REPRODUCTION
RESEARCH

An expanded mouse testis transcriptome and mass spectrometry defines novel proteins

Jaya Gamble[1], Joel Chick[2], Kelly Seltzer[1], Joel H Graber[3], Steven Gygi[2], Robert E Braun[4] and Elizabeth M Snyder[1]

[1]*Department of Animal Sciences, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA,* [2]*Department of Cell Biology, Harvard Medical School, Boston, Massachusetts, USA,* [3]*MDI Biological Laboratory, Bar Harbor, Maine, USA and* [4]*The Jackson Laboratory, Bar Harbor, Maine, USA*

Correspondence should be addressed to E Snyder; Email: elizabeth.snyder@rutgers.edu

Abstract

The testis transcriptome is exceptionally complex. Despite its complexity, previous testis transcriptome analyses relied on a reductive method for transcript identification, thus underestimating transcriptome complexity. We describe here a more complete testis transcriptome generated by combining Tuxedo, a reductive method, and spliced-RUM, a combinatorial transcript-building approach. Forty-two percent of the expanded testis transcriptome is composed of unannotated RNAs with novel isoforms of known genes and novel genes constituting 78 and 9.8% of the newly discovered transcripts, respectively. Across tissues, novel transcripts were predominantly expressed in the testis with the exception of novel isoforms which were also highly expressed in the adult ovary. Within the testis, novel isoform expression was distributed equally across all cell types while novel genes were predominantly expressed in meiotic and post-meiotic germ cells. The majority of novel isoforms retained their protein-coding potential while most novel genes had low protein-coding potential. However, a subset of novel genes had protein-coding potentials equivalent to known protein-coding genes. Shotgun mass spectrometry of round spermatid total protein identified unique peptides from four novel genes along with seven annotated non-coding RNAs. These analyses demonstrate the testis expresses a wide range of novel transcripts that give rise to novel proteins.

Reproduction (2020) **159** 15–26

Introduction

Spermatogenesis involves a myriad of biological processes, many of which are unique to germ cells (Russell *et al.* 1990) and occur during specific phases during development. As a result, the adult mammalian testis contains a wide range of germ cells undergoing mitotic, meiotic, and post-meiotic events. Given this level of cellular complexity and the need for specialized biological events like meiosis, it is unsurprising that even prior to the genomics era, unusual testis-specific transcripts were commonly reported (Kleene 2001). More recently, high-throughput RNA sequencing (RNA-seq) has proven to be a powerful platform for defining transcriptomes (Sultan *et al.* 2008, Wang *et al.* 2008) in a range of cell types (Djebali *et al.* 2012) and tissues (Brawand *et al.* 2011). Several such studies have greatly expanded our understanding of the testis transcriptome and confirmed its previously theorized complexity (Ramskold *et al.* 2009, Merkin *et al.* 2012). Compared with other tissues, the testis expresses a higher number of protein-coding genes (Djureinovic *et al.* 2014) and its transcriptome is less likely than other tissues to be dominated by a few, highly expressed genes (Ramskold *et al.* 2009).

Although previous reports have demonstrated the breadth of testis expressed transcripts, the majority of analyses have focused only on known transcripts or relied on a reductive method, Tuxedo (Trapnell *et al.* 2012), for novel transcript identification. Tuxedo uses a spliced alignment method based on expressed sequence tag (EST) assemblers, in which transcripts are defined via a minimum path coverage method (Trapnell *et al.* 2010). Given the propensity of germ cells to express a large number of alternatively spliced transcripts (Soumillon *et al.* 2013), reductive pipelines underestimate the complexity of the testis transcriptome. In spite of this shortcoming, the reported testis transcriptome is already quite complex. This characteristic is attributed primarily to the germ cells, which are reported to express a large number of both non-coding and intergenic regions (Soumillon *et al.* 2013).

The unique biology of germ cell differentiation suggests germ cells may be reliant on a very wide range of proteins, a conclusion supported by the observation that male germ cells express a large fraction of known protein coding genes and protein expression is often highly cell-type dependent (Djureinovic *et al.* 2014). In particular, meiotic and post-meiotic germ cells appear to

ISSN 1470–1626 (paper) 1741–7899 (online)
https://doi.org/10.1530/REP-19-0092
Online version via https://rep.bioscientifica.com

express an extremely wide range of alternatively spliced transcripts that generate proteins (Naro *et al.* 2017). As such, the male germ cell is an excellent model in which to identify novel proteins. Unfortunately, a majority of proteome studies in the testis to date have utilized either immunohistochemical detection (Djureinovic *et al.* 2014) or mass spectrometry (Guo *et al.* 2008, Wang *et al.* 2014), both of which rely on annotation of known protein-coding genes. As such, they fail to detect novel proteins. A notable exception to this is the report by Chocu *et al.* (2014) which identified a number of novel protein-coding genes using a novel rat testis transcriptome (reviewed in Com *et al.* 2014). Unfortunately, this analysis relied on the reductive Tuxedo method and so likely underestimated the potential for novel protein expression in the testis.

With the goal of identifying a broader range of novel testis proteins, we generated a extended testis transcriptome by combining reductive and combinatorial transcript-building philosophies to a large testis RNA-seq dataset. The expression profile of the identified novel transcripts was determined across tissues and within the testis. Lastly, protein-coding potential was assessed and mass spectrometry of isolated post-meiotic germ cell protein used to analyze novel protein expression in male germ cells.

Materials and methods

Animal care and sample collection

All experimental mice used in this study were cared for in accordance with the 'Guide for the Care and Use of Experimental Animals' established by the National Institutes of Health (NIH) and all protocols approved by the Jackson Laboratory Animal Care and Use Committee. Late juvenile male mice (25 days post-partum) of a mixed C57BL6/J-129S1/SvImJ genetic background ($n=3$) were killed and whole testes was collected. Samples were stored in RNAlater® (Life Technologies) at −20°C until extraction.

RNA sequencing

Paired-end RNA sequencing was performed on an Illumina HiSeq 2000 at The Jackson Laboratory. Total RNA extraction via the mirVana RNA isolation kit (Life Technologies) was performed per manufacturer's instructions. RNA sequencing libraries for 100 bp paired-end sequencing were produced using the TruSeq RNA Sample prep Kit v2 Set A and B (Illumina). Extended materials and methods regarding RNA-sequencing analyses can be found in the Supplementary materials (see section on supplementary materials given at the end of this article).

Tissue and testicular cell-specific expression

Adult mouse 100 bp paired end RNA-seq read data were obtained from ENCODE (www.encodeproject.org) for the following samples and accessions: brain technical replicate 1 (ENCFF286WTQ and ENCFF358NPU), brain technical replicate 2 (ENCFF445AWP and ENCFF958CHE), heart technical replicate 1 (ENCFF871XHK and ENCFF952JKH), heart technical replicate 2 (ENCFF104UFH and ENCFF126WYO), testis technical replicate 1 (ENCFF517RDO and ENCFF786ZKB), testis technical replicate 2 (ENCFF682XSC and ENCFF690HKC), liver technical replicate 1 (ENCFF492PRP and ENCFF581OEV), liver technical replicate 2 (ENCFF161LEK and ENCFF516HOO), and ovary (ENCFF312OKA and ENCFF463WEH). Single-end 76 bp RNA-seq strand-specific reads derived from isolated testicular cell types were obtained from the SRA database (GEO accession numbers GSE43717, GSE43719, and GSE43721 (Soumillon *et al.* 2013)). Individual samples were aligned to the expanded transcriptome and expression estimated via RSEM. For tissue and cell expression, transcripts per million (TPM) was calculated as the average of available technical replicates.

Molecular confirmation of novel transcripts

Novel transcripts for confirmation were selected based on expression and relative abundance (for novel isoforms). Selected transcripts were confirmed via RT PCR using total RNA from C57BL/6J adult whole testis extracted via TRIzol® (Invitrogen) and reverse transcribed using SuperScript® III RT (Life Technology). All cDNA templates were assessed using RT-PCR with *Rps2* as a template quality control (Fig. 2C for example) and all reactions were run in full for each gel. Primers spanning relevant junctions or open reading frames were used to amplify element-specific products (Supplementary Table 1) and sequence confirmed via Sanger sequencing of either PCR amplicons or TA cloned via the TOPO® TA Cloning Kit (Life Technology) amplicons. The resulting sequence was aligned to the GRCm38 genome via BLAT for confirmation.

