Long noncoding RNAs in spermatogenesis: insights from recent high-throughput transcriptome studies

Alfred Chun-Shui Luk, Wai-Yee Chan, Owen M Rennert1 and Tin-Lap Lee

School of Biomedical Sciences, Room 622A, Lo Kwee-Seong Integrated Biomedical Sciences Building, The Chinese University of Hong Kong, Shatin, Hong Kong, China and 1Laboratory of Clinical and Developmental Genomics, National Institute of Child Health & Human Development, National Institutes of Health, Bethesda, Maryland, USA

Correspondence should be addressed to T-L Lee; Email: leetl@cuhk.edu.hk

Abstract

Spermatogenesis is a complex developmental process in which undifferentiated spermatogonia are differentiated into spermatocytes and spermatids through two rounds of meiotic division and finally giving rise to mature spermatozoa (sperm). These processes involve many testis- or male germ cell-specific gene products that undergo strict developmental regulations. As a result, identifying critical, regulatory genes controlling spermatogenesis provide the clues not only to the regulatory mechanism of spermatogenesis at the molecular level, but also to the identification of candidate genes for infertility or contraceptives development. Despite the biological importance in male germ cell development, the underlying mechanisms of stage-specific gene regulation and cellular transition during spermatogenesis remain largely elusive. Previous genomic studies on transcriptome profiling were largely limited to protein-coding genes. Importantly, protein-coding genes only account for a small percentage of transcriptome; the majority are noncoding transcripts that do not translate into proteins. Although small noncoding RNAs (ncRNAs) such as microRNAs, siRNAs, and Piwi-interacting RNAs are extensively investigated in male germ cell development, the role of long ncRNAs (lncRNAs), commonly defined as ncRNAs longer than 200 bp, is relatively unexplored. Herein, we summarize recent transcriptome studies on spermatogenesis and show examples that a subset of noncoding transcript population, known as lncRNAs, constitutes a novel regulatory target in spermatogenesis.

Reproduction (2014) 147 R131–R141

Introduction

Spermatogenesis refers to the continuous production of male gametes (i.e. sperms) from spermatogonial stem cells (SSCs) through a series of tightly regulated developmental programs involving mitosis, meiosis, and differentiation inside the testes. SSCs develop from gonocytes after birth and are able to self-renew and maintain a relatively constant population on the basement membrane of seminiferous tubules throughout reproductive life until the number declines due to aging factors. The microenvironment inside the seminiferous tubules where SSCs proliferate and differentiate represents the ‘niche’, contributed by Sertoli cells, the basement membrane, Leydig cells, and other interstitial cells. The ‘niche’ nourishes the SSC development and supports the whole process of spermatogenesis. A sub-population of SSCs commits differentiation pathway into type A spermatogonia, followed by the formation of type B spermatogonia, preleptotene spermatocytes, leptotene spermatocytes, zygotene spermatocytes, pachytene spermatocytes, round spermatids, elongated spermatocytes, and finally spermatozoa (or sperms; Oatley & Brinster 2012; Fig. 1).

Therefore, studying the developmental processes at various male germ cell stages during spermatogenesis allows a comprehensive understanding of the cellular and molecular regulations that underlie cellular proliferation and differentiation. Identification of key regulators or signaling pathways in these developmental programs will potentially be useful in generating novel hypotheses, designing diagnostic biomarkers, and therapeutic interventions in various reproductive health issues.

Noncoding RNAs: overlooked transcript species in genome biology

The central dogma of molecular biology suggests that proteins perform the most of structural, functional, and regulatory roles in various cellular functions. However, this protein-centric view has recently been challenged, as the number of protein-coding genes in higher organisms is very similar to lower ones, and this is not sufficient to account for their complexity. It was estimated that known genes only account for around 5–10% of the mammalian genome. However, recent
large-scale transcriptome studies have suggested that the majority of genome is transcribed and outnumbers the current gene annotations. Importantly, a very large number of transcripts do not appear to encode for proteins. These transcript species are known as noncoding RNAs (ncRNAs). Previously, they were often considered as ‘junk’ in the genome, but emerging evidence has highlighted that ncRNAs can have a wide variety of functions. Prokaryotes have <25% noncoding DNA; simple eukaryotes have between 25 and 50% noncoding DNA; and more complex fungi, plants, and animals have more than 50%, rising to ~98.5% noncoding DNA in humans (Amaral et al. 2008).

