crhr1	corticotropin releasing hormone receptor 1
chr4	11259081 AATAAA	AATAAA	45	4	NM_172338.2	Dnajc16	DnaJ (Hsp40) homolog, subfamily C, member 16
chr2	3244353 AATAAA	AATAAA	6	6	NM_010074.3	Dpp4	dipeptidylpeptidase 4, transcript variant 1
chr2	3244353 AATAAA	AATAAA	6	6	NM_001159543.1	Dpp4	dipeptidylpeptidase 4, transcript variant 2
chr2	3244353 AATAAA	AATAAA	6	6	XM_006498691.2	Dpp4	dipeptidylpeptidase 4, transcript variant X2
chr2	3244353 AATAAA	AATAAA	6	6	XM_006498692.2	Dpp4	dipeptidylpeptidase 4, transcript variant X3
chr2	3244353 AATAAA	AATAAA	6	6	XM_011239274.1	Dpp4	dipeptidylpeptidase 4, transcript variant X1
chr2	3244353 AATAAA	AATAAA	6	6	XM_006498693.2	Dpp4	dipeptidylpeptidase 4, transcript variant X4
chr11	70341354 AATAAA	AATAAA	7	8	XM_011249261.1	Efcab3	EF-hand calcium binding domain 3, transcript variant X1
chr11	70268648 CATAAA	CATAAA	5	7	XM_011249261.1	Efcab3	EF-hand calcium binding domain 3, transcript variant X1
chr7	28825669 ATTAAA	CATAAA	25	4	NM_001040131.2	Eif4g2	eukaryotic translation initiation factor 4, gamma 2, transcript variant 2
chr7	28825669 ATTAAA	CATAAA	25	4	NM_013507.3	Eif4g2	eukaryotic translation initiation factor 4, gamma 2, transcript variant 1
chr5	31462146 AATAAA	AATAAA	11	11	NM_001161738.1	Glmn	glomulin, FKBP associated protein, transcript variant 2
chr5	31462146 AATAAA	AATAAA	11	11	NM_001161739.1	Glmn	glomulin, FKBP associated protein, transcript variant 3
chr5	31462146 AATAAA	AATAAA	11	11	NM_133248.2	Glmn	glomulin, FKBP associated protein, transcript variant 1
chr5	31462146 AATAAA	AATAAA	11	11	XM_011249410.1	Glmn	glomulin, FKBP associated protein, transcript variant X2
chr5	31462146 AATAAA	AATAAA	11	11	XM_006534800.2	Glmn	glomulin, FKBP associated protein, transcript variant X1
chr10	62021220 AATAAA	AATAAA	12	3	NM_025431.2	Llph	LLP homolog, long-term synaptic facilitation (Aplysia)
chr10	62021286 ACTAAA	AATAAA	6	3	NM_025431.2	Llph	LLP homolog, long-term synaptic facilitation (Aplysia)

Table 10 (continued)

Chr	Testis PolyA-site PAS	Liver PAS	Testis Reads	Liver Reads	Accession Umber	Short Gene Name	Long Gene Name
chr18	70785780 AATAAA	AATAAA	78	43	XM_006525496.2	Me2	malic enzyme 2, NAD(+)-dependent, mitochondrial, transcript variant X1
chr18	70785780 AATAAA	AATAAA	78	43	NM_145494.2	Me2	malic enzyme 2, NAD(+)-dependent, mitochondrial
chr17	29840278 AGTAAA	ATTAAA	3	34	XM_006523751.2	Mep1a	meprin 1 alpha, transcript variant X1
chr17	29840278 AGTAAA	ATTAAA	3	34	NM_008585.2	Mep1a	meprin 1 alpha
chr17	29840303 ATTAAA	ATTAAA	126	34	XM_006523751.2	Mep1a	meprin 1 alpha, transcript variant X1
chr17	29840303 ATTAAA	ATTAAA	126	34	NM_008585.2	Mep1a	meprin 1 alpha
chr9	21560546 ATTAAA	ATTAAA	229	49	XR_379407.2	Myo5c	myosin VC, transcript variant X5
chr9	21560546 ATTAAA	ATTAAA	229	49	XM_006510937.2	Myo5c	myosin VC, transcript variant X4
chr9	21560546 ATTAAA	ATTAAA	229	49	XM_011242693.1	Myo5c	myosin VC, transcript variant X1
chr9	21560546 ATTAAA	ATTAAA	229	49	XM_006510935.2	Myo5c	myosin VC, transcript variant X2
chr9	21560546 ATTAAA	ATTAAA	229	49	XM_006510936.2	Myo5c	myosin VC, transcript variant X3
chr9	21560546 ATTAAA	ATTAAA	229	49	NM_001081322.1	Myo5c	myosin VC
chr9	21560546 ATTAAA	ATTAAA	229	49	XM_011242694.1	Myo5c	myosin VC, transcript variant X8
chr9	45738040 TATAAA	TATAAA	9	1	NM_001165957.1	Nme9	NME/NM23 family member 9
chr11	44268073 AATAAA	AATAAA	3	4	NM_010927.3	Nos2	nitric oxide synthase 2, inducible
chr11	44268073 AATAAA	AATAAA	3	4	XM_006532446.2	Nos2	nitric oxide synthase 2, inducible, transcript variant X1
chr8	4259731 AATAAA	AATAAA	28	40	NM_001083906.1	Nr3c2	nuclear receptor subfamily 3, group C, member 2

Table 10 (continued)

Chr	Testis PolyA-site PAS	Liver PAS	Testis Reads	Liver Reads	Accession Umber	Short Gene Name	Long Gene Name
chr9	37455839 AATATA	AATATA	2	14	XM_006510643.2	Ppp2r1b	protein phosphatase 2, regulatory subunit A, beta, transcript variant X4
chr9	37455839 AATATA	AATATA	2	14	XM_011242614.1	Ppp2r1b	protein phosphatase 2, regulatory subunit A, beta, transcript variant X1
chr9	37455839 AATATA	AATATA	2	14	XM_006510641.2	Ppp2r1b	protein phosphatase 2, regulatory subunit A, beta, transcript variant X2
chr9	37455839 AATATA	AATATA	2	14	XM_006510642.2	Ppp2r1b	protein phosphatase 2, regulatory subunit A, beta, transcript variant X3
chr9	37455839 AATATA	AATATA	2	14	NM_001034085.2	Ppp2r1b	protein phosphatase 2, regulatory subunit A, beta, transcript variant 1
chr9	37455839 AATATA	AATATA	2	14	NM_001286553.1	Ppp2r1b	protein phosphatase 2, regulatory subunit A, beta, transcript variant 3
chr9	37455839 AATATA	AATATA	2	14	NM_028614.3	Ppp2r1b	protein phosphatase 2, regulatory subunit A, beta, transcript variant 2
chr7	16611681 ATTAAA	ATTAAA	4	6	XM_011250461.1	Relb	avian reticuloendotheliosis viral (v-rel) oncogene related B, transcript variant X4
chr7	16611681 ATTAAA	ATTAAA	4	6	XM_011250459.1	Relb	avian reticuloendotheliosis viral (v-rel) oncogene related B, transcript variant X2
chr7	16611681 ATTAAA	ATTAAA	4	6	XM_006539675.1	Relb	avian reticuloendotheliosis viral (v-rel) oncogene related B, transcript variant X1
chr7	16611681 ATTAAA	ATTAAA	4	6	XM_011250460.1	Relb	avian reticuloendotheliosis viral (v-rel) oncogene related B, transcript variant X3
chr7	16611681 ATTAAA	ATTAAA	4	6	XM_011250462.1	Relb	avian reticuloendotheliosis viral (v-rel) oncogene related B, transcript variant X5
chr7	16611681 ATTAAA	ATTAAA	4	6	NM_009046.2	Relb	avian reticuloendotheliosis viral (v-rel) oncogene related B, transcript variant 1
chr7	16611681 ATTAAA	ATTAAA	4	6	NM_001290457.1	Relb	avian reticuloendotheliosis viral (v-rel) oncogene related B, transcript variant 2
chr9	7933613 TATAAA	TATAAA	3	5	XM_006510575.2	Slc44a2	solute carrier family 44, member 2, transcript variant X1
chr9	7933613 TATAAA	TATAAA	3	5	NM_001199186.1	Slc44a2	solute carrier family 44, member 2, transcript variant 1
chr9	7933613 TATAAA	TATAAA	3	5	XM_011242593.1	Slc44a2	solute carrier family 44, member 2, transcript variant X2
chr9	7933613 TATAAA	TATAAA	3	5	NM_152808.3	Slc44a2	solute carrier family 44, member 2, transcript variant 2
chr9	7933613 TATAAA	TATAAA	3	5	XM_006510576.1	Slc44a2	solute carrier family 44, member 2, transcript variant X3

Table 10 (continued)

Chr	Testis PolyA-site PAS	Liver PAS	Testis Reads	Liver Reads	Accession Umber	Short Gene Name	Long Gene Name
chr5	8069175 AATAAA	AATAAA	5	12	NM_009274.2	Srpk2	serine/arginine-rich protein specific kinase 2
chr5	8069175 AATAAA	AATAAA	5	12	XM_011249757.1	Srpk2	serine/arginine-rich protein specific kinase 2, transcript variant X8
chr5	8069175 AATAAA	AATAAA	5	12	XM_011249752.1	Srpk2	serine/arginine-rich protein specific kinase 2, transcript variant X2
chr5	8069175 AATAAA	AATAAA	5	12	XM_011249754.1	Srpk2	serine/arginine-rich protein specific kinase 2, transcript variant X5
chr5	8069175 AATAAA	AATAAA	5	12	XM_011249751.1	Srpk2	serine/arginine-rich protein specific kinase 2, transcript variant X1
chr5	8069175 AATAAA	AATAAA	5	12	XM_011249758.1	Srpk2	serine/arginine-rich protein specific kinase 2, transcript variant X10
chr5	8069175 AATAAA	AATAAA	5	12	XM_006535660.2	Srpk2	serine/arginine-rich protein specific kinase 2, transcript variant X4
chr5	8069175 AATAAA	AATAAA	5	12	XM_011249756.1	Srpk2	serine/arginine-rich protein specific kinase 2, transcript variant X7
chr5	8069175 AATAAA	AATAAA	5	12	XM_011249753.1	Srpk2	serine/arginine-rich protein specific kinase 2, transcript variant X3
chr5	8069175 AATAAA	AATAAA	5	12	XM_011249755.1	Srpk2	serine/arginine-rich protein specific kinase 2, transcript variant X6
chr5	8069175 AATAAA	AATAAA	5	12	XM_006535661.2	Srpk2	serine/arginine-rich protein specific kinase 2, transcript variant X9
chr5	8069175 AATAAA	AATAAA	5	12	XM_011249759.1	Srpk2	serine/arginine-rich protein specific kinase 2, transcript variant X11
chr2	30200330 AATAAA	AATAAA	501	106	NM_019667.2	Stam2	signal transducing adaptor molecule (SH3 domain and ITAM motif) 2
chr4	43579931 AATAAA	AATAAA	45	34	XM_006538046.2	Tbc1d2	TBC1 domain family, member 2, transcript variant X1
chr4	43579931 AATAAA	AATAAA	45	34	NM_198664.3	Tbc1d2	TBC1 domain family, member 2
chr4	43579931 AATAAA	AATAAA	45	34	XM_006538048.1	Tbc1d2	TBC1 domain family, member 2, transcript variant X3
chr4	43579931 AATAAA	AATAAA	45	34	XM_006538047.2	Tbc1d2	TBC1 domain family, member 2, transcript variant X2
chr11	62418307 AATAAA	AATACA	676	18	NM_019507.2	Tbx21	T-box 21
chr11	62418387 AATACA	AATACA	87	18	NM_019507.2	Tbx21	T-box 21