Fluorescent activated cell sorting (FACs) and protein isolation

FACs was performed on adult whole testis as in Gaysinskaya *et al.* (2014) with the following modifications: 5 µg of Hoechst 33342 was used per 6 mL cell suspension and was allowed to proceed for 30 min at 37°C. Preliminary analyses included quantification of cell purity by DAPI staining and morphology analysis. Approximately 3.5 million isolated round spermatids were solubilized in RIPA buffer with protease inhibitor cocktail (Sigma-Aldrich), quantified by DC Protein Assay (BioRad), and 2 µg of total protein processed for mass spectrometry (Supplementary materials). Following MS identification, targets were selected for further orthogonal confirmation (Supplementary materials).

Results

Two transcript-building tools utilizing different strategies were applied to RNA-seq data of 25 days post-partum (dpp) whole testis to identify potential protein-coding transcripts (Supplementary Fig. 1). This time point was

selected in order to capture only actively transcribed mRNAs. In the background selected, 25 dpp testis contain spermatids up to step 10, the point at which transcription halts (Namekawa *et al.* 2006). Pipeline-specific transcriptomes were assessed and compared and the resulting expanded testis transcriptome analyzed for possible novel protein-coding transcripts.

The Tuxedo-defined testis transcriptome

Of the 87,613 transcripts contained within the known transcriptome (Ensembl 38 release 68), Tuxedo identified 65.6% as expressed in the late juvenile testis (Fig. 1). Tuxedo defined an equally large number of novel transcripts, generally with fewer exons per transcript than the known transcriptome and consistent with the reductionist nature of Tuxedo. Of the more than 53,000 novel Tuxedo-defined transcripts, roughly half were defined as novel genes (not containing any known junctions). Of the novel genes, over 75% contained only a single exon. Reads aligning to single exon transcripts informed on over 29,000 putative transcripts but made up less than 1% of the total aligned reads, suggesting they were not major contributors to testis transcriptome complexity. Additional analysis showed promiscuous alignment of single exon transcript reads and overall low expression of single exon transcripts (Supplementary Fig. 2); thus, they were eliminated from further study.

The resulting Tuxedo-derived transcriptome contained a total of 23,739 novel transcripts, of which approximately 60% were novel isoforms of known genes. The remaining major classes of transcripts were defined as either novel genes (sharing no junctions with known transcripts), intronic (contained entirely within the intron of a known transcript), or antisense (derived from the opposite strand of a known transcript) (Table 1).

The combinatorially defined testis transcriptome

Spliced RUM (SR) is a post-sequencing aligner and transcript-building methodology with a combinatorial building philosophy (see Supplementary materials for details). Application of SR to the same late juvenile whole testis RNA-seq data analyzed by Tuxedo identified a total of 19,386 novel transcripts. Of the novel SR-derived transcripts, the majority (85.3%) were novel isoforms of known genes (Table 1). SR-defined transcripts contained a much larger number of exons per transcript and a greater number of transcripts per gene than either the Tuxedo derived transcriptome or the known expressed transcriptome (Fig. 1). Analysis of individual genes demonstrated this to be a function of SR's combinatorial approach.

The expression of novel SR-derived transcripts was compared to the expression of known transcripts (Supplementary Fig. 2) and the distribution and median expression of SR-derived transcripts found to be similar to that observed for known transcripts, suggesting that SR identified a set of novel transcripts with biologically relevant expression. Thus, the 19,386 novel SR-defined transcripts represent potentially relevant additions to the known testis transcriptome. In order to determine if the SR-defined transcripts represented a unique addition to the previously defined Tuxedo-derived transcriptome, the two were compared.

An expanded testis transcriptome

Overlap between the Tuxedo and SR pipelines was assessed (Fig. 1) and found to be limited. While the different transcript-building approaches (reductionist versus combinatorial) played a part in the differences, additional approach-specific biases were also observed including preference for alternative exon usage (Tuxedo) over exon skipping (SR). Given their different but complementary natures, the union of both Tuxedo- and SR-defined transcriptomes along with the entirety of known transcripts was used for downstream analyses. In addition to the 87,610 transcripts contained within the Ensembl dataset, this expanded testis transcriptome included an additional 41,293 transcripts (Table 1). A set of eight novel isoforms derived from seven genes along with eight novel genes were selected for confirmation (Supplementary Tables 2 and 3). Of these, 14 (87.5%) were detectible by PCR and confirmed via Sanger sequencing.

Confirmation of combined pipeline efficacy by reannotation

As a broader measure of the expanded transcriptome accuracy, it was reannotated using a more recent Ensembl gene annotation (December 2017, release 91) (Table 1) that included a large mouse gene update. This comparison showed 49% of the previously identified novel genes had been independently identified and validated in the updated mouse gene annotation demonstrating the effectiveness of the combined transcript building approach to identify new transcripts. Comparison across annotations found that the increase in novel isoforms was due to the reassignment of many novel genes, intronic, and antisense transcripts to novel isoforms of known and newly annotated genes, further demonstrating the utility of the transcript building pipeline for identification of bona fide transcripts. Based on final annotation derived from the Ensembl gene release 91, each transcript was assigned a transcript class and unique identifier (*Txt###* for Testis expressed transcript) for use prior to official annotation. These classes and identifiers were then used for the remainder of analyses.