ncRNAs can be generally classified according to length. The short class refers to ncRNAs with <200 bp; anything longer than that is referred to as long ncRNA (lncRNA). Examples of short ncRNAs include rRNA, microRNA (miRNA), siRNA, and Piwi-interacting RNA (piRNA). Application of short RNAs has contributed significantly in biology and medicine in recent years. Short ncRNAs involved in spermatogenesis have been systematically reviewed previously (Bettegowda & Wilkinson 2010, Chuma & Nakano 2013, Yadav & Kotaja 2014) and have proved to be indispensible in various developmental processes in male germ cell development.

A novel regulatory layer in cellular development: lncRNAs

One of the major breakthroughs in recent high-throughput transcriptomic studies and genome-wide association studies is the recognition on multiple regulatory roles played by lncRNAs (Djebali et al. 2012). Many lncRNAs are transcribed by RNA polymerase II just as a mRNA without protein coding ability, or formed by processing other transcripts. While short RNA relies largely on partnership with protein, such as argonaut (Yadav & Kotaja 2014) and piwi proteins (Chuma & Nakano 2013), for regulatory functions, there is no specific group of protein regulator identified for universal processing of lncRNA transcripts. Recently, a number of hallmark reports have revealed the versatile molecular mechanisms of regulation exerted by lncRNA in human or rodent embryonic stem cell (ESC) models, such as interacting with proteins and regulating their function (Hung et al. 2011, Ng et al. 2013), regulating the action of chromatin-modifying factors by directing interaction (Klattenhoff et al. 2013), inactivating X chromosome by direct binding (Ohhata et al. 2011), promoting organelle formation (Chen & Carmichael 2009), inhibiting antisense mRNA expression (Santoro et al. 2013), promoting antisense mRNA degradation (Kraus et al. 2013), regulating mRNA export from nucleus (Chen & Carmichael 2009), acting as competing endogenous RNAs (ceRNAs) to inhibit the function of miRNA (Cheng & Lin 2013), and acting as a decoy which binds DNA-binding proteins and prevents them from approaching target sites in genomic DNA (Kino et al. 2010, Redon et al. 2010). Given such a wide variety of regulatory roles, the exact molecular regulations conferred by lncRNAs remain to be elucidated. In brief, lncRNAs can be classified based on

---

**Figure 1** The complexity of mouse spermatogenesis involves different cell types, cellular stages at various developmental time-frames. **Top layer:** Diagrams showing the key developmental stages in mouse spermatogenesis - gonocytes (light pink), spermatogonia (dark pink), spermatocytes (purple), spermatids and sperm (cyan). **Second layer:** cell types development at each stage. **Third layer:** Developmental processes in spermatogenesis. **Forth layer:** Ploidity of each stage. **Fifth layer:** Postnatal day in which some of the representative cell types arisen in first wave of mouse spermatogenesis.
Genomic locations in terms of transcript associated or nontranscript associated (intergenic) and the direction of transcription (sense vs antisense). Transcript-associated candidates can be further divided based on genomic features that are classified as promoter-, exon-, intron-, or UTR-associated IncRNAs. Based on recent studies, the molecular mechanisms of IncRNA regulation can mainly be summarized into three modes of regulations, including competitor, recruiter/activator, and precursor (Fig. 2). First, as a competitor, IncRNA can bind to DNA-binding proteins, such as transcription factors (Hung 2011, Ng 2013), and inhibit their attachment to target DNA. They can also bind complementarily to some structures in DNA molecules and hinder protein binding to specific antisense mRNA (Kraus et al. 2013), or to miRNAs for post-transcriptional regulation (Zhang et al. 2010). Secondly, IncRNA can activate epigenetic modifiers or recruit them to specific target sites, thus enhancing DNA methylation (Berghoff et al. 2013) and histone modifications (Yap et al. 2010, Klattenhoff et al. 2013). IncRNAs are suggested to increase the proximity of those proteins to their reaction sites, thus enhancing the rate of reactions such as DNA methylation and histone modifications. Finally, IncRNAs can be processed by certain RNase such as Drosha and Dicers to form shorter RNAs, which in turn may carry out further post-transcriptional regulation–examples include H19 (Keniry et al. 2012) and HongrES2 (Ni et al. 2011).