Table 10 (continued)

Chr	Testis PolyA-site PAS	Liver PAS	Testis Reads	Liver Reads	Accession Umber	Short Gene Name	Long Gene Name
chr12	60128929 AATATA	AAAAAG	8	1	XM_006516305.2	Tc2n	tandem C2 domains, nuclear, transcript variant X6
chr2	69709896 ATTAAA	ATTAAA	4	3	XM_006500243.1	Tmem87b	transmembrane protein 87B, transcript variant X1
chr2	69709896 ATTAAA	ATTAAA	4	3	XM_011239791.1	Tmem87b	transmembrane protein 87B, transcript variant X3
chr2	69709896 ATTAAA	ATTAAA	4	3	XM_011239790.1	Tmem87b	transmembrane protein 87B, transcript variant X2
chr2	69709896 ATTAAA	ATTAAA	4	3	NM_028248.2	Tmem87b	transmembrane protein 87B
chr4	43040893 AATAAA	AATAAA	14	20	XM_006537772.2	Tmod1	tropomodulin 1, transcript variant X1
chr4	43040893 AATAAA	AATAAA	14	20	NM_021883.2	Tmod1	tropomodulin 1
chr4	43040893 AATAAA	AATAAA	14	20	XM_006537773.1	Tmod1	tropomodulin 1, transcript variant X2
chr16	13776825 AATAAA	AATAAA	3	2	NM_178855.4	Tmprss15	transmembrane protease, serine 15, transcript variant 2
chr16	13776825 AATAAA	AATAAA	3	2	NM_008941.3	Tmprss15	transmembrane protease, serine 15, transcript variant 1
chr13	21827706 AATAAA	AATAAA	9	4	XM_006517233.1	Ttc37	tetratricopeptide repeat domain 37, transcript variant X2
chr13	21827706 AATAAA	AATAAA	9	4	XM_006517232.1	Ttc37	tetratricopeptide repeat domain 37, transcript variant X1
chr13	21827706 AATAAA	AATAAA	9	4	NM_001081352.1	Ttc37	tetratricopeptide repeat domain 37
chr13	21827706 AATAAA	AATAAA	9	4	XM_006517234.2	Ttc37	tetratricopeptide repeat domain 37, transcript variant X3
chr13	21827706 AATAAA	AATAAA	9	4	XM_006517235.2	Ttc37	tetratricopeptide repeat domain 37, transcript variant X4
chr1	44059182 ATTAAA	ATTAAA	19	20	XM_011238547.1	Unc80	unc-80 homolog (C. elegans), transcript variant X9
chr1	44059182 ATTAAA	ATTAAA	19	20	XM_011238540.1	Unc80	unc-80 homolog (C. elegans), transcript variant X1
chr1	44059182 ATTAAA	ATTAAA	19	20	XM_011238544.1	Unc80	unc-80 homolog (C. elegans), transcript variant X6

Table 10 (continued)

Chr	Testis PolyA-site PAS	Liver PAS	Testis Reads	Liver Reads	Accession Umber	Short Gene Name	Long Gene Name
chr1	44059182 ATTAAA	ATTAAA	19	20	XM_011238545.1	Unc80	unc-80 homolog (C. elegans), transcript variant X7
chr1	44059182 ATTAAA	ATTAAA	19	20	XM_011238541.1	Unc80	unc-80 homolog (C. elegans), transcript variant X2
chr1	44059182 ATTAAA	ATTAAA	19	20	XM_011238546.1	Unc80	unc-80 homolog (C. elegans), transcript variant X8
chr1	44059182 ATTAAA	ATTAAA	19	20	XM_006496086.2	Unc80	unc-80 homolog (C. elegans), transcript variant X5
chr1	44059182 ATTAAA	ATTAAA	19	20	XM_006496092.2	Unc80	unc-80 homolog (C. elegans), transcript variant X10
chr1	44059182 ATTAAA	ATTAAA	19	20	XM_011238542.1	Unc80	unc-80 homolog (C. elegans), transcript variant X3
chr1	44059182 ATTAAA	ATTAAA	19	20	XM_011238543.1	Unc80	unc-80 homolog (C. elegans), transcript variant X4
chr1	44059182 ATTAAA	ATTAAA	19	20	XM_006496093.2	Unc80	unc-80 homolog (C. elegans), transcript variant X11
chr1	44059182 ATTAAA	ATTAAA	19	20	NM_175510.3	Unc80	unc-80 homolog (C. elegans)
chr12	74644387 AATAAA	AATAAA	3	12	XR_381520.2	Wdr60	WD repeat domain 60, transcript variant X2
chr12	74644387 AATAAA	AATAAA	3	12	XM_006515780.2	Wdr60	WD repeat domain 60, transcript variant X1
chr12	74644387 AATAAA	AATAAA	3	12	XM_011244097.1	Wdr60	WD repeat domain 60, transcript variant X3
chr12	74644387 AATAAA	AATAAA	3	12	NM_146039.3	Wdr60	WD repeat domain 60
chr2	60090049 AATAAA	AATAAA	43	9	NM_028054.3	Zfyve19	zinc finger, FYVE domain containing 19, transcript variant 1
chr2	60090049 AATAAA	AATAAA	43	9	NM_001164827.1	Zfyve19	zinc finger, FYVE domain containing 19, transcript variant 2
chr2	60090049 AATAAA	AATAAA	43	9	XM_006500207.2	Zfyve19	zinc finger, FYVE domain containing 19, transcript variant X6
chr2	60090049 AATAAA	AATAAA	43	9	XM_006500204.2	Zfyve19	zinc finger, FYVE domain containing 19, transcript variant X3
chr2	60090049 AATAAA	AATAAA	43	9	XM_006500203.2	Zfyve19	zinc finger, FYVE domain containing 19, transcript variant X1
chr2	60090049 AATAAA	AATAAA	43	9	XM_006500205.1	Zfyve19	zinc finger, FYVE domain containing 19, transcript variant X4
chr2	60090049 AATAAA	AATAAA	43	9	XM_011239781.1	Zfyve19	zinc finger, FYVE domain containing 19, transcript variant X2
chr2	60090049 AATAAA	AATAAA	43	9	XM_006500206.2	Zfyve19	zinc finger, FYVE domain containing 19, transcript variant X5

Chr	Testis PolyA-site	Liver PolyA-site	Testis PAS	Liver PAS	Signal Match	Testis Reads	Liver Reads	Abs Diff	Ratio Ranking
chr16	10530301	10530347	ATTAAA	CATAAA	FALSE	1834	1	1833	1834
chr2	179910989	179911079	AATAAA	ATTAAA	FALSE	2712	7	2705	387.428571
chr15	76176363	76176413	AATAAA	ACTAAA	FALSE	283	1	282	283
chr2	29946464	29946528	ATTAAA	AATAAA	FALSE	238	1	237	238
chr8	73928447	73928530	AATAAA	ACTAAA	FALSE	817	4	813	204.25
chr16	20605545	20605503	AATAAA	CATAAA	FALSE	685	4	681	171.25
chr11	109529107	109529046	AATAAA	ACTAAA	FALSE	332	2	330	166
chr11	4793651	4793687	AATAAA	CATAAA	FALSE	133	1	132	133
chr17	80600650	80600707	AATAAA	AGTAAA	FALSE	132	1	131	132
chr19	6985513	6985467	ATTAAA	AATATA	FALSE	1210	10	1200	121
chr11	115939887	115939930	AGTAAA	AATAAA	FALSE	327	3	324	109
chr17	23961710	23961702	AGTAAA	AATAAA	FALSE	5481	52	5429	105.403846
chr10	77028419	77028367	AATAAA	TATAAA	FALSE	286	3	283	95.3333333
chr2	26313425	26313510	AATAAA	AATACA	FALSE	86	1	85	86
chr9	14315005	14315069	AATAAA	TATAAA	FALSE	235	3	232	78.3333333
chr11	70268691	70268644	ATTAAA	CATAAA	FALSE	515	7	508	73.5714286
chr3	57638751	57638686	AATAAA	AATATA	FALSE	68	1	67	68
chr4	106722297	106722253	ATTAAA	AAAAAG	FALSE	197	3	194	65.6666667
chr7	29161807	29161861	AATAAA	AATATA	FALSE	233	4	229	58.25
chr7	104561446	104561435	ATTAAA	AATAAA	FALSE	44	1	43	44

Table 11. Mouse honed transcript subset

Mouse honed transcripts, where testis and liver transcripts are matched to the same polyA-site within a 100bp upstream and 100bp downstream, and extracted to retain transcripts using different PAS, sorted by absolute difference in read counts in testis and liver.

	Testis	Liver
Transcripts with Isoforms	61	83
Transcripts with 2 Isoforms	44	64
Transcripts with 3 or More Isoforms	17	19

Table 12. Individual testis and liver tissue transcript annotation numerical data from mouse

PolyA-seq tissue files from mouse were individually annotated to identify potential transcript isoforms using different polyA-sites in alternative polyadenylation.

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr2	9837701	9.133	AATAAA	4	-	NM_198001.2	1110008P14Rik	RIKEN cDNA 1110008P14 gene
chr9	37351784	9.053	ATTAAA	29	+	XM_006510578.2	1110032A03Rik	RIKEN cDNA 1110032A03 gene, transcript variant X2
chr11	6515601	5.396	AATAAA	206	-	XM_006514705.1	Abca13	ATP-binding cassette, sub-family A (ABC1), member 13, transcript variant X1
chr11	6158572	3.071	AATATA	9	-	XM_006514705.1	Abca13	ATP-binding cassette, sub-family A (ABC1), member 13, transcript variant X1
chr11	75387292	6.812	AATAAA	51	+	XM_006533063.1	Abca8a	ATP-binding cassette, sub-family A (ABC1), member 8a, transcript variant X3
chr8	13884794	5.685	AATAAA	9	-	XR_387742.2	Abcc12	ATP-binding cassette, sub-family C (CFTR/MRP), member 12, transcript variant X2
chr5	9119039	7.446	AATAAA	1	+	NM_001190443.1	Abcf2	ATP-binding cassette, sub-family F (GCN20), member 2, transcript variant 2
chr5	8068539	5.164	AATAAA	4	+	XM_011248231.1	Acad12	acyl-Coenzyme A dehydrogenase family, member 12, transcript variant X1
chr2	32090041	3.79	AATAAA	99	+	XM_011239257.1	Acp2	acid phosphatase 2, lysosomal, transcript variant X1
chr2	32090041	3.79	AATAAA	99	+	XM_011239257.1	Acp2	acid phosphatase 2, lysosomal, transcript variant X1
chr7	27969892	10.12	TATAAA	1	-	XM_006540843.2	Adamts17	a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 17, transcript variant X2
chr2	4543105	3.347	AGTAAA	12	-	NM_029981.1	Adamtsl2	ADAMTS-like 2
chr2	4543264	7.798	ATTAAA	4	-	NM_029981.1	Adamtsl2	ADAMTS-like 2
chr2	4543264	7.798	ATTAAA	4	-	NM_029981.1	Adamtsl2	ADAMTS-like 2
chr5	36916738	7.348	ATTAAA	6	-	NM_177078.4	Adrbk2	adrenergic receptor kinase, beta 2, transcript variant 1
chr2	4053110	9.125	ATTAAA	15	-	NM_026212.2	Agpat2	1-acylglycerol-3-phosphate O-acyltransferase 2 (lysophosphatidic acid acyltransferase, beta)
chr4	8794797	3.233	AATAAA	5	+	XR_390721.2	Akr7a5	aldo-keto reductase family 7, member A5 (aflatoxin aldehyde reductase), transcript variant X1