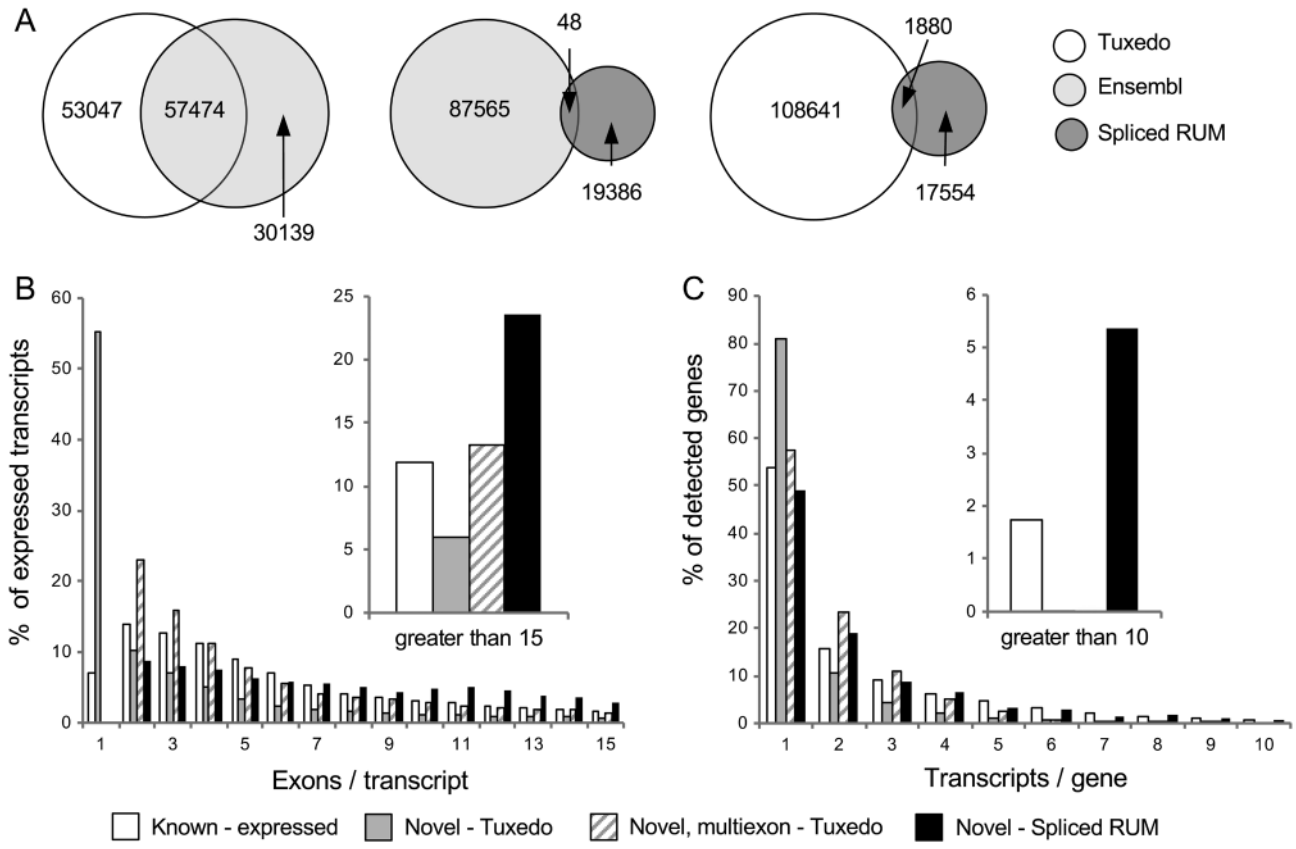

Figure 1 Tuxedo and Spliced RUM identify different sets of novel transcripts not present in Ensembl. (A) Overlap of transcripts contained within the known transcriptome (Ensembl GRCm38.75) or identified by Tuxedo and/or SR as testis expressed. (B) Exons per transcript by transcriptome. (C) Transcripts per gene by transcriptome. Expressed – TPM > 0, calculated by RSEM across all biological replicates ($n = 3$).

Expression profile of the expanded testis transcriptome across tissues and within the testis

The unusually large number of novel transcripts in the expanded testis transcriptome suggested they may have unique or limited expression profiles. To determine if this was the case, publicly available raw RNA-sequencing data of adult mouse tissues (www.encodeproject.org/) was aligned to the expanded transcriptome and novel transcript expression quantified (Fig. 2A). Testis-discovered novel transcripts were highly enriched in the testis relative to other adult tissues with the exception of novel isoforms, which were also enriched in the adult ovary relative to other tissues. These findings were further confirmed and expanded upon by reverse transcriptase PCR (RT-PCR) of selected novel isoforms and novel genes (Fig. 2B and C). In multiple cases, additional products were detected in non-testis tissues. These products likely represent tissue-specific novel isoforms that went undetected in our testis-centric transcriptome and further support the argument for defining tissue-specific transcriptomes.

To define the expression profile of the expanded testis-transcriptome within the testis, raw RNA-sequencing data from isolated testicular cell types (Soumillon *et al.* 2013) was obtained and transcript expression examined (Fig. 3). Cell types analyzed were somatic cells (Sertoli cells), the mitotic germ cell population (undifferentiated and differentiating spermatogonia), the meiotic germ cell population (pachytene spermatocytes), and two populations of post-meiotic germ cells (round spermatids and vas deferens isolated spermatozoa). As

Table 1 Annotation and reannotation of novel testis-detected transcripts.

	Ensembl 38.68				Ensembl 38.90			
	Tuxedo	Spliced RUM	Both	Total	Tuxedo	Spliced RUM	Both	Total
Novel gene	6093	1069	673	7835	3548	445	1	3994
Novel isoform	13455	15630	898	29983	15277	17225	974	33476
Intronic	842	191	140	1173	830	132	116	1078
Antisense	1517	664	121	2302	1482	407	98	1987
				41293				40535

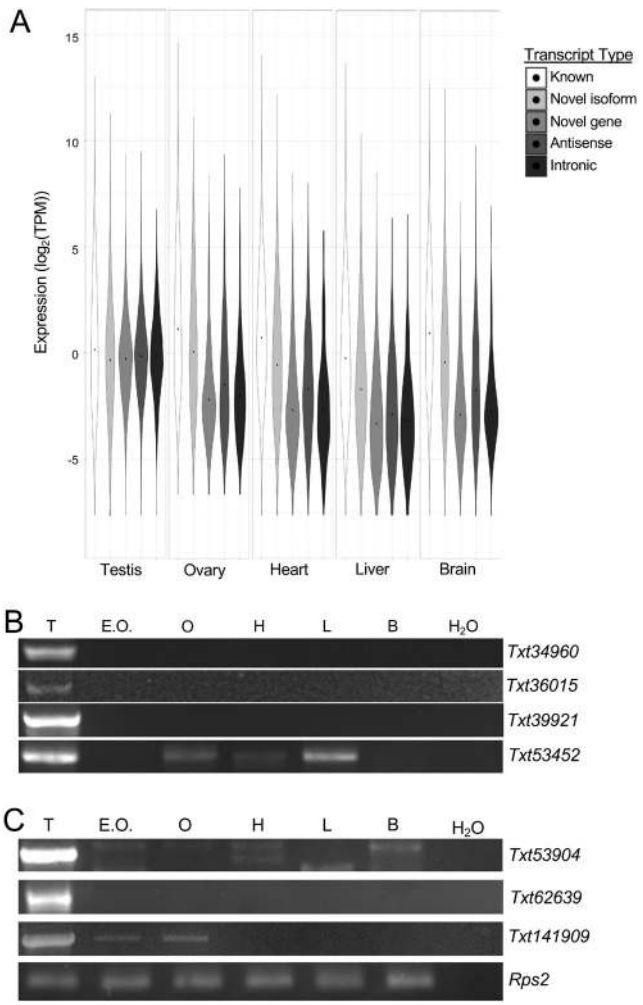

Figure 2 Novel transcripts display testis-enriched or -specific expression. (A) Expression of the five predominant transcript types in the expanded testis transcriptome across select adult tissues by RNA sequencing. Median expression indicated by black dot. Reverse transcriptase PCR of select (B) novel isoforms and (C) genes across tissues. *Rps2* (ribosomal protein S2) shown as a cDNA integrity and loading control. T, testis. E.O., 17.5 days post-coitum (dpc) ovary. O, ovary; H, heart; L, liver; B, brain; H_2O, water template (negative control). Tissues from adult C57Bl6/J unless indicated. $n = 3$. Representative gels shown.

previously reported (Soumillon *et al.* 2013), the majority of novel transcript expression in the transcriptome defined herein was observed in germ cell populations. However, contrary to previous reports, expression of novel genes, intronic, and antisense transcripts was highly enriched in meiotic and post-meiotic germ cells. A similar, but less pronounced, expression profile was observed for novel isoforms.

Protein coding of novel isoforms

As an initial step in identifying novel protein coding transcripts, the protein coding potential of the expanded transcriptome was assessed computationally (Fig. 4). As protein coding potential is correlated to open reading frame (ORF) length, ORF length across each class of transcript was calculated. Additionally, a more robust estimate of protein-coding capacity was calculated for each transcript type by Coding-Potential Assessment Tool (CPAT) (Wang *et al.* 2013). This particular tool was selected because it does not rely on sequence alignment, which may be biased against novel protein-coding transcripts.

Novel isoforms were found to have an almost identical ORF length distribution as that for known protein coding genes and only a slight reduction in calculated protein-coding potential suggesting the majority of novel isoforms are of protein-coding genes, and they retain their protein-coding potential. Ontology analysis of genes with identified novel isoforms demonstrated a large number of genes important to germ cell biology express previously uncharacterized transcripts (Table 2). Additionally, many genes display cell-type-dependent expression of novel isoforms. For example, *Cpeb2* (cytoplasmic polyadenylation element binding protein 2 (Kurihara *et al.* 2003)) expresses two isoforms (Supplementary Fig. 3A). One is highly enriched in spermatocytes while another equally enriched in spermatids.

In order to determine if novel isoforms had significant impacts on encoded proteins, the location of novel exons were mapped to the annotated regions of their parent gene (Supplementary Fig. 3B). Roughly half of the novel exons reside outside of their gene's annotated ORF suggesting they may impact post-transcriptional regulation.