Public resources for IncRNA prediction and analysis

The locations and configuration of IncRNA genes in the genome, known as annotation, are updated in various public databases regularly, providing the basis for IncRNA prediction. Currently, the majority of mammalian IncRNA annotation is based on humans and rodents. As listed in Table 1, some databases contain both coding RNA and ncRNA annotations, while some provided ready-to-use datasets of IncRNAs. One can download the IncRNA datasets in Browser Extensible Data (BED) or Gene Transfer Format (GTF) format to visualize them in genome browsers (e.g. UCSC genome browser: http://genome.ucsc.edu) or perform further analyses (for details, please visit: http://genome.ucsc.edu/FAQ/FAQ-format.html). Furthermore, disease association and molecular interaction and function are also important.

Figure 2 Examples on molecular regulation of IncRNAs. The regulations can be summarized to three major categories: (A) competitor, (B) recruiter/activator, and (C) precursor.
issues in IncRNA studies. Some databases have taken the pioneer roles to systematically summarize relevant studies and to provide tools to predict potential interactions and functions of IncRNAs.

### IncRNAs in mammalian spermatogenesis

#### Transcriptome studies

Identification of functional IncRNA was technically challenging before the advent of high-throughput genome-wide expression assays. Only few IncRNAs, such as H19 (Pachnis et al. 1984), XIST (Brown et al. 1991), and KCNQ1OT1 (Lee et al. 1999), were discovered and characterized. Those discoveries were sporadic and confined to candidates with relatively high expression levels or close proximity to a cluster of known genes. Unbiased identification of noncoding transcripts in the genome is impossible without prior knowledge of genome sequence and high-throughput genomic analysis techniques. These limitations were eventually