Table 13. Testis tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr1	36584059	4.796	AATAAA	31	+	NM_175200.4	Als2cr11	amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 11 (human)
chr1	36595984	5.005	AATAAA	6	-	NM_175200.4	Als2cr11	amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 11 (human)
chr1	36595984	5.005	AATAAA	6	-	XM_011238605.1	Als2cr11	amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 11 (human), transcript variant X1
chr3	57262638	5.236	AATAAA	9	-	XM_006500868.2	Ampd2	adenosine monophosphate deaminase 2, transcript variant X1
chr7	25426119	5.098	AATAAA	6	+	XM_006540573.2	Apba2	amyloid beta (A4) precursor protein-binding, family A, member 2, transcript variant X1
chr2	17918048	7.251	ATTAAA	10	-	XM_006497541.2	Arhgap21	Rho GTPase activating protein 21, transcript variant X4
chr7	13494431	7.196	AATAAA	36	+	NM_172739.4	Arhgap35	Rho GTPase activating protein 35
chr7	13494431	7.196	AATAAA	36	+	NM_172739.4	Arhgap35	Rho GTPase activating protein 35
chr13	21881080	4.578	AATAAA	14	+	XM_006517628.2	Arl15	ADP-ribosylation factor-like 15, transcript variant X1
chr2	30650820	4.681	GATAAA	8	-	NM_022989.4	Arl6ip6	ADP-ribosylation factor-like 6 interacting protein 6
chr7	19683896	3.969	TATAAA	6	+	XM_006540581.2	Atp10a	ATPase, class V, type 10A, transcript variant X2
chr7	19723603	8.193	TATAAA	60	+	XM_006540581.2	Atp10a	ATPase, class V, type 10A, transcript variant X2
chr7	19683896	3.969	TATAAA	6	+	XM_006540581.2	Atp10a	ATPase, class V, type 10A, transcript variant X2
chr7	19723603	8.193	TATAAA	60	+	XM_006540581.2	Atp10a	ATPase, class V, type 10A, transcript variant X2
chr1	16122628	8.049	ATTAAA	2	+	NM_178405.3	Atp1a2	ATPase, Na+/K+ transporting, alpha 2 polypeptide
chr6	83002564	6.645	AATAAA	1	+	XM_011241162.1	Atp2b2	ATPase, Ca++ transporting, plasma membrane 2
chr8	47082259	3.826	ATTAAA	3	-	XM_006531344.1	Atp2c2	ATPase, Ca++ transporting, type 2C, member 2
chr11	59735390	5.478	AATAAA	2	+	NM_001177890.1	Cacna1g	calcium channel, voltage-dependent, T type, alpha 1G subunit, transcript variant 4
chr11	59784116	8.685	AATAAA	16	-	NM_001177890.1	Cacna1g	calcium channel, voltage-dependent, T type, alpha 1G subunit, transcript variant 4
chr11	59775705	5.369	ATTAAA	6	+	NM_001177890.1	Cacna1g	calcium channel, voltage-dependent, T type, alpha 1G subunit, transcript variant 4

Table 13. Testis tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr10	4547347	7.907	AATAAA	5	+	NM_026201.3	Ccar1	cell division cycle and apoptosis regulator 1
chr10	4547380	5.292	AATAAA	4	+	NM_026201.3	Ccar1	cell division cycle and apoptosis regulator 1
chrX	19674643	6.073	AATAAA	1	+	NM_001024624.2	Cdkl5	cyclin-dependent kinase-like 5
chrX	19674508	3.245	ATTAAA	2	+	NM_001024624.2	Cdkl5	cyclin-dependent kinase-like 5
chr1	34530629	7.946	AATAAA	1	+	XM_011247976.1	Cfap221	cilia and flagella associated protein 221, transcript variant X1
chr1	34502591	9.349	AGTAAA	42	+	XM_011247976.1	Cfap221	cilia and flagella associated protein 221, transcript variant X1
chr1	34490051	7.354	GATAAA	2	-	XM_011247976.1	Cfap221	cilia and flagella associated protein 221, transcript variant X1
chr17	35804672	8.341	AATAAA	15	-	XM_006525090.2	Daam2	dishevelled associated activator of morphogenesis 2, transcript variant X10
chr17	35818609	3.445	ATTAAA	1	+	XM_006525090.2	Daam2	dishevelled associated activator of morphogenesis 2, transcript variant X10
chr2	32819563	8.03	AATAAA	55	+	NM_138306.2	Dgkz	diacylglycerol kinase zeta, transcript variant 2
chr2	32820008	3.313	AATAAA	20	+	NM_138306.2	Dgkz	diacylglycerol kinase zeta, transcript variant 2
chr2	32819563	8.03	AATAAA	55	+	NM_138306.2	Dgkz	diacylglycerol kinase zeta, transcript variant 2
chr2	32820008	3.313	AATAAA	20	+	NM_138306.2	Dgkz	diacylglycerol kinase zeta, transcript variant 2
chr12	76551868	6.777	AATAAA	4	-	NM_010060.3	Dnah11	dynein, axonemal, heavy chain 11
chr12	76553476	3.426	AATAAA	25	-	NM_010060.3	Dnah11	dynein, axonemal, heavy chain 11
chr2	9778698	4.096	AATAAA	1	-	XM_006497653.2	Dnm1	dynamin 1, transcript variant X7
chr2	9778937	8.176	ATTAAA	9	-	XM_006497653.2	Dnm1	dynamin 1, transcript variant X7
chr2	9778698	4.096	AATAAA	1	-	XM_006497653.2	Dnm1	dynamin 1, transcript variant X7
chr2	9778937	8.176	ATTAAA	9	-	XM_006497653.2	Dnm1	dynamin 1, transcript variant X7

Table 13. Testis tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr11	70341354	7.122	AATAAA	7	+	XM_011249261.1	Efcab3	EF-hand calcium binding domain 3
chr11	70430382	7.371	AATAAA	4	+	XM_011249261.1	Efcab3	EF-hand calcium binding domain 3
chr11	70268691	6.759	ATTAAA	515	-	XM_011249261.1	Efcab3	EF-hand calcium binding domain 3
chr11	70268648	8.453	CATAAA	5	-	XM_011249261.1	Efcab3	EF-hand calcium binding domain 3X1
chr7	28826731	6.082	AATAAA	336	-	NM_001040131.2	Eif4g2	eukaryotic translation initiation factor 4, gamma 2
chr7	28825669	5.683	ATTAAA	25	-	NM_001040131.2	Eif4g2	eukaryotic translation initiation factor 4, gamma 2
chr16	18220846	6.13	AATAAA	63	-	NM_010143.1	Ephb3	Eph receptor B3
chr16	18222321	7.349	ATTAAA	201	-	NM_010143.1	Ephb3	Eph receptor B3
chr5	4081139	9.566	ATTAAA	4	-	NM_001243123.1	Fam193a	family with sequence similarity 193, member A
chr5	4081299	10.55	ATTAAA	10	-	NM_001243123.1	Fam193a	family with sequence similarity 193, member A
chr5	4081139	9.566	ATTAAA	4	-	NM_001243123.1	Fam193a	family with sequence similarity 193, member A
chr5	4081299	10.55	ATTAAA	10	-	NM_001243123.1	Fam193a	family with sequence similarity 193, member A
chr5	34334229	7.196	AATAAA	3	-	XM_011249494.1	Fbrsl1	fibrosin-like 1, transcript variant X13
chr5	34349836	7.15	AGTAAA	5	-	XM_011249494.1	Fbrsl1	fibrosin-like 1, transcript variant X13
chr5	31458378	4.808	AATAAA	10	-	NM_001161738.1	Glmn	glomulin, FKBP associated protein
chr5	31462146	7.73	AATAAA	11	-	NM_001161738.1	Glmn	glomulin, FKBP associated protein
chr5	31459581	7.342	AGTAAA	15	-	NM_001161738.1	Glmn	glomulin, FKBP associated protein
chr5	31458981	6.007	ATTAAA	24	-	NM_001161738.1	Glmn	glomulin, FKBP associated protein
chr13	14717853	6.225	AATAAA	1264	+	XR_873900.1	Gm30411	predicted gene, 30411, transcript variant X4
chr13	14717773	3.263	ACTAAA	1	-	XR_873900.1	Gm30411	predicted gene, 30411, transcript variant X4
chr13	14717939	8.688	ATTAAA	8	+	XR_873900.1	Gm30411	predicted gene, 30411, transcript variant X4
chr18	50355837	6.924	AATAAA	157	+	XM_006526407.2	Gm33438	predicted gene, 33438, transcript variant X2
chr18	50355915	10.43	AATAAA	8	+	XM_006526407.2	Gm33438	predicted gene, 33438, transcript variant X2

Table 13. Testis tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr7	31202601	3.259	AATAAA	91	+	NM_008155.4	Gpi1	glucose phosphate isomerase 1
chr7	31202731	7.696	AATAAA	129	-	NM_008155.4	Gpi1	glucose phosphate isomerase 1
chr7	31202735	6.007	AATAAA	8	+	NM_008155.4	Gpi1	glucose phosphate isomerase 1
chr7	31202601	3.259	AATAAA	91	+	NM_008155.4	Gpi1	glucose phosphate isomerase 1
chr7	31202731	7.696	AATAAA	129	-	NM_008155.4	Gpi1	glucose phosphate isomerase 1
chr7	31202735	6.007	AATAAA	8	+	NM_008155.4	Gpi1	glucose phosphate isomerase 1
chr11	72989006	5.947	AGTAAA	4	-	XM_006534505.2	Helz	helicase with zinc finger domain, transcript variant X2
chr11	72989698	8.984	ATTAAA	100	-	XM_006534505.2	Helz	helicase with zinc finger domain, transcript variant X2
chr7	17080387	5.191	AATAAA	25	-	NM_010418.2	Herc2	hect (homologous to the E6-AP (UBE3A) carboxyl terminus) domain and RCC1 (CHC1)-like domain (RLD) 2
chr7	17021041	9.105	CATAAA	1	+	NM_010418.2	Herc2	hect (homologous to the E6-AP (UBE3A) carboxyl terminus) domain and RCC1 (CHC1)-like domain (RLD) 2
chr7	19576908	5.719	AATAAA	17	-	NM_010567.2	Inppl1	inositol polyphosphate phosphatase-like 1, transcript variant 1
chr7	19574578	7.454	ATTAAA	37	-	NM_010567.2	Inppl1	inositol polyphosphate phosphatase-like 1, transcript variant 1
chr7	19575195	6.171	ATTAAA	1	-	NM_010567.2	Inppl1	inositol polyphosphate phosphatase-like 1, transcript variant 1
chr16	31153603	6.337	AATAAA	2	-	XM_006522383.2	Kalrn	kalirin, RhoGEF kinase, transcript variant X25
chr16	31153435	8.592	ATTAAA	1	-	XM_006522383.2	Kalrn	kalirin, RhoGEF kinase, transcript variant X25
chr16	31514010	5.374	ATTAAA	9	+	XM_006522363.2	Kalrn	kalirin, RhoGEF kinase, transcript variant X6
chr17	39984006	9.19	AAAAAG	5	+	XM_006523820.2	Kat2b	K(lysine) acetyltransferase 2B, transcript variant X2
chr17	39983866	7.32	AATACA	17	+	XM_006523820.2	Kat2b	K(lysine) acetyltransferase 2B, transcript variant X2
chr2	3359478	7.091	AATAAA	11	+	XM_006497891.2	Kcnt1	potassium channel, subfamily T, member 1, transcript variant X12
chr2	3365863	4.699	AATAAA	2	-	XM_006497891.2	Kcnt1	potassium channel, subfamily T, member 1, transcript variant X12
chr2	3359478	7.091	AATAAA	11	+	XM_006497891.2	Kcnt1	potassium channel, subfamily T, member 1, transcript variant X12
chr2	3365863	4.699	AATAAA	2	-	XM_006497891.2	Kcnt1	potassium channel, subfamily T, member 1, transcript variant X12