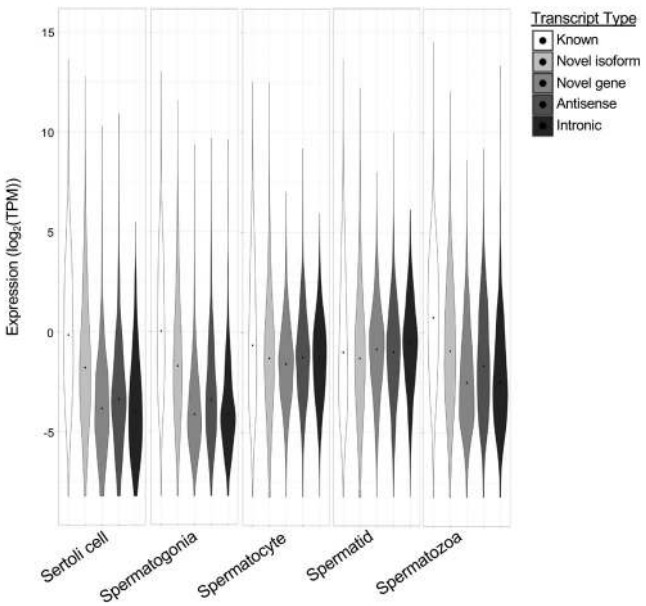

Figure 3 Novel transcripts are predominantly expressed in meiotic and post-meiotic male germ cells. Average expression of the five predominant transcript types in isolated testicular somatic and germ cell populations by RNA-sequencing. $n = 3$. Black dot – median expression.

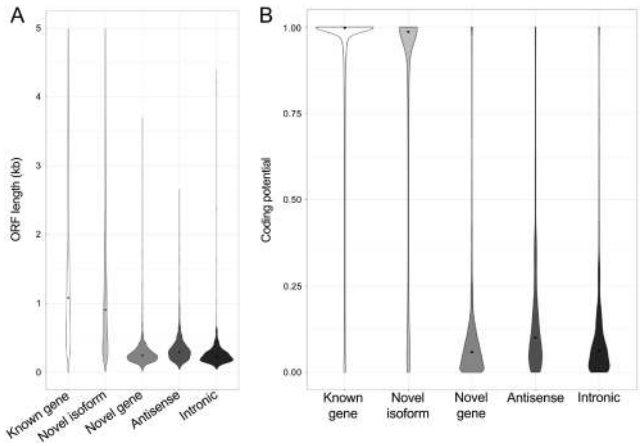

Figure 4 The expanded testis transcriptome contains both coding and non-coding novel transcripts. (A) ORF length per transcript by transcript type (protein coding refers to all known protein coding transcripts detected in the testis). (B) Protein coding potential (0 to 1 scale) as assessed by CPAT.

For example, the two *Cpeb2* isoforms mentioned above differ only in their 3′ UTR. The remaining half of novel exons overlapped with their gene's start codon, ORF, or stop codon. As such, they could potentially impact the isoform's coding potential. These findings correlate well with previous results showing meiotic and post-meiotic germ cells have high rates of intron retention which may generate novel proteins (Naro *et al.* 2017). While at least some of novel exons are likely to have a deleterious effect on protein coding, some novel isoforms may encode new peptides. *Txt4736*, for example, is a novel isoform of *Vash2* (vasohibin-2) that appears to encode an in-frame peptide insertion (Supplementary Fig. 4). VASH2 is a potent regulator of angiogenesis (Xue *et al.* 2013), defects in which are known to cause male infertility (Brennan *et al.* 2003). *Txt4736* contains two previously uncharacterized exons which result in a forty amino acid insertion near the C-terminus. While there are no reported mouse ESTs or mRNAs that include the identified novel exons, multiple human *Vash2* isoforms exist (Shibuya *et al.* 2006) that impact the coding region. *Txt4736* was found to be expressed exclusively in meiotic and post-meiotic germ cells and is significantly more abundant than either of the known *Vash2* isoforms suggesting it may play an important role in germ cell biology.

Novel protein-coding genes

Unlike novel isoforms, novel genes as well as antisense and intronic transcripts were found to have overall short open-reading frames and low protein-coding potential. However, a subset of novel genes was found to have protein-coding potential similar to that of known protein-coding genes (Supplementary Fig. 3C). This analysis generated a non-conservative estimate of approximately 200 novel protein-coding genes. Given their high expression in meiotic and post-meiotic germ cells and their relatively high protein-coding potential, these novel genes represented a tractable model to test whether novel testis-expressed transcripts could generate protein.

Novel protein identification by shotgun mass spectrometry

Discovery, or shotgun, mass spectrometry (MS), requires a database of novel peptides against which to query the raw MS data and allows the detection of novel proteins from complex peptide mixes. Thus, prior to MS analysis, *in silico* translation of novel transcripts was used to generate an expanded testis proteome. However, the majority of *in silico* translation tools default to a minimum peptide length of 100 amino acids, excluding many proteins important to post-meiotic germ cell biology (for example, protamine 1–51 amino acids). In order to overcome this limitation, proteins of 50 amino acids or greater derived from in silico translation of novel transcripts were appended to the reported Ensembl proteome to generate the database for novel protein discovery (Ensembl + novel).

Fluorescent-activated cell sorting (FACs) and mass spectrometry of round spermatid proteins

Protein from FACs isolated round spermatids was subjected to shotgun MS. Prior to MS analysis, isolated cells were

Table 2 Ontology analysis of genes encoding novel isoforms.

Biological process	Gene ontology ID	No. genes	Fold enrichment	Select genes with known roles in reproduction
Histone acetylation	GO:0043967	41	2.69	*Brd4*, Chd5*, Ncoa1*
Protein monoubiquitination	GO:0045724	21	2.61	*Fancl, Neurl1a*, Pcgf5, Scml2*, Trim37*, Ube2w*
Nuclear envelope organization	GO:0006998	27	2.32	*Lmna*, Spag4*, Spast, Sun1**
DNA methylation	GO:0006306	30	2.28	*Asz1*, Ctcfl, Dnmt3a*, Fkbp6*, Kmt2a, Mael*, Trim28*
Synapsis	GO:0007129	34	2.28	*Dmc1*, Hormad1*, Mcmdc2*, Mlh1*, Msh4*, Sypc1*, Terb1**
Regulation of translation initiation	GO:0006446	37	2.18	*Boll*, Dazl*, Fmr1, Khdrbs1*, Paip2**
Centrosome cycle	GO:0007098	48	2.09	*Cdk5rap2*, Cep63*, Odf2*, Plk4*
RNP complex export from the nucleus	GO:0071426	39	2.01	*Alkbh5*, Nup107, Nxf2**

*Knockout results in male infertility (http://www.informatics.jax.org/).

assessed for DNA content and morphology (Supplementary Fig. 5A and B). FACs generated a highly enriched population of round spermatids with limited contamination, primarily from elongating spermatids and cellular debris. MS of the resulting protein lysate identified unique peptides assigned to over 1400 proteins including known round spermatid proteins like HSPA2 (Govin *et al.* 2006) and IPO5 (Loveland *et al.* 2006). Additionally, peptides unique to newly annotated protein coding transcripts previously identified as novel by the combined analysis were also identified. These proteins include CATSPERE1 (*Txt5393*), Gm16486 (*Txt153989*), and AC164099.2 (*Txt50054*). Ontology analysis of MS-detected, known proteins showed enrichment for processes specific or important in round spermatids such as translation and protein transport (Supplementary Fig. 5C and D).

In addition to known proteins, query of our Ensembl + novel peptide database identified a total of 1049 peptides that appeared to be derived from novel transcripts. In order to determine whether these peptides had been reported in other protein databases, a second peptide database (UniProt, https://www.uniprot.org/) was searched for the 1049 putative novel peptides. Of these, 296 were detectible in the UniProt database, underscoring the need for multiple database queries to ensure confidence when assigning a peptide as novel.