---

**Table 1** Online resources for IncRNA annotation and characterization.

<table>
<thead>
<tr>
<th>Database</th>
<th>Web addresses</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>General IncRNA annotation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference Sequence Database (RefSeq), Pruitt et al. (2012)</td>
<td><a href="http://www.ncbi.nlm.nih.gov/refseq/">http://www.ncbi.nlm.nih.gov/refseq/</a></td>
<td>ReSeq provides RNA information on over 31 000 organisms. The related protein products, if any, are shown in corresponding RNA transcripts. Potential IncRNAs do not contain any coding sequence (CDS), and probably have accession prefix of NR or XR. Genomic information and expression evidence are provided for each transcript</td>
</tr>
<tr>
<td>Encyclopedia of DNA Elements (ENCODE) project, ENSEMBL, Derrien et al. (2012)</td>
<td><a href="http://ensembl.org/index.html">http://ensembl.org/index.html</a></td>
<td>RNA annotations of 78 species are available. Gene type of each transcript is shown. Long transcripts in 3’-overlapping ncRNA, antisense, IncRNA, polymorphic pseudogene, processed pseudogene, processed transcript, pseudogene, sense intronic, and sense overlapping may all contribute to IncRNAs. Gene information and expression evidence are provided for each transcript</td>
</tr>
<tr>
<td>USCS Genes, Meyer et al. (2013)</td>
<td><a href="http://genome.ucsc.edu/">http://genome.ucsc.edu/</a></td>
<td>UCSC genes is a summary of transcript annotations from other databases. IncRNAs found under this annotation are either annotated by RefSeq, or by GenBank plus one other source. The complete annotation can be directly visualized in UCSC genome browser or downloaded from UCSC table browsers for further analysis</td>
</tr>
<tr>
<td>ncRNA Expression Database (NRED), Dinger et al. (2009)</td>
<td><a href="http://nred.matticklab.com/cgi-bin/nccrna.pl">http://nred.matticklab.com/cgi-bin/nccrna.pl</a></td>
<td>NRED provides probe-centric expression information of mouse and human IncRNAs in over 50 tissues or stages. The data is based on results by custom ncRNA microarray experiments investigating six normal development backgrounds. Also, genomic information and evolutionary conservation are provided for each transcript</td>
</tr>
<tr>
<td>NONCODE, Bu et al. (2012)</td>
<td><a href="http://www.noncode.org/">http://www.noncode.org/</a></td>
<td>NONCODE contains annotations of non-coding RNAs from eight species. IncRNAs can be easily selected by size</td>
</tr>
<tr>
<td>Functional RNA Database (fRNAdb), Kin et al. (2007)</td>
<td><a href="http://www.nrcrna.org/frnadb/">http://www.nrcrna.org/frnadb/</a></td>
<td>fRNAdb contains noncoding RNA annotations of ten species. It provides information on transcription factors, associated diseases, sequence ontology, and gene ontology for some IncRNAs</td>
</tr>
<tr>
<td>Characterized IncRNA annotation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long noncoding RNA Database (IncRNAdb), Amaral et al. (2011)</td>
<td><a href="http://www.lncrnadb.org/">http://www.lncrnadb.org/</a></td>
<td>This database provides expression, literature, and some other characterizations of 50 characterized IncRNAs in various species</td>
</tr>
<tr>
<td>The IncRNA and Disease Database (IncRNADisease), Chen et al. (2013)</td>
<td><a href="http://cmbi.bjmu.edu.cn/IncRNADisease">http://cmbi.bjmu.edu.cn/IncRNADisease</a></td>
<td>This website provides both experimentally supported and predicted IncRNA-human disease relationships, based on hundreds of publications. It also shows relevant IncRNA–protein interaction, where applicable</td>
</tr>
<tr>
<td>Function and interaction prediction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noncoding RNA Functional Annotation Server (ncFANs), Liao et al. (2011)</td>
<td><a href="http://www.bioinfo.org/ncfans/">http://www.bioinfo.org/ncfans/</a></td>
<td>This tool provides a platform for obtaining IncRNA expression information from human and mouse microarray data. It also predicts the functions of IncRNAs via four algorithms under user’s statistics settings</td>
</tr>
<tr>
<td>ChIPBase, Yang et al. (2013)</td>
<td><a href="http://deepbase.sysu.edu.cn/chipbase/">http://deepbase.sysu.edu.cn/chipbase/</a></td>
<td>ChIPBase provides expression profiles of human IncRNA and transcription factor binding maps of IncRNAs from six species. All the information provided is based on ChIP-Seq experiments</td>
</tr>
<tr>
<td>starBase, Yang et al. (2011)</td>
<td><a href="http://starbase.sysu.edu.cn/">http://starbase.sysu.edu.cn/</a></td>
<td>This database provides information on IncRNA–miRNA and IncRNA–protein interactions. It also predicts ceRNA networks for IncRNA function prediction. All the information provided is based on CLIP-Seq studies</td>
</tr>
<tr>
<td>RNA–Protein Interaction Prediction (RPISeq), Muppilla et al. (2011)</td>
<td><a href="http://pridb.gdcb.iastate.edu/RPISeq/">http://pridb.gdcb.iastate.edu/RPISeq/</a></td>
<td>This online tool provides a means to predict any possible interaction between IncRNA of known sequence and proteins</td>
</tr>
</tbody>
</table>
removed by the genome projects and large-scale expression studies on various male germ cell development models (Table 2). The common platforms for delineating the transcriptome landscape on spermatogenesis are summarized below.

**cDNA and expressed sequence tag libraries**

The era of high-throughput datasets began with the construction of cDNA or expressed sequence tag (EST) libraries. The idea of this technology was to clone all mature mRNAs in a sample with a pull-down target, most commonly a poly-A tail, into a collection of cloning vectors. Although at that time researchers usually focused on cloning coding genes, the libraries also captured mature lncRNA with poly-A tail, and thus is an evidence for lncRNA expression.

**McCarrey et al. (1999)** prepared 23 comprehensive testis-derived libraries at $2 \times 10^5$ plaque-forming units (PFU) for whole testes, peritubular cells, Sertoli cells, and nine representative spermatogenic cell types of several strains of mice and rats. The cDNA is unidirectionally cloned to the expression vector. After validation with the presence of some known genes, the library can be used for screening the expression of any transcripts at each cell source with sequence-specific probes, for subsequent sequencing or overexpressing the whole transcript. **Wang et al. (2001)** also applied the cDNA library subtraction approach to identify genes specifically expressed in spermatogonia. Owing to sensitivity, only 25 genes could be identified, half of which resided in sex chromosomes, while the other half was distributed evenly on autosomes. Although these cloned libraries theoretically contain all mRNA transcripts with poly-A tails, detecting the presence or abundance of a specific transcript still relies on low-throughput procedures using northern blotting and Sanger sequencing. Another limitation of this technology is its low detection sensitivity due to the requirement for abundantly expressed transcripts in the cloning procedure.