Table 13. Testis tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr9	21723148	6.3	AATAAA	12	-	NM_001039522.1	Leo1	Leo1, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)
chr9	21725453	5.206	AATAAA	51	-	NM_001039522.1	Leo1	Leo1, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)
chr9	21723148	6.3	AATAAA	12	-	NM_001039522.1	Leo1	Leo1, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)
chr9	21725453	5.206	AATAAA	51	-	NM_001039522.1	Leo1	Leo1, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)
chr10	62021220	5.765	AATAAA	12	-	NM_025431.2	Llph	LLP homolog, long-term synaptic facilitation (Aplysia)
chr10	62021286	4.173	ACTAAA	6	-	NM_025431.2	Llph	LLP homolog, long-term synaptic facilitation (Aplysia)
chr17	29840278	5.263	AGTAAA	3	-	XM_006523751.2	Mep1a	meprin 1 alpha, transcript variant X1
chr17	29840303	8.118	ATTAAA	126	+	XM_006523751.2	Mep1a	meprin 1 alpha, transcript variant X1
chr6	91438451	7.43	AATAAA	9	+	NR_027619.1	Mug-ps1	murinoglobulin, pseudogene 1
chr6	91439302	3.192	AATAAA	2	-	NR_027619.1	Mug-ps1	murinoglobulin, pseudogene 1
chr17	57343392	7.794	AATAAA	3736	-	XM_006524822.1	Myl12a	myosin, light chain 12A, regulatory, non-sarcomeric, transcript variant X1
chr17	57343222	5.157	ATTAAA	3	-	XM_006524822.1	Myl12a	myosin, light chain 12A, regulatory, non-sarcomeric, transcript variant X1
chr17	57321041	6.123	ATTAAA	1	+	NM_023402.2	Myl12b	myosin, light chain 12B, regulatory
chr11	3420749	4.939	AAAAAG	4	-	XM_006514702.1	Myo1g	myosin IG, transcript variant X3
chr11	3417525	7.665	GATAAA	15	-	XM_006514702.1	Myo1g	myosin IG, transcript variant X3
chr19	3407783	7.144	AATAAA	9	+	XM_011247222.1	Myrf	myelin regulatory factor, transcript variant X2
chr19	3409917	7.285	AATAAA	13	-	XM_011247222.1	Myrf	myelin regulatory factor, transcript variant X2
chr2	29594973	5.953	AATAAA	12	-	XM_006497751.1	Neb	nebulin, transcript variant X3
chr2	29778055	4.836	AATAAA	20	+	XM_006497751.1	Neb	nebulin, transcript variant X3
chr2	29787264	5.579	AATAAA	12	+	XM_006497751.1	Neb	nebulin, transcript variant X3
chr7	19881507	7.307	AATAAA	14	+	XM_006507896.1	Nup98	nucleoporin 98, transcript variant X2
chr7	19888045	7.508	AATAAA	2	-	XM_006507896.1	Nup98	nucleoporin 98, transcript variant X2
chr7	19881507	7.307	AATAAA	14	+	XM_006507896.1	Nup98	nucleoporin 98, transcript variant X2

Table 13. Testis tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr2	26803584	8.82	ATTAAA	3	-	NM_146584.2	Olfr1026	olfactory receptor 1026
chr2	30881726	8.985	AATAAA	158	-	NM_146974.1	Olfr1262	olfactory receptor 1262
chr19	5273932	6.75	AATAAA	1204	-	NM_146809.2	Olfr1426	olfactory receptor 1426
chr5	32387281	5.096	AATAAA	7	+	XM_006535208.2	Pcgf3	polycomb group ring finger 3, transcript variant X1
chr5	32387317	5.401	AATAAA	399	+	XM_006535208.2	Pcgf3	polycomb group ring finger 3, transcript variant X1
chr19	10648166	5.924	ATTAAA	4	-	NM_001190483.1	Pcsk5	proprotein convertase subtilisin/kexin type 5, transcript variant 1
chr19	10980019	3.19	ATTAAA	26	+	NM_001190483.1	Pcsk5	proprotein convertase subtilisin/kexin type 5, transcript variant 1
chr16	10847799	3.493	AATAAA	1	+	XM_011246053.1	Pdxdc1	pyridoxal-dependent decarboxylase domain containing 1, transcript variant X1
chr16	10840465	3.434	ATTAAA	7	+	XM_011246053.1	Pdxdc1	pyridoxal-dependent decarboxylase domain containing 1, transcript variant X1
chr8	4247895	8.623	AATAAA	761	+	XM_006509228.2	Plekha2	pleckstrin homology domain-containing, family A (phosphoinositide binding specific) member 2, transcript variant X2
chr8	4259731	5.676	AATAAA	28	-	XM_006509228.2	Plekha2	pleckstrin homology domain-containing, family A (phosphoinositide binding specific) member 2, transcript variant X2
chr7	25370460	6.626	AATAAA	8	-	XM_006539437.2	Plekhg2	pleckstrin homology domain containing, family G (with RhoGef domain) member 2, transcript variant X3
chr9	38550660	3.326	ACTAAA	6	+	NM_001195084.1	Plscr2	phospholipid scramblase 2, transcript variant A
chr9	38550511	6.71	TATAAA	2	+	NM_001195084.1	Plscr2	phospholipid scramblase 2, transcript variant A
chr7	16903551	5.614	AATAAA	19	+	NM_027514.2	Pvr	poliovirus receptor
chr7	16905822	4.132	ATTAAA	5	-	NM_027514.2	Pvr	poliovirus receptor
chr16	4040027	5.665	AATACA	45	-	XM_006522171.2	Rbfox1	RNA binding protein, fox-1 homolog (C. elegans) 1, transcript variant X15
chr16	4038350	7.275	AGTAAA	11	-	XM_006522192.2	Rbfox1	RNA binding protein, fox-1 homolog (C. elegans) 1, transcript variant X35
chr16	4039974	6.109	ATTAAA	28	-	XM_006522171.2	Rbfox1	RNA binding protein, fox-1 homolog (C. elegans) 1, transcript variant X15

Table 13. Testis tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr1	13129199	5.59	AATAAA	1	-	XR_865164.1	Rdh10	retinol dehydrogenase 10 (all-trans), transcript variant X1
chr1	13129239	4.651	AATAAA	1	-	XR_865164.1	Rdh10	retinol dehydrogenase 10 (all-trans), transcript variant X1
chr1	13131364	3.835	AATAAA	5	-	XR_865164.1	Rdh10	retinol dehydrogenase 10 (all-trans), transcript variant X1
chr7	16611681	5.929	ATTAAA	4	+	XM_011250461.1	Relb	avian reticuloendotheliosis viral (v-rel) oncogene related B, transcript variant X4
chr11	84797937	9.286	AGTAAA	2	-	XM_006534630.2	Rnf213	ring finger protein 213, transcript variant X1
chr11	84797991	7.227	ATTAAA	15	-	XM_006534630.2	Rnf213	ring finger protein 213, transcript variant X1
chr7	20216933	4.926	AATAAA	12	-	NM_009103.3	Rrm1	ribonucleotide reductase M1
chr7	20191567	9.61	AATAAA	58	-	NM_009103.3	Rrm1	ribonucleotide reductase M1
chr7	20216933	4.926	AATAAA	12	-	NM_009103.3	Rrm1	ribonucleotide reductase M1
chr19	40813957	3.213	AATAAA	5	+	NM_009289.3	Slk	STE20-like kinase, transcript variant 1
chr19	40814024	3.409	GATAAA	7	+	NM_009289.3	Slk	STE20-like kinase, transcript variant 1
chr5	8069175	9.224	AATAAA	5	+	NM_009274.2	Srpk2	serine/arginine-rich protein specific kinase 2
chr5	8067773	3.542	TATAAA	53	+	NM_009274.2	Srpk2	serine/arginine-rich protein specific kinase 2
chr5	8067858	3.175	TATAAA	28	+	NM_009274.2	Srpk2	serine/arginine-rich protein specific kinase 2
chr3	88392340	8.976	AATAAA	464	+	XR_375536.2	Stpg2	sperm tail PG rich repeat containing 2, transcript variant X1
chr3	88887897	7.014	AATAAA	8	-	XR_375536.2	Stpg2	sperm tail PG rich repeat containing 2, transcript variant X1
chr3	88888252	6.74	AATAAA	14	-	XM_006501628.1	Stpg2	sperm tail PG rich repeat containing 2, transcript variant X3
chr3	88392340	8.976	AATAAA	464	+	XR_375536.2	Stpg2	sperm tail PG rich repeat containing 2, transcript variant X1
chr3	88887897	7.014	AATAAA	8	-	XR_375536.2	Stpg2	sperm tail PG rich repeat containing 2, transcript variant X1
chr3	88888252	6.74	AATAAA	14	-	XM_006501628.1	Stpg2	sperm tail PG rich repeat containing 2, transcript variant X3

Table 13. Testis tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr7	27992192	5.12	AATAAA	52	-	NM_009302.3	Swap70	SWA-70 protein
chr7	27992068	7.603	AATACA	6	+	NM_009302.3	Swap70	SWA-70 protein
chr7	27992268	5.032	AGTAAA	3	+	NM_009302.3	Swap70	SWA-70 protein
chr5	33976815	7.34	AATAAA	7	-	XM_011240758.1	Tbc1d1	TBC1 domain family, member 1, transcript variant X3
chr5	33971346	5.59	ATTAAA	4	-	XM_011240758.1	Tbc1d1	TBC1 domain family, member 1, transcript variant X3
chr4	43579800	6.99	AATAAA	147	-	XM_006538046.2	Tbc1d2	TBC1 domain family, member 2, transcript variant X1
chr4	43579931	8.811	AATAAA	45	+	XM_006538046.2	Tbc1d2	TBC1 domain family, member 2, transcript variant X1
chr4	43578421	4.635	ATTAAA	11	-	XM_006538046.2	Tbc1d2	TBC1 domain family, member 2, transcript variant X1
chr11	62418307	5.24	AATAAA	676	+	NM_019507.2	Tbx21	T-box 21
chr11	62418387	3.647	AATACA	87	-	NM_019507.2	Tbx21	T-box 21
chr2	25322231	6.731	AATAAA	20	-	XM_006499150.1	Tfpi	tissue factor pathway inhibitor, transcript variant X6
chr2	25354175	10.17	ATTAAA	1	-	XM_006499145.1	Tfpi	tissue factor pathway inhibitor, transcript variant X1
chr4	43040893	5.503	AATAAA	14	-	XM_006537772.2	Tmod1	tropomodulin 1, transcript variant X1
chr4	43066124	5.093	TATAAA	14	+	NM_021883.2	Tmod1	tropomodulin 1
chr6	94554524	6.404	AATAAA	31	-	NM_011609.4	Tnfrsf1a	tumor necrosis factor receptor superfamily, member 1a
chr6	94554563	6.374	ATTAAA	6	+	NM_011609.4	Tnfrsf1a	tumor necrosis factor receptor superfamily, member 1a
chr14	56375931	8.86	AATAAA	3	+	NM_009429.3	Tpt1	tumor protein, translationally-controlled 1
chr14	56375975	6.891	ATTAAA	3	+	NM_009429.3	Tpt1	tumor protein, translationally-controlled 1
chr7	19833213	6.727	AATAAA	151	+	NM_001109897.2	Trpc2	transient receptor potential cation channel, subfamily C, member 2
chr7	19842280	5.183	AATAAA	774	-	NM_001109897.2	Trpc2	transient receptor potential cation channel, subfamily C, member 2
chr7	19833213	6.727	AATAAA	151	+	NM_001109897.2	Trpc2	transient receptor potential cation channel, subfamily C, member 2
chr7	19842280	5.183	AATAAA	774	-	NM_001109897.2	Trpc2	transient receptor potential cation channel, subfamily C, member 2