Identification of novel proteins from annotated non-coding RNAs by mass spectrometry

Of the 296 peptides detected in the UniProt database, but not Ensembl, a number were found to be encoded by transcripts currently annotated as non-coding. A total of seven annotated non-coding RNAs (ncRNAs) or their isoforms were identified as putative protein-coding transcripts in this way. Several of these were selected for molecular confirmation by RT-PCR and Sanger sequencing followed by molecular analysis of tissue specificity (Fig. 5). Included among the selected transcripts was *Txt37298* which encodes a 58 amino acid peptide, the C-terminus of which was identified five independent times in the MS analysis. Unlike the majority of transcripts encoding MS-identified peptides which have enriched expression in meiotic and post-meiotic germ cells (Supplementary Fig. 6A), *Txt37298* is expressed nearly equally across all cell types examined. Although currently annotated as ncRNA *AC121965.1*, there is ample evidence that *Txt37298* is a genuine protein-coding transcript. Homologues in five other species, including human, have been identified for this transcript and all generate a known protein, NDUFB1, which is a nuclear genome-encoded subunit of mitochondrial complex I required for complex assembly (Stroud *et al.* 2016). Additionally, peptides nearly identical to those identified in our analyses by MS have been reported in non-Ensembl databases (jPOST (Okuda *et al.* 2017), PRIDE (Jones *et al.* 2008), and PhosphoSitePlus (Hornbeck *et al.* 2015)). To validate the protein level expression of mNDUFB1 across tissues and within the testis, a rabbit antibody was generated against the C′ terminus of the computed protein and tested using Western blot and immunofluorescence. These analyses found mNDUFB1 to be detectible in a range of adult mouse tissues (see Supplementary materials for discussion) and to colocalize with a known mitochondrial protein, COX IV. Taken together, these analyses confirm *Txt37298* as a genuine protein coding transcript for a conserved mitochondrial protein.

Identification of novel proteins from novel genes by mass spectrometry

The 753 detected peptides found in neither the Ensembl or UniProt databases represented a total of 243 novel transcripts, 64 of which encoded proteins detected by three or more unique peptides. Of these, four were selected for further study. In spite of having a similar number of transcripts with high protein coding potential scores, no peptides from antisense or intronic transcripts were detected. RT-PCR and Sanger sequencing of three novel protein coding genes confirmed their computed sequence and tissue specificity (Fig. 2C). Further analysis demonstrated all four to have highly enriched expression in the meiotic and post-meiotic germ cell populations (Supplementary Fig. 6A).

Four peptides of the 1119 amino acid protein encoded by *Txt41000* were detected by MS (Fig. 6A). Domain analysis identified a conserved kinase domain in the computed protein that is shared with a family of proteins associated with sperm motility. The member of this family with the highest identity (69.6%) with the *Txt41000* protein was Chinese hamster sperm motility kinase X. Limited identity was found for the mouse (41.4%) and rat (40.1%) sperm motility kinase X (SMOKX). SMOKX is a member of a family of kinases known as the sperm motility kinases, encoded for by a total of 8 *Smok* genes (www.informatics.jax.org). The founding member of this family, *Smok1*, is associated with t-complex associated transmission ratio distortion and likely functions to regulate flagellar function in sperm (Herrmann *et al.* 1999). The expression of *Txt41000* is similar to that of other detected *Smok* genes (Fig. 6B), suggesting the protein encoded for by *Txt41000* (termed here putative sperm motility kinase or pSMK) may be functionally similar to other sperm motility kinases. Given the potential relevance to male fertility, an peptide polyclonal antibody against pSMOK was generated in mouse. Western blotting with post-immune sera showed pSMOK to be detectible in the testis (Fig. 6C) with a distinct round spermatids localization pattern.

The protein encoded by *Txt141909* was also detected by MS, which identified a total of 11 peptides spanning almost 16% of the 358 amino acid protein (Fig. 6B). A WD repeat domain was identified in the *Txt141909* protein

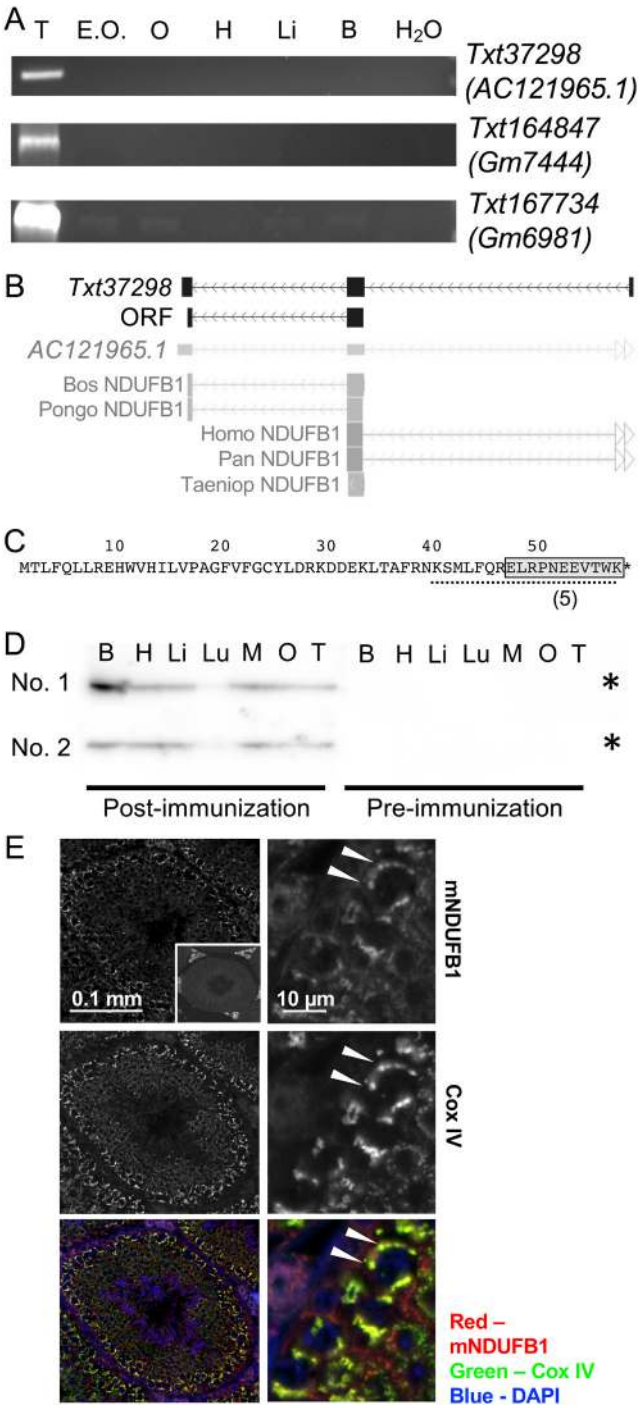

Figure 5 Annotated ncRNAs generate peptides detectible by MS and polyclonal antibody. (A) Reverse transcriptase PCR detection of transcripts annotated as non-coding but with MS detectible peptides. PCR product spans the entire computed ORF and a portion of the 5′ and 3′ UTRs. *Txt* – transcript ID within expanded transcriptome. ncRNA accession in parentheses. H_2O – water template (negative control). Tissues from adult C57Bl6/J unless indicated. $n=3$. Representative gels shown. (B) Transcript and ORF structure of *Txt37298* relative to its annotated ncRNA and non-mouse homologues (grey). Black and grey boxes indicate exons. Arrow heads along lines indicate introns and transcript direction within the genome. Open arrows indicate transcript continues beyond frame. (C) In silico translation of the *Txt37298* ORF. MS detected peptides highlighted in box. Number of detected peptides in parenthesis. Dashed lined - peptide used for antibody production. (D) Western blot detection of mNDUFB1 in adult mouse tissues by pre- and post-immune serum derived from two individuals (No. 1 and No. 2). Asterisks indicate approximately 8 kDa. $n=3$, representative blots shown. (E) Immunofluorescent detection of mNDUFB1 (inset – no antibody control) with COX IV. Arrow heads highlight regions of co-localization. T, testis; E.O., 17.5 dpc ovary; O, ovary; H, heart; Li, liver; Lu, lung; M, muscle; B, brain.

and a close protein homologue (rat WD repeat domain 88) was identified. Similar identity scores were found for WD repeat domain 88 proteins in other species as well, demonstrating the protein encoded by *Txt141909* is likely a new member of the WDR protein family.