Other than full-length cDNA libraries, a relatively rapid approach could be achieved by EST library construction.

### Table 2: High-throughput studies with data available for lncRNA identification in male germ cell development.

<table>
<thead>
<tr>
<th>References</th>
<th>Platform</th>
<th>Sample</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>McCarrey et al. (1999)</td>
<td>cDNA library</td>
<td>Whole testes, primitive type A, types A and B spermatogonia, preleptotene, leptotene, zygote and pachytene spermatocytes, round spermatids, and Sertoli cells from mice and rats</td>
<td></td>
</tr>
<tr>
<td>Wang et al. (2001)</td>
<td>cDNA library subtraction</td>
<td>Primitive type A from 6-day-old CD-1 mice, types A and B spermatogonia from 8-day-old CD-1 mice</td>
<td></td>
</tr>
<tr>
<td>Liu et al. (2007)</td>
<td>EST library</td>
<td>Mouse type A spermatogonia, pachytene spermatocytes, round spermatids, and Sertoli cells</td>
<td></td>
</tr>
<tr>
<td>Yao et al. (2004)</td>
<td>SAGE</td>
<td>Whole testes from 3- and 29-month-old BDF1 mice and 14-month-old SAMP1 mice</td>
<td>61 929 SAGE tags for three libraries</td>
</tr>
<tr>
<td>Divina et al. (2005)</td>
<td>SAGE</td>
<td>Whole testes of adult mouse</td>
<td>76 854 SAGE tags for two libraries</td>
</tr>
<tr>
<td>Wu et al. (2004), Chan et al. (2006) and Lee et al. (2009, 2012)</td>
<td>SAGE</td>
<td>Type A spermatogonia from PND 6 mice, pachytene spermatocytes, and round spermatids from PND 60 mice</td>
<td>333 470 SAGE tags for three libraries</td>
</tr>
<tr>
<td>Xiao et al. (2008)</td>
<td>Microarray (Genechip Mouse 430 v.2, Affymetrix)</td>
<td>Whole testes of mice aged 4, 9, 18, 35, 54 days, and 6 months</td>
<td></td>
</tr>
<tr>
<td>Waldman Ben-Asher et al. (2010)</td>
<td>Microarray (MGU74 v.2, Affymetrix)</td>
<td>Whole testes of mice at PNDs 7, 10, 12, 14, and 17</td>
<td></td>
</tr>
<tr>
<td>Bao et al. (2013)</td>
<td>Microarray (Mouse LncRNA Array v2.0 8×60k, Arraystar)</td>
<td>Whole testes of mice at six time points from E12.5 to adult</td>
<td></td>
</tr>
<tr>
<td>Sun et al. (2013)</td>
<td>Microarray (Mouse LncRNA Array v2.0 8×60k, Arraystar)</td>
<td>Whole testes from mouse at 6 days and 8 weeks</td>
<td></td>
</tr>
<tr>
<td>Das et al. (2013)</td>
<td>Microarray, RNA-Seq</td>
<td>Fresh sperms and testes of adult horses</td>
<td>rRNA is naturally depleted in sperms</td>
</tr>
<tr>
<td>Lesch et al. (2013)</td>
<td>ChIP-Seq for epigenetic marks, RNA-Seq</td>
<td>Flow-sorted male germ cells in mouse embryo at E12.5–E14.5, pachytene spermatocytes and round spermatids were isolated from adult testes</td>
<td>Poly-adenylated transcripts for RNA-Seq</td>
</tr>
<tr>
<td>Schmidt et al. (2013)</td>
<td>RNA-Seq</td>
<td>Whole testes of mice at PNDs 6 and 21</td>
<td>Poly-adenylated transcripts only</td>
</tr>
<tr>
<td>Gong et al. (2013)</td>
<td>RNA-Seq</td>
<td>Whole testes of mouse at 6 days, 4 weeks, and 10 weeks</td>
<td>Poly-adenylated transcripts only</td>
</tr>
<tr>
<td>Gan et al. (2013)</td>
<td>RNA-Seq/R5hmC-Seq</td>
<td>Isolated spermatogenic cells of eight stages, including primitive type A, types A and B spermatogonia, leptotene and pachytene spermatocytes, round and elongated spermatids, and spermatooza. Sertoli cells were also included for reference</td>
<td>Poly-adenylated transcripts for RNA-Seq</td>
</tr>
<tr>
<td>Laiho et al. (2013)</td>
<td>RNA-Seq</td>
<td>Whole testes of mice at PNDs 7, 14, 17, 21, and 28</td>
<td>rRNA removal with rRNA-specific probes</td>
</tr>
</tbody>
</table>