Table 13. Testis tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr5	34694135	5.585	AAAAAG	15	+	XM_006534851.2	Ulk1	unc-51 like kinase 1, transcript variant X3
chr7	34969647	7.521	ATTAAA	5	-	XM_006539676.2	Uri1	URI1, prefoldin-like chaperone, transcript variant X1
chr10	21088592	8.57	AATAAA	1	+	NM_175936.1	Vmn2r81	vomeronasal 2, receptor 81
chr10	21088688	5.255	AATAAA	6	+	NM_175936.1	Vmn2r81	vomeronasal 2, receptor 81
chr5	34906201	3.272	AATAAA	5	+	XM_011240713.1	Wdr19	WD repeat domain 19, transcript variant X1
chr5	34906286	5.444	AATAAA	10	+	XM_011240713.1	Wdr19	WD repeat domain 19, transcript variant X1
chr7	19833213	6.727	AATAAA	151	+	NR_104582.1	Xntrpc	Xndc1-transient receptor potential cation channel, subfamily C, member 2 readthrough, transcript variant 2
chr7	19842280	5.183	AATAAA	774	-	NR_104582.1	Xntrpc	Xndc1-transient receptor potential cation channel, subfamily C, member 2 readthrough, transcript variant 2
chr17	32550039	6.27	ATTAAA	18	-	NM_028198.2	Xpo5	exportin 5
chr17	32553462	6.879	ATTAAA	35	-	NM_028198.2	Xpo5	exportin 5
chr6	87042190	8.119	AATAAA	15	+	XM_006506015.2	Zfp637	zinc finger protein 637, transcript variant X1
chr6	87038276	6.172	AGTAAA	2	+	XM_006506015.2	Zfp637	zinc finger protein 637, transcript variant X1

Table 13. Testis tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr8	13884794	5.685	AATAAA	12	-	XR_387742.2	Abcc12	ATP-binding cassette, sub-family C (CFTR/MRP), member 12, transcript variant X2
chr8	13885506	3.284	ATTAAA	1	-	XR_387742.2	Abcc12	ATP-binding cassette, sub-family C (CFTR/MRP), member 12, transcript variant X2
chr5	9118869	3.982	AATAAA	9	-	NM_001190443.1	Abcf2	ATP-binding cassette, sub-family F (GCN20), member 2, transcript variant 2
chr5	9118989	7.59	AATAAA	4	+	NM_001190443.1	Abcf2	ATP-binding cassette, sub-family F (GCN20), member 2, transcript variant 2
chr5	9119039	7.446	AATAAA	23	+	NM_001190443.1	Abcf2	ATP-binding cassette, sub-family F (GCN20), member 2, transcript variant 2
chr5	9119122	4.327	ATTAAA	1	+	NM_001190443.1	Abcf2	ATP-binding cassette, sub-family F (GCN20), member 2, transcript variant 2
chr17	17403310	5.184	ATTAAA	12	+	NM_009593.2	Abcg1	ATP-binding cassette, sub-family G (WHITE), member 1
chr17	17456764	5.184	ATTAAA	2	+	NM_009593.2	Abcg1	ATP-binding cassette, sub-family G (WHITE), member 1
chr2	32090041	3.79	AATAAA	51	+	NM_007387.2	Acp2	acid phosphatase 2, lysosomal
chr2	32087222	4.308	AATATA	1	+	NM_007387.2	Acp2	acid phosphatase 2, lysosomal
chr12	29367036	5.107	AATAAA	984	+	NM_019785.2	Actr10	ARP10 actin-related protein 10
chr2	104610368	5.802	ATTAAA	3	-	NM_001272052.1	Ada	adenosine deaminase, transcript variant 1
chr2	104610431	8.556	ATTAAA	1	-	NM_001272052.1	Ada	adenosine deaminase, transcript variant 1
chr1	36584059	4.796	AATAAA	19	+	NM_175200.4	Als2cr11	amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 11 (human)
chr1	36595984	5.005	AATAAA	12	-	NM_175200.4	Als2cr11	amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 11 (human)
chr1	51797358	7.459	AATAAA	5	-	NM_029711.1	Arpc2	actin related protein 2/3 complex, subunit 2
chr1	51789624	7.189	ATTAAA	21	-	NM_029711.1	Arpc2	actin related protein 2/3 complex, subunit 2

Table 14. Liver tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr5	8959224	7.001	AATAAA	17	+	XR_880822.1	Asic3	acid-sensing (proton-gated) ion channel 3, transcript variant X4
chr5	8959281	5.66	AATATA	2	+	XR_880822.1	Asic3	acid-sensing (proton-gated) ion channel 3, transcript variant X4
chr7	19723762	6.93	AATAAA	19	-	XM_006540581.2	Atp10a	ATPase, class V, type 10A, transcript variant X2
chr7	19683898	5.134	TATAAA	2	+	XM_006540581.2	Atp10a	ATPase, class V, type 10A, transcript variant X2
chr7	19723603	8.193	TATAAA	3	+	XM_006540581.2	Atp10a	ATPase, class V, type 10A, transcript variant X2
chr8	47082210	3.608	AATAAA	18	+	NM_026922.1	Atp2c2	ATPase, Ca++ transporting, type 2C, member 2
chr8	47083460	4.236	AATACA	1	-	NM_026922.1	Atp2c2	ATPase, Ca++ transporting, type 2C, member 2
chr11	66850492	7.45	AATAAA	4	+	XR_879509.1	Brca1	breast cancer 1, transcript variant X2
chr11	66850447	3.664	GATAAA	4	+	XR_879509.1	Brca1	breast cancer 1, transcript variant X2
chr19	30172932	6.01	AATAAA	34	-	NM_001080706.1	Btaf1	BTAF1 RNA polymerase II, B-TFIID transcription factor-associated, (Mot1 homolog, S. cerevisiae)
chr19	30131551	6.044	ATTAAA	4	+	NM_001080706.1	Btaf1	BTAF1 RNA polymerase II, B-TFIID transcription factor-associated, (Mot1 homolog, S. cerevisiae)
chr1	10126849	6.484	AATAAA	7	+	XM_006496519.2	C130026I21Rik	RIKEN cDNA C130026I21 gene, transcript variant X9
chr1	10126752	7.306	ATTAAA	9	+	XM_006496519.2	C130026I21Rik	RIKEN cDNA C130026I21 gene, transcript variant X9
chr11	59784116	8.685	AATAAA	1504	-	NM_001177890.1	Cacna1g	calcium channel, voltage-dependent, T type, alpha 1G subunit, transcript variant 4
chr11	59775705	5.369	ATTAAA	7	+	NM_001177890.1	Cacna1g	calcium channel, voltage-dependent, T type, alpha 1G subunit, transcript variant 4
chr5	8960541	4.99	AATAAA	2	-	XR_389739.2	Cdk5	cyclin-dependent kinase 5, transcript variant X3
chr5	8963418	3.737	AATAAA	1	-	XM_006535626.2	Cdk5	cyclin-dependent kinase 5, transcript variant X1
chr4	3717410	8.608	AATAAA	1	+	NM_001081047.1	Cnksr1	connector enhancer of kinase suppressor of Ras 1
chr4	3718458	4.402	ATTAAA	4	+	NM_001081047.1	Cnksr1	connector enhancer of kinase suppressor of Ras 1

Table 14. Liver tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr6	5116584	7.108	AGTAAA	13	-	XM_011241052.1	Col28a1	collagen, type XXVIII, alpha 1, transcript variant X1
chr6	5112325	4.218	ATTAAA	4	-	XM_011241052.1	Col28a1	collagen, type XXVIII, alpha 1, transcript variant X1
chr6	5118096	5.181	ATTAAA	19	-	XM_011241052.1	Col28a1	collagen, type XXVIII, alpha 1, transcript variant X1
chr17	32848847	3.614	AATAAA	2	+	NM_001081335.2	Cul9	cullin 9
chr17	32847281	3.804	AGTAAA	3	+	NM_001081335.2	Cul9	cullin 9
chr19	3808832	7.722	AATAAA	5	+	XM_006526639.1	Ddb1	damage specific DNA binding protein 1, transcript variant X1
chr19	3816303	5.28	GATAAA	13	+	XM_006526639.1	Ddb1	damage specific DNA binding protein 1, transcript variant X1
chr2	32819563	8.03	AATAAA	27	+	NM_138306.2	Dgkz	diacylglycerol kinase zeta, transcript variant 2
chr2	32820008	3.313	AATAAA	18	+	NM_138306.2	Dgkz	diacylglycerol kinase zeta, transcript variant 2
chr12	76551869	5.402	AATAAA	5	-	NM_010060.3	Dnah11	dynein, axonemal, heavy chain 11
chr12	76553476	3.426	AATAAA	58	-	NM_010060.3	Dnah11	dynein, axonemal, heavy chain 11
chr11	70336456	4.146	AATAAA	4	+	XM_011249261.1	Efcab3	EF-hand calcium binding domain 3, transcript variant X1
chr11	70341360	9.013	AATAAA	8	+	XM_011249261.1	Efcab3	EF-hand calcium binding domain 3, transcript variant X1
chr11	70268691	6.759	ATTAAA	187	-	XM_011249261.1	Efcab3	EF-hand calcium binding domain 3, transcript variant X1
chr11	70268644	10.94	CATAAA	7	-	XM_011249261.1	Efcab3	EF-hand calcium binding domain 3, transcript variant X1
chr7	28826731	6.082	AATAAA	77	-	NM_001040131.2	Eif4g2	eukaryotic translation initiation factor 4, gamma 2, transcript variant 2
chr7	28825669	5.683	ATTAAA	8	-	NM_001040131.2	Eif4g2	eukaryotic translation initiation factor 4, gamma 2, transcript variant 2
chr7	28825725	5.91	CATAAA	4	-	NM_001040131.2	Eif4g2	eukaryotic translation initiation factor 4, gamma 2, transcript variant 2
chr4	3718458	4.402	ATTAAA	4	+	NM_001081047.1	Cnksr1	connector enhancer of kinase suppressor of Ras 1