Unlike *Txt37298* and *Txt141909*, which appear to encode conserved proteins, the other two novel genes identified by MS as putative protein coding genes (*Txt62639* and *Txt13871*) encode proteins with no known domains or homologues outside of mouse. *Txt62639* generates a protein 271 amino acids in length, 47 amino acids of which was spanned by seven MS-identified peptides (Supplementary Fig. 6B). The *Txt62639* protein shares nearly 100% identity to an uncharacterized mouse protein (Q9D4B5_MOUSE) encoded by a RIKEN gene *4933404G15Rik*. Detailed alignment suggests *Txt62639* and *4933404G15Rik* are the same gene but *4933404G15Rik* was excluded from the ENSEMBL annotation. The remaining gene identified by MS, *Txt13871*, contains no large open reading frame. However, a small open reading frame encoding an 87 amino acids peptide was identified computationally and two peptides from this ORF were detected by MS, indicating *Txt13871* may generate a short peptide.

Discussion

To address whether the testis transcriptome encodes novel proteins, an expanded testis transcriptome was generated and its expression and protein coding capacity described. These analyses demonstrated a large portion of novel testis-expressed transcripts are exclusive to the testis and predominantly expressed in meiotic and post-meiotic germ cells. Further, the majority of novel isoforms and a subset of novel genes within the expanded testis transcriptome have protein coding probabilities similar to known protein coding genes, suggesting they generate proteins. While roughly half of the alternative isoforms contain novel exons that may impact protein coding, the remaining contain novel exons outside of the open reading frame suggesting a possible role in post-transcriptional regulation. To test whether novel genes with high protein coding potential generated protein products, shotgun mass spectrometry (MS) of isolated round spermatid proteins was used

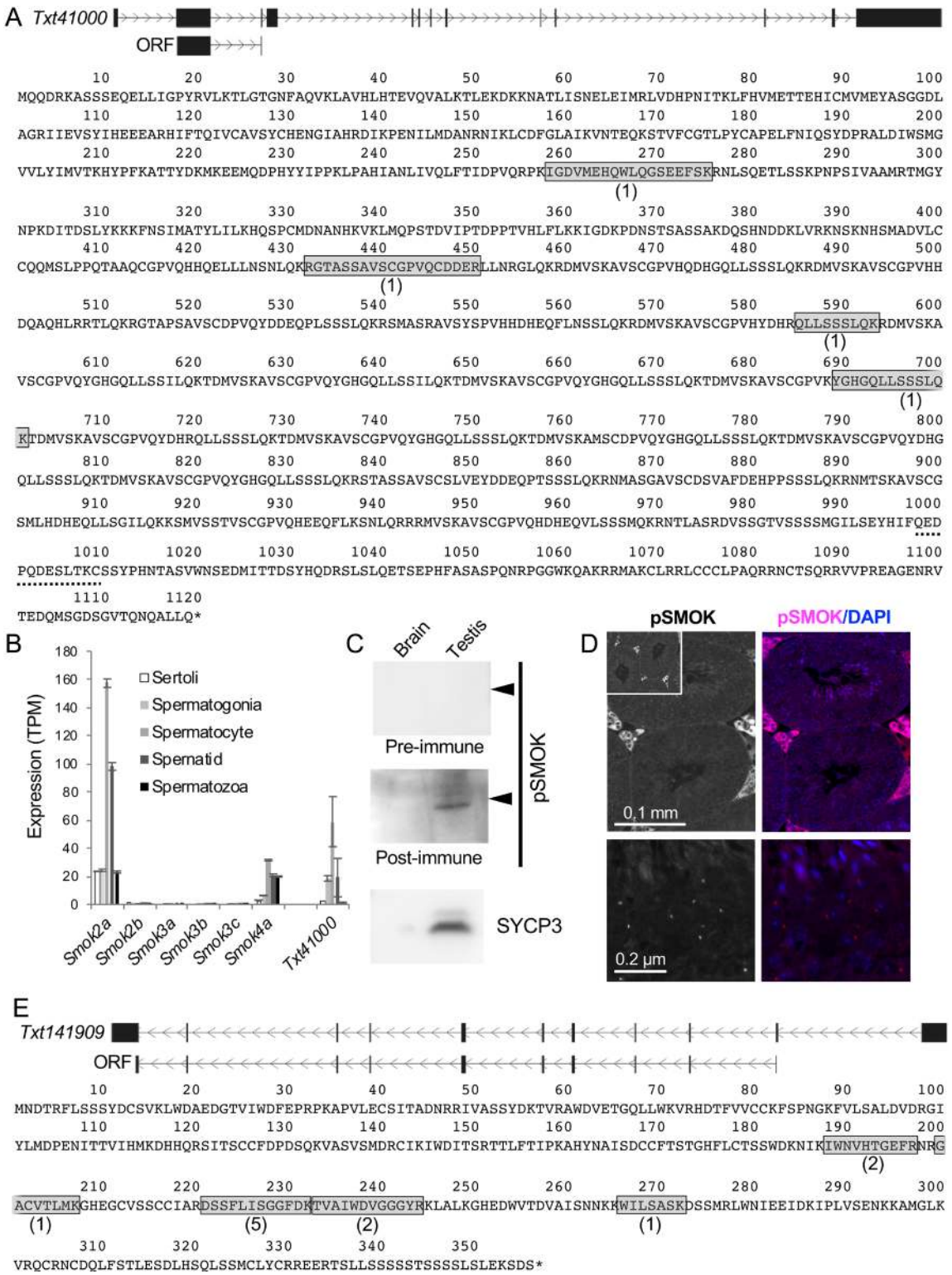

Figure 6 Novel protein-coding genes detected by MS and polyclonal antibody. (A) Transcript and ORF structure relative to the mm10 genome and computed protein for *Txt41000* Transcript coordinates: Chr13:49,814,407-49,896,053, dotted line indicates peptide used for antibody generation. (B) Cell type expression profile of *Txt41000* compared to known sperm motility kinases. (C) Western blot detection of a putative sperm motility kinase (pSMOK) encoded by *Txt41000*. Arrowhead indicates approximately 130 kDa. SYCP3 shown as a positive control for the testis. (D) Immunofluorescent detection of pSMOK (magenta) in adult testis. (E) *Txt141909* (transcript coordinates: Chr7:35,235,341-35,275,554). Black boxes indicate exons. Arrow heads along lines indicate introns and transcript direction within the genome. MS detected peptides highlighted in box. Number of detected peptides in parenthesis.

and identified a number of novel proteins derived from novel transcripts. MS also identified peptides encoded by a number of annotated ncRNAs. These findings were confirmed via the generation and analysis of antibodies against multiple targets (see Supplementary material for an expanded discussion).

Novel isoforms impact gene function on multiple levels

A large number of novel isoforms were identified in this study, findings reflective of early genome scale analyses suggesting the testis generates a large number of alternatively-spliced transcripts (Shima *et al.* 2004, Wang *et al.* 2008, Soumillon *et al.* 2013). Many novel isoforms identified herein are derived from genes known to be fundamental for male germ cell differentiation and retain high protein coding potential. These novel isoforms may impact gene function by alterating either the encoded proteins or post-transcriptional regulation of the transcript. Roughly half of the novel exons identified in this study reside within the open reading frame. As such, it is likely at least some isoforms generate novel peptides, as observed in *Vash2*. These novel peptides may drive novel functions or be differentially regulated on the protein level. Previous reports have demonstrated that many alternative isoforms give rise to alternatively phosphorylated proteins (Wang *et al.* 2008). A similarly large fraction of novel exons identified in this study are located outside of the open reading frame. This suggests many novel isoforms impact untranslated regions (UTRs) or are the result of alternative transcription start site selection or transcript polyadenylation. Thus, these transcripts may alter the encoded protein or impact post-transcriptional regulation.