Table 14. Liver tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr16	18220846	6.13	AATAAA	8	-	NM_010143.1	Ephb3	Eph receptor B3
chr16	18222321	7.349	ATTAAA	1	-	NM_010143.1	Ephb3	Eph receptor B3
chr5	4081141	7.584	ATTAAA	6	-	NM_001243123.1	Fam193a	family with sequence similarity 193, member A
chr5	4081301	9.566	ATTAAA	2	-	NM_001243123.1	Fam193a	family with sequence similarity 193, member A
chr2	12269220	8.528	AATAAA	8	-	NM_001164770.1	Fbxw2	F-box and WD-40 domain protein 2, transcript variant 4
chr2	12269431	3.755	ATTAAA	1	-	NM_001164770.1	Fbxw2	F-box and WD-40 domain protein 2, transcript variant 4
chr7	27823724	8.041	ACTAAA	7	+	XM_011250554.1	Ffar2	free fatty acid receptor 2, transcript variant X7
chr7	27823851	6.575	GATAAA	10	+	XM_011250554.1	Ffar2	free fatty acid receptor 2, transcript variant X7
chr5	31458378	4.808	AATAAA	10	-	NM_001161738.1	Glmn	glomulin, FKBP associated protein, transcript variant 2
chr5	31462146	7.73	AATAAA	11	-	NM_001161738.1	Glmn	glomulin, FKBP associated protein, transcript variant 2
chr5	31458981	6.007	ATTAAA	15	-	NM_001161738.1	Glmn	glomulin, FKBP associated protein, transcript variant 2
chr14	35225295	4.516	AATAAA	3	-	XM_006519864.2	Gm29776	predicted gene, 29776, transcript variant X2
chr14	35224245	8.582	AGTAAA	1	-	XM_006519864.2	Gm29776	predicted gene, 29776, transcript variant X2
chr13	14717853	6.225	AATAAA	2489	+	XR_873900.1	Gm30411	predicted gene, 30411, transcript variant X4
chr13	14717939	8.688	ATTAAA	32	+	XR_873900.1	Gm30411	predicted gene, 30411, transcript variant X4
chr7	31202601	3.259	AATAAA	19	+	NM_008155.4	Gpi1	glucose phosphate isomerase 1
chr7	31202729	9.25	AATAAA	3	-	NM_008155.4	Gpi1	glucose phosphate isomerase 1
chr10	24493619	4.413	AATAAA	4	-	XM_011243546.1	Hcfc2	host cell factor C2, transcript variant X7
chr10	24492888	5.231	ATTAAA	19	-	XR_871726.1	Hcfc2	host cell factor C2, transcript variant X9

Table 14. Liver tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr11	72989023	3.632	AATAAA	18	-	XM_006534505.2	Helz	helicase with zinc finger domain, transcript variant
chr11	72989698	8.984	ATTAAA	208	-	XM_006534505.2	Helz	helicase with zinc finger domain, transcript variant
chr13	3645770	9.748	AATAAA	5	-	XM_011244681.1	Iqgap2	IQ motif containing GTPase activating protein 2,
chr13	3653067	8.708	AATAAA	2	+	XM_011244681.1	Iqgap2	IQ motif containing GTPase activating protein 2,
chr18	20832824	6.976	AATAAA	371	+	NM_001162942.1	Mapre2	microtubule-associated protein, RP/EB family, member 2, transcript variant 3
chr18	20832910	6.286	AATAAA	18	+	NM_001162942.1	Mapre2	microtubule-associated protein, RP/EB family, member 2, transcript variant 3
chr6	5166126	3.789	AATAAA	11	-	XM_006505086.2	Mios	missing oocyte, meiosis regulator, homolog (Drosophila), transcript variant X1
chr6	5184289	7.77	ATTAAA	11	-	XM_006505086.2	Mios	missing oocyte, meiosis regulator, homolog
chr9	4386888	4.394	AATAAA	2	+	XM_006509892.1	Mmp1b	matrix metallopeptidase 1b (interstitial collagenase), transcript variant X1
chr9	4386712	7.588	AATACA	9	+	XM_006509892.1	Mmp1b	matrix metallopeptidase 1b (interstitial collagenase), transcript variant X1
chr6	91438451	7.43	AATAAA	3	+	NR_027619.1	Mug-ps1	murinoglobulin, pseudogene 1
chr6	91439296	8.479	AATAAA	9	-	NR_027619.1	Mug-ps1	murinoglobulin, pseudogene 1
chr17	57343392	7.794	AATAAA	206523	-	XM_006524822.1	Myl12a	myosin, light chain 12A, regulatory, non-sarcomeric,
chr17	57343222	5.157	ATTAAA	9	-	XM_006524822.1	Myl12a	myosin, light chain 12A, regulatory, non-sarcomeric, transcript variant X1
chr17	57343479	7.05	CATAAA	20	+	XM_006524822.1	Myl12a	myosin, light chain 12A, regulatory, non-sarcomeric, transcript variant X1
chr8	4259731	5.676	AATAAA	40	-	XM_006530574.2	Nr3c2	nuclear receptor subfamily 3, group C, member 2, transcript variant X1
chr8	4256485	4.268	TATAAA	3	+	XM_006530577.2	Nr3c2	nuclear receptor subfamily 3, group C, member 2, transcript variant X3

Table 14. Liver tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr1	30859902	7.005	AGTAAA	3	-	NR_104361.1	Osgepl1	O-sialoglycoprotein endopeptidase-like 1, transcript variant 3
chr1	30859665	4.227	TATAAA	1	-	XR_373262.2	Osgepl1	O-sialoglycoprotein endopeptidase-like 1, transcript variant X1
chr19	5106679	6.35	AAAAAG	1	+	NM_172635.3	Patl1	protein associated with topoisomerase II homolog 1 (yeast)
chr19	5107747	6.829	AATAAA	17	-	NM_172635.3	Patl1	protein associated with topoisomerase II homolog 1 (yeast)
chr19	5131574	7.006	AATAAA	1	-	NM_172635.3	Patl1	protein associated with topoisomerase II homolog 1 (yeast)
chr19	10664318	6.678	ATTAAA	8	+	NM_001190483.1	Pcsk5	proprotein convertase subtilisin/kexin type 5, transcript variant 1
chr19	10980026	8.908	ATTAAA	9	+	NM_001190483.1	Pcsk5	proprotein convertase subtilisin/kexin type 5, transcript variant 1
chr15	37305745	7.283	AATAAA	7	-	NM_011117.2	Plec	plectin, transcript variant 1
chr15	37296015	7.391	ATTAAA	2	-	NM_011117.2	Plec	plectin, transcript variant 1
chr16	34012000	3.439	AATAAA	36	-	XM_006522734.2	Polq	polymerase (DNA directed), theta, transcript variant X2
chr16	34013228	6.358	AATAAA	6	-	XM_006522734.2	Polq	polymerase (DNA directed), theta, transcript variant X2
chr9	37459367	3.908	AATAAA	4	+	XM_006510643.2	Ppp2r1b	protein phosphatase 2, regulatory subunit A, beta, transcript variant X4
chr9	37455839	3.041	AATATA	14	+	XM_006510643.2	Ppp2r1b	protein phosphatase 2, regulatory subunit A, beta, transcript variant X4
chr19	3385731	5.505	AATAAA	673	+	NM_011224.1	Pygm	muscle glycogen phosphorylase
chr19	3387996	3.602	AATAAA	3	+	NM_011224.1	Pygm	muscle glycogen phosphorylase
chr15	60743948	7.196	AATAAA	1	-	NM_001253809.1	Racgap1	Rac GTPase-activating protein 1, transcript variant 3
chr15	60748566	5.202	AATAAA	444	-	NM_001253809.1	Racgap1	Rac GTPase-activating protein 1, transcript variant 3
chr15	60748818	5.238	AATATA	15	-	NM_001253809.1	Racgap1	Rac GTPase-activating protein 1, transcript variant 3

Table 14. Liver tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr16	4040027	5.665	AATACA	147	-	XM_006522171.2	Rbfox1	RNA binding protein, fox-1 homolog (C. elegans) 1, transcript variant X15
chr16	4040018	4.8	AGTAAA	17	+	XM_006522171.2	Rbfox1	RNA binding protein, fox-1 homolog (C. elegans) 1, transcript variant X15
chr16	4039974	6.109	ATTAAA	191	-	XM_006522171.2	Rbfox1	RNA binding protein, fox-1 homolog (C. elegans) 1, transcript variant X15
chr1	13129239	4.651	AATAAA	15	-	XR_865164.1	Rdh10	retinol dehydrogenase 10 (all-trans), transcript variant X1
chr1	13131349	7.341	AATAAA	6	-	XR_865164.1	Rdh10	retinol dehydrogenase 10 (all-trans), transcript variant X1
chr11	84799763	5.723	AATAAA	4	-	XM_006534630.2	Rnf213	ring finger protein 213, transcript variant X1
chr11	84797944	8.78	AGTAAA	23	-	XM_006534630.2	Rnf213	ring finger protein 213, transcript variant X1
chr11	84797991	7.227	ATTAAA	125	-	XM_006534630.2	Rnf213	ring finger protein 213, transcript variant X1
chr16	8673174	9.264	AATAAA	1	+	XM_006523038.2	Robo2	roundabout homolog 2 (Drosophila), transcript
chr16	8621658	3.626	ATTAAA	5	+	XM_006523038.2	Robo2	roundabout homolog 2 (Drosophila), transcript variant X4
chr7	20191571	8.395	AATAAA	6	-	NM_009103.3	Rrm1	ribonucleotide reductase M1
chr7	20216929	5.309	AATAAA	13	-	NM_009103.3	Rrm1	ribonucleotide reductase M1
chr17	42929899	10.05	AATAAA	72	-	XM_006524159.1	Safb2	scaffold attachment factor B2, transcript variant X3
chr17	42930148	5.291	AATAAA	22	-	XM_006524157.2	Safb2	scaffold attachment factor B2, transcript variant X2
chr17	42910066	7.548	ATTAAA	1	-	XM_006524157.2	Safb2	scaffold attachment factor B2, transcript variant X2
chr2	104505580	8.538	AATAAA	15	+	NM_012032.4	Serinc3	serine incorporator 3
chr2	104510480	7.152	AATAAA	6	-	NM_012032.4	Serinc3	serine incorporator 3

Table 14. Liver tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr17	72008368	7.88	AATAAA	66	+	XM_011246335.1	Six2	sine oculis-related homeobox 2, transcript variant X2
chr17	72008327	7.448	TATAAA	1	+	XM_011246335.1	Six2	sine oculis-related homeobox 2, transcript variant X2
chr2	43538567	5.439	AATAAA	7	+	XM_006499047.2	Slc1a2	solute carrier family 1 (glial high affinity glutamate transporter), member 2, transcript variant X2
chr2	43537925	3.613	AATACA	1	+	XM_006499047.2	Slc1a2	solute carrier family 1 (glial high affinity glutamate transporter), member 2, transcript variant X2
chr11	62272947	9.323	AATAAA	19	+	XM_006534520.2	Sp2	Sp2 transcription factor, transcript variant X3
chr11	62274331	5.675	AATAAA	5	+	XM_006534520.2	Sp2	Sp2 transcription factor, transcript variant X3
chr5	8069175	9.224	AATAAA	12	+	NM_009274.2	Srpk2	serine/arginine-rich protein specific kinase 2
chr5	8067773	3.542	TATAAA	2	+	NM_009274.2	Srpk2	serine/arginine-rich protein specific kinase 2
chr5	8067865	8.007	TATAAA	6	+	NM_009274.2	Srpk2	serine/arginine-rich protein specific kinase 2
chr3	88392340	8.976	AATAAA	316	+	XR_375536.2	Stpg2	sperm tail PG rich repeat containing 2, transcript variant X1
chr3	88888252	6.74	AATAAA	7	-	XM_006501628.1	Stpg2	sperm tail PG rich repeat containing 2, transcript variant X3
chr7	27992192	5.12	AATAAA	14	-	NM_009302.3	Swap70	SWA-70 protein
chr7	27992068	7.603	AATACA	1	+	NM_009302.3	Swap70	SWA-70 protein
chr19	32023062	8.732	ATTAAA	12	+	NM_145952.3	Tbc1d12	TBC1D12: TBC1 domain family, member 12
chr19	32022878	3.61	CATAAA	5	+	NM_145952.3	Tbc1d12	TBC1D12: TBC1 domain family, member 12
chr4	43579800	6.99	AATAAA	14	-	XM_006538046.2	Tbc1d2	TBC1 domain family, member 2, transcript variant X1
chr4	43579931	8.811	AATAAA	34	+	XM_006538046.2	Tbc1d2	TBC1 domain family, member 2, transcript variant X1

Table 14. Liver tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr11	62418307	5.24	AATAAA	278	+	NM_019507.2	Tbx21	T-box 21
chr11	62418387	3.647	AATACA	18	-	NM_019507.2	Tbx21	T-box 21
chr12	60128971	3.81	AAAAAG	1	+	XM_006516305.2	Tc2n	tandem C2 domains, nuclear, transcript variant X6
chr12	60128929	5.119	AATATA	7	+	XM_006516305.2	Tc2n	tandem C2 domains, nuclear, transcript variant X6
chr12	60144326	5.535	ATTAAA	5	-	XM_011244191.1	Tc2n	tandem C2 domains, nuclear, transcript variant X3
chr2	25322231	6.731	AATAAA	46	-	XM_006499150.1	Tfpi	tissue factor pathway inhibitor, transcript variant X6
chr2	25354175	10.17	ATTAAA	325	-	XM_006499145.1	Tfpi	tissue factor pathway inhibitor, transcript variant X1
chr4	53873462	7.806	AATAAA	52	+	XM_011250037.1	Tmem245	transmembrane protein 245, transcript variant X1
chr4	53874889	4.668	ATTAAA	8	+	XM_011250037.1	Tmem245	transmembrane protein 245, transcript variant X1
chr4	11101523	3.713	AATAAA	15	+	XR_390735.1	Tmem82	transmembrane protein 82, transcript variant X1
chr4	11102358	4.099	AATAAA	10	+	XR_390735.1	Tmem82	transmembrane protein 82, transcript variant X1
chr4	43040893	5.503	AATAAA	20	-	XM_006537772.2	Tmod1	tropomodulin 1, transcript variant X1
chr4	43066124	5.093	TATAAA	1	+	NM_021883.2	Tmod1	tropomodulin 1
chr2	62169689	4.691	AGTAAA	11	-	XM_011239584.1	Trp53bp1	transformation related protein 53 binding protein 1, transcript variant X6
chr2	62169740	3.634	ATTAAA	7	-	XM_011239584.1	Trp53bp1	transformation related protein 53 binding protein 1, transcript variant X6
chr7	19833213	6.727	AATAAA	86	+	NM_001109897.2	Trpc2	transient receptor potential cation channel, subfamily C, member 2
chr7	19842280	5.183	AATAAA	14	-	NM_001109897.2	Trpc2	transient receptor potential cation channel, subfamily C, member 2
chr1	44059182	3.374	ATTAAA	20	+	XM_011238547.1	Unc80	unc-80 homolog (C. elegans), transcript variant X9
chr1	44160740	8.761	CATAAA	8	-	XM_011238547.1	Unc80	unc-80 homolog (C. elegans), transcript variant X9

Table 14. Liver tissue annotation from mouse identifying transcript isoforms

Chr	PolyA-site	PolyA-score	Testis PAS	Reads	Strand	Accession Number	Short Gene Name	Full Gene Name
chr17	6040130	5.244	AATAAA	5	+	NM_001104564.1	Vmn2r102	vomeronasal 2, receptor 102
chr17	6039167	6.767	TATAAA	10	+	NM_001104564.1	Vmn2r102	vomeronasal 2, receptor 102
chr10	21088585	6.742	AATAAA	18	+	NM_175936.1	Vmn2r81	vomeronasal 2, receptor 81
chr10	21088688	5.255	AATAAA	50	+	NM_175936.1	Vmn2r81	vomeronasal 2, receptor 81
chr11	23389741	6.001	AATAAA	12	+	XM_006514801.2	Vrk2	vaccinia related kinase 2, transcript variant X1
chr11	23389742	6.637	AATAAA	3	-	XM_006514801.2	Vrk2	vaccinia related kinase 2, transcript variant X1
chr11	23389801	9.622	TATAAA	6	+	XM_006514801.2	Vrk2	vaccinia related kinase 2, transcript variant X1
chr2	91563825	4.125	AATAAA	57	+	NM_054068.2	Vsx1	visual system homeobox 1 homolog (zebrafish)
chr2	91562640	3.845	ATTAAA	1	+	NM_054068.2	Vsx1	visual system homeobox 1 homolog (zebrafish)
chr2	91564870	8.66	ATTAAA	75	+	NM_054068.2	Vsx1	visual system homeobox 1 homolog (zebrafish)
chr4	8070517	4.724	AATAAA	6	-	XM_006539245.2	Vwa5b1	von Willebrand factor A domain containing 5B1, transcript variant X6
chr4	8072409	5.393	AATAAA	10	-	XM_006539245.2	Vwa5b1	von Willebrand factor A domain containing 5B1, transcript variant X6
chr12	74644394	7.688	AATAAA	12	-	XR_381520.2	Wdr60	WD repeat domain 60, transcript variant X2
chr12	74645033	8.417	AATAAA	3	-	XR_381520.2	Wdr60	WD repeat domain 60, transcript variant X2
chr12	74644786	7.579	GATAAA	7	-	XR_381520.2	Wdr60	WD repeat domain 60, transcript variant X2
chr17	32550043	5.417	ATTAAA	13	-	NM_028198.2	Xpo5	exportin 5
chr17	32553462	6.879	ATTAAA	32	-	NM_028198.2	Xpo5	exportin 5
chr2	60090049	8.037	AATAAA	3138	-	NM_028054.3	Zfyve19	zinc finger, FYVE domain containing 19, transcript variant 1
chr2	60090135	7.184	AATAAA	9	-	NM_028054.3	Zfyve19	zinc finger, FYVE domain containing 19, transcript variant 1
chr2	60090259	7.647	TATAAA	23	-	NM_028054.3	Zfyve19	zinc finger, FYVE domain containing 19, transcript variant 1

Table 14. Liver tissue annotation from mouse identifying transcript isoforms

Category	Number of Genes	Category Hit Against Total Genes	Subcategories
cellular process	14	16.66666667	cell communication (9), cell cycle (4), cellular component movement (2), cytokinesis (2)
metabolic process	14	16.66666667	biosyntheric process (2), catabolic process (1), generation of precursor metabolites and energy (1), nitrogen compound process (3), phosphate-containing compound metabolic process (2), primary metabolic process (10)
localization	12	14.28571429	RNA localization (1), transport
biological regulation	9	10.71428571	regulation of biological processes (3), regulation of molecular function (7)
developmental process	7	8.333333333	anatomical structure morphogenesis (3), cell differentiation (1), death (1), ectoderm development (1), mesoderm development (4),
response to stimulus	7	8.333333333	cellular defense response (1), immune response (2), response to biotic stimulus (1), response to external stimulus (2), response to
cellular component organization or biogenesis	6	7.142857143	cellular component biogenesis (1), cellular component organization (6)
immune system process	6	7.142857143	immune response (2)
multicelluar organismal process	6	7.142857143	single-multicellular organism process (1)
apoptotic process	1	1.19047619	[no hit for lower level categories]
biological adhesion	1	1.19047619	cell adhesion (1)
reproduction	1	1.19047619	gamete generation (1)

Table 15. Gene ontology for biological processes in mouse annotated transcripts

Gene ontology (GO) was performed on comprehensive mouse annotated transcripts using PANTHER Classification System to identify represented biological processes.

Gene Short Name	Full Gene Name	Cell Specificity	Base Pairs (bps)	Round Spermatids	Spermatocytes
C-Kit	kit kit oncogene	spermatogonia	145	-	-
GFRα-1	glial cell line derived neurotropic factor family receptor alpha 1	spermatogonia	608	-	-
P450scc	cytochrome P450, family 11, subfamily A, polypeptide 1	Leydig	169	-	-
SYCP3	synaptonemal complex protein 3	Spermatocyte	469	-	+
Dbil5	diazepam binding inhibitor-like 5	round spermatid	121	+	-
PRM2	protamine 2	round spermatid	276	+	-
FSHR	follicle stimulating hormone receptor	Sertoli cell	356	-	-

Table 16. Isolated male germ cell purity validated with germ cell-specific primers

"+" = gene amplification, indicated by banding on gels, following RT-PCR

"-" = no gene amplification

Gene Short Name	Accession Number	bps	Forward Primer 5' --> 3'	Reverse Primer 5' --> 3'
C-Kit	X65998.1	145	CGCTCCAA-GAATGTAAGTGGC	CCCTGCTTTCGGTTAC-CACA
GFRα-1	NM_010279.2	608	GCGACAGAC-TATCGTCCCTG	CACCAGCGAGAC-CATCCTTT
P450scc	NM_019779.3	169	GGTTCCACTCCTCAAA-GCCA	AAAGAAGCCCG-GATCTCGAC
SYCP3	NM_011517.2	469	CCTCAGATGCTTCGAGG GTG	CCCACTGCTG-CAACACATTC
Dbil5	NM_021294.2	121	CCCAGGGCGACTG-TAACATC	GCAATGTAGATCCTCATG GCAT
PRM2	NM_008933.1	276	ACAAGACCATGAAC-GCGAGG	GAGGCTTAG-TGATGGTGCCT
FSHR	NM_013523.3	356	TCATT-GAGGCCAGCCTTACC	CACTGTGGTGTTCCCAG-TGA

Table 17. Germ cell-specific primers used in PCR validation

	Input	Output	Percent Yield
Mouse Testis RNA	5µg	15.71ng	0.314%
Male Germ Cell RNA	2.5-5µg	0.8-2ng	0.04%

Table 18. Production of polyA+ mRNA

Yield of polyA+ mRNA from total RNA of male germ cell populations using the Oligotex mRNA Mini kit.

BIBLIOGRAPHY

Agarwal A, Makker K, Sharma R. Clinical relevance of oxidative stress in male-factor infertility: An update. Am J Repro Immunol 2008; 59: 2-11.

Balhorn R. A model for the structure of chromatin in mammalian sperm. Jour of Cell Bio 1982; 93: 298-305.

Bartel DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell 2004; 116: 281-297.

Bastos H, Lassalle B, Chicheportiche A, Riou L, Testart J, Allemand I, Fouchet P. Flow cytometric characterization of viable meiotic and postmeiotic cells by Hoechst 33342 in mouse spermatogenesis. ISAC 2005; 65A: 40-49.

Bellve AR, Cavicchia JC, Millette CF, O'Brien DA, Bhatnagar YM, Dym M. Spermatogenic cells of prepuberal mouse isolation and morphological characterization. JCB 1977; 74: 68-85.