Given many isoforms of reproductively important genes are capable of altering gene function, these isoforms may have cell-dependent functions and thus be drivers of germ cell differentiation. Supporting this notion, over 3000 genes expressed in the testis generate two or more novel isoforms that have different cell expression profiles, as seen in *Cpeb2*. Whether it be by changing a gene's protein coding potential or altering its post-transcriptional regulation, this apparent isoform switching suggests novel isoforms may be a mechanism to modulate gene function throughout germ cell development. These conclusions are buoyed by many single gene examples (see Foulkes *et al.* 1992, Goodson *et al.* 1995, Wang *et al.* 2013 for examples).

The testis as a site for novel protein discovery

Multiple reports have suggested the testis is a site of particularly high proteomic diversity in regards to known protein-coding transcripts (Djureinovic *et al.* 2014, Fagerberg *et al.* 2014, Uhlen *et al.* 2015). As a result, the testis (Melaine *et al.* 2018) and individual testicular cells such as spermatozoa (Jumeau *et al.* 2015) have been shown to be key sites for mass spectrometry-based discovery of previously annotated but undetected ("missing") proteins. The identification of additional novel protein-coding transcripts in the testis and detection of proteins derived from annotated ncRNAs suggests testis proteome diversity may be even greater and highlights the testis as a particularly unique site for both transcript and protein variant discovery. The idea that developing germ cells leverage novel components of the genome is not a new one. It was proposed nearly a decade ago that the testis is a site of gene evolution (Kaessmann 2010). The observation that novel, testis-specific genes give rise to new proteins is strong evidence supporting the idea of the testis as a birthplace for novel genes.

Given the expression of novel genes in meiotic and post-meiotic germ cells and the novel biology of each, it is tempting to hypothesize these newly discovered genes play specialized roles in either meiosis or post-meiotic events. Since the majority of novel proteins discovered have relatively high homology to known gene families, gene duplication and germ cell specific repurposing may be a major driver of novel gene expression in meiotic and post-meiotic germ cells. These observations fit well with a rich body of literature describing germ-cell specific protein and transcript isoforms (see Uhlen *et al.* 2015 for an overview and Dass *et al.* 2007, Sun *et al.* 2010, Ueda *et al.* 2017 for specific examples). As a result of their restricted expression and potentially novel functions, newly discovered protein-coding genes represent excellent avenues for infertility or contraceptive research.

Based on the findings reported here, the testis appears to be an excellent tissue for the detection of novel proteins from either unknown genes or genes misannotated as non-coding. This analysis showed many novel protein-coding transcripts are predominantly expressed in a single germ cell type and certain cell types express a higher frequency of novel protein coding genes. The expected outcome is a highly individualized proteome that facilitates the diversity of germ cell functions. With the advent of more advanced transcript discovery (Haas *et al.* 2013, Pertea *et al.* 2016) and proteomic approaches, the findings reported herin will likely be expanded upon. These efforts should focus on the development of more tissue and cell-specific transcriptome discovery analyses coupled to stringent database curation, efforts that will better connect tissue, cell, or organism transcriptomes to proteomes. Ultimately, these analysis are likely to uncover novel facets of known protein function or regulation as well as identifying completely novel proteins.

Supplementary materials

This is linked to the online version of the paper at https://doi.org/10.1530/REP-19-0092.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding

This work was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NIH-NICHD F32 HD072628 and K99/R00 HD083521 to E M S, NIH-NICHD HD027215 to R E B, and NIH P50 GM076468 to J H G). The authors would also like to thank their non-federal funding support. EMS – The Jackson Laboratory and Rutgers University.

Author contribution statement

J G conducted, analyzed, and summarized the novel transcript molecular confirmation studies. J C conducted and analyzed all aspect of the proteomics analysis and assisted with manuscript editing and revision. K S optimized and conducted portions of the novel transcript molecular confirmation studies including the generation of a custom antibody against pSMOK. J H G designed the Spliced RUM analysis pipeline and applied it to testis RNA sequencing data. S G provided support and intellectual input for the proteomics studies. R B provided support and intellectual input for the transcriptomics studies, suggested experimental approaches and designs, and manuscript editing. J H G and E S conceived of the initial experimental design. E S conducted the majority of the transcriptomic analyses; designed and assisted with all molecular confirmation studies; and drafted, edited, and revised the manuscript.

Acknowledgments

The authors would like to sincerely thank Drs Robyn Ball, Nazira Bektassova, and Lucie Hutchins for their computational and statistical assistance. Drs Steven Munger and Mary Ann Handel for their critical evaluation of this manuscript; members of the Braun Laboratory (Alexandra Lyahkovich and Christopher McCarty) for their molecular analysis efforts; and members of the Snyder Laboratory (Kelly Seltzer, Lauren Chukralluh, and Gabriella Acoury) for their critical feedback and efforts with molecular analyses.

References

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M *et al.* 2011 The evolution of gene expression levels in mammalian organs. *Nature* **478** 343–348. (https://doi.org/10.1038/nature10532)

Brennan J, Tilmann C & Capel B 2003 Pdgfr-alpha mediates testis cord organization and fetal Leydig cell development in the XY gonad. *Genes and Development* **17** 800–810. (https://doi.org/10.1101/gad.1052503)

Chocu S, Evrard B, Lavigne R, Rolland AD, Aubry F, Jegou B, Chalmel F & Pineau C 2014 Forty-four novel protein-coding loci discovered using a proteomics informed by transcriptomics (PIT) approach in rat male germ cells. *Biology of Reproduction* **91** 123. (https://doi.org/10.1095/biolreprod.114.122416)

Com E, Melaine N, Chalmel F & Pineau C 2014 Proteomics and integrative genomics for unraveling the mysteries of spermatogenesis: the strategies of a team. *Journal of Proteomics* **107** 128–143. (https://doi.org/10.1016/j.jprot.2014.04.013)

Dass B, Tardif S, Park JY, Tian B, Weitlauf HM, Hess RA, Carnes K, Griswold MD, Small CL & Macdonald CC 2007 Loss of polyadenylation protein tauCstF-64 causes spermatogenic defects and male infertility. *PNAS* **104** 20374–20379. (https://doi.org/10.1073/pnas.0707589104)

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F *et al.* 2012 Landscape of transcription in human cells. *Nature* **489** 101–108. (https://doi.org/10.1038/nature11233)

Djureinovic D, Fagerberg L, Hallstrom B, Danielsson A, Lindskog C, Uhlen M & Ponten F 2014 The human testis-specific proteome defined by transcriptomics and antibody-based profiling. *Molecular Human Reproduction* **20** 476–488. (https://doi.org/10.1093/molehr/gau018)

Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K *et al.* 2014 Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular and Cellular Proteomics* **13** 397–406. (https://doi.org/10.1074/mcp.M113.035600)

Foulkes NS, Mellstrom B, Benusiglio E & Sassone-Corsi P 1992 Developmental switch of CREM function during spermatogenesis: from antagonist to activator. *Nature* **355** 80–84. (https://doi.org/10.1038/355080a0)

Gaysinskaya V, Soh IY, Van Der Heijden GW & Bortvin A 2014 Optimized flow cytometry isolation of murine spermatocytes. *Cytometry: Part A* **85** 556–565. (https://doi.org/10.1002/cyto.a.22463)

Goodson ML, Park-Sarge OK & Sarge KD 1995 Tissue-dependent expression of heat shock factor 2 isoforms with distinct transcriptional activities. *Molecular and Cellular Biology* **15** 5288–5293. (https://doi.org/10.1128/mcb.15.10.5288)

Govin J, Caron C, Escoffier E, Ferro M, Kuhn L, Rousseaux S, Eddy EM, Garin J & Khochbin S 2006 Post-meiotic shifts in HSPA2/HSP70.2 chaperone activity during mouse spermatogenesis. *Journal of Biological Chemistry* **281** 37888–37892. (https://doi.org/10.1074/jbc.M608147200)

Guo X, Zhang P, Huo R, Zhou Z & Sha J 2008 Analysis of the human testis proteome by mass spectrometry and bioinformatics. *Proteomics: Clinical Applications* **2** 1651–1657. (https://doi.org/10.1002/prca.200780120)

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M *et al.* 2013 De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8** 1494–1512. (https://doi.org/10.1038/nprot.2013.084)

Herrmann BG, Koschorz B, Wertz K, Mclaughlin KJ & Kispert A 1999 A protein kinase encoded by the t complex responder gene causes non-Mendelian inheritance. *Nature* **402** 141–146. (https://doi.org/10.1038/45970)

Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V & Skrzypek E 2015 PhosphoSitePlus, 2014: mutations, PTMS and recalibrations. *Nucleic Acids Research* **43** D512–D520. (https://doi.org/10.1093/nar/gku1267)

Jones P, Cote RG, Cho SY, Klie S, Martens L, Quinn AF, Thorneycroft D & Hermjakob H 2008 PRIDE: new developments and new datasets. *Nucleic Acids Research* **36** D878–D883. (https://doi.org/10.1093/nar/gkm1021)

Jumeau F, Com E, Lane L, Duek P, Lagarrigue M, Lavigne R, Guillot L, Rondel K, Gateau A, Melaine N *et al.* 2015 Human spermatozoa as a model for detecting missing proteins in the context of the chromosome-centric human proteome project. *Journal of Proteome Research* **14** 3606–3620. (https://doi.org/10.1021/acs.jproteome.5b00170)

Kaessmann H 2010 Origins, evolution, and phenotypic impact of new genes. *Genome Research* **20** 1313–1326. (https://doi.org/10.1101/gr.101386.109)

Kleene KC 2001 A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. *Mechanisms of Development* **106** 3–23. (https://doi.org/10.1016/s0925-4773(01)00413-0)

Kurihara Y, Tokuriki M, Myojin R, Hori T, Kuroiwa A, Matsuda Y, Sakurai T, Kimura M, Hecht NB & Uesugi S 2003 CPEB2, a novel putative translational regulator in mouse haploid germ cells. *Biology of Reproduction* **69** 261–268. (https://doi.org/10.1095/biolreprod.103.015677)

Loveland KL, Hogarth C, Szczepny A, Prabhu SM & Jans DA 2006 Expression of nuclear transport importins beta 1 and beta 3 is regulated during rodent spermatogenesis. *Biology of Reproduction* **74** 67–74. (https://doi.org/10.1095/biolreprod.105.042341)

Melaine N, Com E, Bellaud P, Guillot L, Lagarrigue M, Morrice NA, Guevel B, Lavigne R, Velez De La Calle JF, Dojahn J *et al.* 2018 Deciphering the dark proteome: use of the testis and characterization of two dark proteins. *Journal of Proteome Research* **17** 4197–4210. (https://doi.org/10.1021/acs.jproteome.8b00387)

Merkin J, Russell C, Chen P & Burge CB 2012 Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338** 1593–1599. (https://doi.org/10.1126/science.1228186)

Namekawa SH, Park PJ, Zhang LF, Shima JE, Mccarrey JR, Griswold MD & Lee JT 2006 Postmeiotic sex chromatin in the male germline of mice. *Current Biology* **16** 660–667. (https://doi.org/10.1016/j.cub.2006.01.066)

Naro C, Jolly A, Di Persio S, Bielli P, Setterblad N, Alberdi AJ, Vicini E, Geremia R, De La Grange P & Sette C 2017 An orchestrated intron retention program in meiosis controls timely usage of transcripts during germ cell differentiation. *Developmental Cell* **41** 82.e4–93.e4. (https://doi.org/10.1016/j.devcel.2017.03.003)

Okuda S, Watanabe Y, Moriya Y, Kawano S, Yamamoto T, Matsumoto M, Takami T, Kobayashi D, Araki N, Yoshizawa AC *et al.* 2017 jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Research* **45** D1107–D1111. (https://doi.org/10.1093/nar/gkw1080)

Pertea M, Kim D, Pertea GM, Leek JT & Salzberg SL 2016 Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols* **11** 1650–1667. (https://doi.org/10.1038/nprot.2016.095)

Ramskold D, Wang ET, Burge CB & Sandberg R 2009 An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Computational Biology* **5** e1000598. (https://doi.org/10.1371/journal.pcbi.1000598)

Russell L, Sinha Hikim A, Ettlin R & Clegg E 1990 *Histological and Histopathological Evalutation of the Te4stis*. St. Louis, MO: Cache River Press. (https://doi.org/10.1111/j.1365-2605.1993.tb01156.x)

Shibuya T, Watanabe K, Yamashita H, Shimizu K, Miyashita H, Abe M, Moriya T, Ohta H, Sonoda H, Shimosegawa T *et al.* 2006 Isolation and characterization of vasohibin-2 as a homologue of VEGF-inducible endothelium-derived angiogenesis inhibitor vasohibin. *Arteriosclerosis, Thrombosis, and Vascular Biology* **26** 1051–1057. (https://doi.org/10.1161/01.ATV.0000216747.66660.26)

Shima JE, Mclean DJ, Mccarrey JR & Griswold MD 2004 The murine testicular transcriptome: characterizing gene expression in the testis during the progression of spermatogenesis. *Biology of Reproduction* **71** 319–330. (https://doi.org/10.1095/biolreprod.103.026880)

Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthes P, Kokkinaki M, Nef S, Gnirke A *et al.* 2013 Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Reports* **3** 2179–2190. (https://doi.org/10.1016/j.celrep.2013.05.031)

Stroud DA, Surgenor EE, Formosa LE, Reljic B, Frazier AE, Dibley MG, Osellame LD, Stait T, Beilharz TH, Thorburn DR *et al.* 2016 Accessory subunits are integral for assembly and function of human mitochondrial complex I. *Nature* **538** 123–126. (https://doi.org/10.1038/nature19754)

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D *et al.* 2008 A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321** 956–960. (https://doi.org/10.1126/science.1160342)

Sun F, Palmer K & Handel MA 2010 Mutation of Eif4g3, encoding a eukaryotic translation initiation factor, causes male infertility and meiotic arrest of mouse spermatocytes. *Development* **137** 1699–1707. (https://doi.org/10.1242/dev.043125)

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ & Pachter L 2010 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28** 511–515. (https://doi.org/10.1038/nbt.1621)

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL & Pachter L 2012 Differential gene and transcript expression analysis of Rna-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7** 562–578. (https://doi.org/10.1038/nprot.2012.016)

Ueda J, Harada A, Urahama T, Machida S, Maehara K, Hada M, Makino Y, Nogami J, Horikoshi N, Osakabe A *et al.* 2017 Testis-specific histone variant H3t gene is essential for entry into spermatogenesis. *Cell Reports* **18** 593–600. (https://doi.org/10.1016/j.celrep.2016.12.065)

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjostedt E, Asplund A *et al.* 2015 Proteomics. Tissue-based map of the human proteome. *Science* **347** 1260419. (https://doi.org/10.1126/science.1260419)

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP & Burge CB 2008 Alternative isoform regulation in human tissue transcriptomes. *Nature* **456** 470–476. (https://doi.org/10.1038/nature07509)

Wang L, Park HJ, Dasari S, Wang S, Kocher JP & Li W 2013 CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Research* **41** e74. (https://doi.org/10.1093/nar/gkt006)

Wang J, Xia Y, Wang G, Zhou T, Guo Y, Zhang C, An X, Sun Y, Guo X, Zhou Z *et al.* 2014 In-depth proteomic analysis of whole testis tissue from the adult rhesus macaque. *Proteomics* **14** 1393–1402. (https://doi.org/10.1002/pmic.201300149)

Xue X, Gao W, Sun B, Xu Y, Han B, Wang F, Zhang Y, Sun J, Wei J, Lu Z *et al.* 2013 Vasohibin 2 is transcriptionally activated and promotes angiogenesis in hepatocellular carcinoma. *Oncogene* **32** 1724–1734. (https://doi.org/10.1038/onc.2012.177)

Received 26 February 2019
First decision 26 March 2019
Revised manuscript received 24 October 2019
Accepted 31 October 2019