Braun RE. Packaging paternal chromosomes with protamine. Nature 2001; 8: 10-12.

Bryant JM, Meyer-Ficca ML, Dang VM, Berger SL, Meyer RG. Separation of spermatogenic cell types using STA-PUT velocity sedimentation. J Vis Exp 2013; 80: 1-9.

Carrell DT, Emery BR, Hammoud S. Altered protamine expression and diminished spermatogenesis: what is the link? Hum Reprod Update 2007; 13: 313-327.

Carrell DT, Liu L. Altered protamine 2 expression is uncommon in donors of known fertility, but common among men with poor fertilizing capacity, and may reflect other abnormalities of spermiogenesis. J Androl 2001; 22: 604-610.

Chang YF, Lee-Chang JS, Panneerdoss S, MacLean JA, Rao MK. Isolation of Sertoli,

Leydig, and spermatogenic cells from the mouse testis. BioTechniques 51:341-344.

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 2009; 7:e1000112.

Chomczynski P. A reagent for the single-step simultaneous isolation of RNA, DNA and proteins from cell and tissue samples. BioTechniques 1993; 15: 532-537.

Darmon SK, Lutz CS. Novel upstream and downstream sequence elements contribute to polyadenylation efficiency. RNA Biol 2012; 10: 1255-1265.

Dass B, Tardif S, Park JY, Tian B, Weitlauf HM, Hess RA, Carnes K, Griswold MD, Small CL, MacDonald CC. Loss of polyadenylation protein τCstF-64 causes spermatogenic defects and male infertility. Proc of the National Acad of Sci 2007; 104: 20374-20379.

Derti A, Garrett-Engele P, MacIsaas KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. A quantitative atlas of polyadenylation in five mammals. Genome Res 2012; 22: 1173-1183.

Di Giammartino DC, Nishida K, Manley JL. Mechanisms and consequences of alternative polyadenylation. Mol Cell 2011; 43: 853-866.

Eddy EM. Regulation of gene expression during spermatogenesis. Semin Cell Dev Biol 1998; 9: 451-457.

Eddy EM. Male germ cell gene expression. Recent Prog Horm Res 2002; 57: 103-128.

Elkon R, Ugalde AP, Agami R. Alternative cleavage and polyadenylation: extent, regulation and function. Nat Rev Genet 2013; 14: 496-506.

Ferlin A, Raicu F, Gatta V, Zuccarello D, Palka G, Foresta C. Male infertility: role of genetic background. Reproductive BioMedicine Online 2007; 14: 734-745.

Fox-Walsh K, Davis-Turak J, Zhou Y, Li H, Fu XD. A multiplex RNA-seq strategy to profile poly(A+)RNA: Application to analysis of transcription response and 3'end formation. Genomics; 98:266-271.

Garrido N, Martínez-Conejero JA, Jauregui J, Horcajadas JA, Simón C, Remohí J, Meseguer M. Microarray analysis in sperm from fertile and infertile men without basic sperm analysis abnormalities reveals a significantly different transcriptome. Amer Society for Repro Med 2009; 91: 1307-1310.

Getun IV, Torres B, Bois PRJ (2011): Flow cytometry purification of mouse meiotic cells. JoVE 2011; 50: 1-3.

Gilbert SF. Developmental Biology. 6th ed. Sunderland, MA: Sinauer Associates; 2000. Spermatogenesis. Available from: http://www.ncbi.nlm.nih.gov/books/NBK10095/.

Gupta I, Clauder-Münster S, Klaus B, Järvelin AI, Aiyar RS, Benes V, Wilkening S, Huber W, Pelechano V, Steinmetz LM. Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions. Mol Syst Biol 2014; 10(2): 719.

Hafez D, Ni T, Mukherjee S, Zhu J, Ohler U. Genome-wide identification and predictive modeling of tissue-specific alternative polyadenylation. ISMB/ECCB 2013; 29: i108-i116.

Hogue M, Li W, Tian B. Accurate mapping of cleavage and polyadenylation sites by 3'region extraction and deep sequencing. Methods Mol Biol 2014; 1125:119-129.

Hoyert DL, Xu J. Deaths: Preliminary Data for 2011. NVSS 2012; 61: 1-52.

Jamsai D, O'Bryan MK. Mouse models in male fertility research. Asian Jour of Andrology 2011; 13: 139-151.

Ji A, Luo W, Li W, Hoque M, Pan Z, Zhao Y, Tian B. Transcriptional activity regulates alternative cleavage and polyadenylation. Molecular Systems Biology 2011; 7:534-546.

Kleene KC. Connecting cis-elements and trans-factors with mechanisms of development regulation of mRNA translation in meiotic and haploid mammalian spermatogenic cells. Reproduction 2013; 146(1): 1-19.

Lai DP, Tan S, Kang YN, Wu J, Ooi HS, Chen J, Shen TT, Qi Y, Zhang X, Guo Y, Zhu T, Liu B, Shao Z, Zhao X. Genome-wide profiling of polyadenylation sites reveals a link between selective polyadenylation and cancer metastasis. Hum Mol Genet 2015; 24(9): 1-8.

Lee LK, Foo KY. Recent insights on the significance of transcriptomic and metabolic analysis of male factor infertilty. Clin Biochem 2014; 47(2014): 973-982.

Lee K, Haugen HS, Clegg CH, Braun RE. Premature translation of protamine 1 mRNA causes precocious nuclear condensation and arrests spermatid differentiation in mice. Proc Natl Acad Sci 1995; 92: 12451-12455.

Legendre M, Ritchie W, Lopez F, Gautheret D. Differential repression of alternative transcripts: A screen for miRNA targets. PLOS 2006; 2: 1-10.

Lewis SEM. Is sperm evaluation useful in predicting human fertility? Jour of the Soc for Reprod and Fertility 2007; 134: 31-40.

Lodish H, Berk A, Zipursky SL, et al. Molecular Cell Biology. 4th edition. New York: W.H. Freeman; 2000.

Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. Genes & Development 2013; 27: 1-16.

Lin Y, Li Z, Ozsolak F, Kim SW, Arango-Argoty G, Liu TT, Tenenbaum SA, Bailey T, Monaghan AP, Milos PM, John B. An in-depth map of polyadenylation sites in cancer. Nuc Acids Res 2012; 40: 8460-8469.

Liu D, Brockman M, Dass B, Hutchins LN, Singh P, McCarrey JR, MacDonald CC, Graber JH. Systematic variation in mRNA 3'-processing signals during mouse spermatogenesis. Nuc Acids Research 2007; 35: 234-246.

Liu Y, Niu M, Yao C, Hai Y, Yuan Q, Liu Y, Guo Y, Li Z, He Z. Fractionation of human spermatogenic cells using STA-PUT gravity sedimentation and their miRNA profiling. Sci Rep 2015; Jan 30;5:8084. doi: 10.1038/srep08084.

Lu J, Bushel PR. Dynamic expression of 3'UTRs revealed by Poisson hidden Markov modeling of RNA-Seq: implications in gene expression profiling. Gene 2013; 527(2): 616-623.

Mandel CR, Bai Y, Tong L. Protein factors in pre-mRNA 3'-end processing. Cell Mol Life Sci 2008; 65: 1099-122.

Martin KC, Casadio A, Zhu H, E Y, Rose JC, Chen M, Bailey CH, Kandel ER.

Synapse-specific, long-term facilitation of aplysia sensory to motor synapses: a function for local protein synthesis in memory storage. Cell 1997; 91(7): 927-938.

Mayr C, Bartel DP. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 2009; 138: 673-684.

Matzuk MM, Lamb DJ. The biology of infertility: research advances and clinical challenges. Nature 2008; 14: 1197-1210.

McGarvey KM, Goldfarb T, Cox E, Farrell CM, Gupta T, Joardar vS, Kodali VK, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Webb D, Wright MW, Murphy TD, Pruitt KD. Mouse genome annotation by the RefSeq project. Mamm Genome 2015; 26(9-10): 379-390.

McMahon KW, Hirsch BA, MacDonald CC. Differences in polyadenylation site choice between somatic and male germ cells. BMC Molecular Biology 2006; 7: 1471-2199.

Mitra M, Johnson EL, Coller HA. Alternative polyadenylation can regulate post-translational membrane localization. Trends Cell Mol Biol 2015; 10: 37-47.

Ni T, Yang Y, Hafez D, Yang W, Klesewetter K, Wakabayashi Y, Ohler U, Peng W and Zhu J. Distinct polyadenylation landscapes of diverse human tissues revealed by modified PA-seq strategy. BMC Genomics 2013; 14:615.

Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell 2010; 143:1018-1029.

Patel ZP, Niederberger CS. Male factor assessment in infertility. Med Clin of North

Amer 2011; 95: 223-234.

Pang ALY, Johnson W, Ravindrananth N, Dym M, Rennert OM, Chan WY. Expression profiling of purified male germ cells: stage-specific expression patterns related to meiosis and postmeiotic development. Physiol Genomics 2006; 24: 75-85.

Rosner B. Fundamentals of biostatistics, third edition. Duxbury Press; 1990.

Roy A, Lin Y, Matzuk MM. Genetics of idiopathic male infertility. The Genetics of Male Infertility 2007; 99-111.

Saner-Amigh KJ, Halvorson LM. Andrology and fertility assessment. LabMedicine 2011; 42: 41-50.

Sanger PL. Pathways to pregnancy and parturition. Washington: Current Conceptions; 2003.

Sartini BL, Wang H, Wang W, Millette CF, Kilpatrick DL. Pre-messenger RNA cleavage factor I (CFm): Potential role in alternative polyadenylation during spermatogenesis. Biology of Reproduction 2008; 78: 472-482.

Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA 2011; 17:761-772.

Shi Y. Alternative polyadenylation: New insights from global analyses. RNA Jour 2012; 18: 2105-2117.

Sun J, Zhong L, Zhu Y, Liu G. Research on the isolation of mouse Leydig cells using differential digestion with a low concentration of collagenase. J Reprod Dev 2011; 57: 433–436.

Sun Y, Fu Y, Li Y, Xu A. Genome-wide alternative polyadenylation in animals: insights from high throughput technologies. Jour of Mol Cell Bio 2012; 0:1-10.

Tian B, Hu J, Zhang H, Lutz CS. A large-scale analysis of mRNA polyadenylation of humans and mouse genes. Nucleic Acids Res 2005; 33:201-212.

Tseden K, Topaloglu O, Meinhardt A, Dev A, Adham I, Müller C, Wolf S, Böhm D, Schlüter G, Engel W, Navernia K. Premature translation of transition protein 2 mRNA causes sperm abnormalities and male infertility. Mol Reprod Dev 2007; 3:273-279.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoforms regulation in human tissue transcriptomes. Nature 2008; 456:470-476.

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. PMC 2009; 10:57-63.

White-Cooper H, Davidson I. Unique aspects of transcription regulation in male germ cells. CSH Perspect 2011; 1-17. Doi:10.1101/cshperspect.a002626.

Yang C-K, Yen P. Differential Translation of Dazap1 Transcripts during Spermatogenesis. Yan W, ed. PLoS ONE 2013; 8(4):e60873. doi:10.1371/journal.pone.0060873.

Zhou Y, Li HR, Huang J, Jin G, Fu XD. Multiplex analysis of PolyA-linked sequences (MAPS): an RNA-Seq strategy to profile Poly(A+) RNA. Methods Mol Biol 2014; 1125:169-178.