[www.reproduction-online.org](http://www.reproduction-online.org)
EST libraries were also used to study different transcriptomes of different germ cell stages during spermatogenesis by means of 3'-processing sites. EST tags were a few hundred base pair long sequences from either end of transcripts in a cDNA library. It is particularly useful to detect RNA isoforms of the same gene, yet, with a relatively strong background noise and vector contamination issue. Liu et al. (2007) built EST libraries from primitive type A spermatogonia, pachytene spermatocytes, round spermatids, and Sertoli cells of mice, and focused on the 3' UTR of the transcripts. They found that even for the same transcript, the 3’ UTR location and length varied at different stages in spermatogenesis.

An important part of current transcription evidence on lncRNAs came from cDNA or EST libraries. Back then, ncRNAs were not the focus and so no IncRNA candidates were reported. Another limitation of these libraries was their bias towards RNA transcripts with poly-A signals. Although poly-A is very useful to remove rRNAs and tRNAs, this method also screened off potential transcripts without poly-A features, leaving the IncRNA landscape incomplete.

**Microarrays**

Based on genomic information gathered from previous cDNA and EST libraries, specific cDNA or oligonucleotide probes can be designed for known RNA transcripts. This gave rise to the microarray platform, which is a hybridization-based technology to detect the presence of labeled cDNA targets on immobilized high-density array of synthetic DNA probes. The detection coverage of microarray completely depends on the content of probes, which can target exons, promoters, 3’-UTRs, or the full genes.

There were a number of studies using microarray to investigate male germ cell development. In early studies, custom-made microarrays were used to detect gene expression (Sha et al. 2002, Tanaka et al. 2002). Later, commercial microarrays such as GeneFilters GF400 (Pang et al. 2003), Affymetrix Genechips that covered mouse genome using U74 series (Su et al. 2002, Shima et al. 2004, Waldman Ben-Asher et al. 2010), 430 series (Xiao et al. 2008), and human genome U95 series (Su et al. 2002) were applied in different experimental setups for studying spermatogenesis.

Early microarray platforms were biased towards the 3'-end of a transcript. A more comprehensive microarray platform, known as tiling microarray, provided an unbiased approach in analyzing the whole transcriptome. Unlike usual microarray, which covers only known transcripts from cDNA clones, probes in whole-genome tiling array cover the whole genome by evenly spaced probes along the genomic coordinates. Since the probes are designed in high-density and distributed evenly throughout the genome, all transcripts, both known and unknown, will be covered. We have demonstrated the use of the Affymetrix mouse tiling 1.1R microarray set to screen through transcriptomes of type A spermatogonia, pachytene spermatocytes, and round spermatids. In this study, rRNA was depleted using the biotinylated rRNA-specific probes approach, which can retain IncRNAs without poly-A tails for microarray analysis. Nearly a half of the transcripts we detected had not been characterized before, revealing a large hidden layer of molecular regulation (Lee et al. 2012 a).

To date, microarray still offers an ideal platform to access the whole transcriptome in a relatively quick and economical approach compared with sequencing-based methods. Although microarray may not be able to distinguish different gene isoforms and orientation of the transcripts, it does pinpoint the regions of expression in gene expression study or enriched targets in immunoprecipitation (IP) setup. These regions are then subject to subsequent low-throughput validation assay using qPCR or ChIP–PCR. Another main advantage of using microarray data to identify IncRNAs is the availability of public data repositories such as Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) or ArrayExpress (www.ebi.ac.uk/arrayexpress).

Researchers are able to access the original datasets for performing analysis tailored to their interest. Some of the Affymetrix microarray data can be reanalyzed with online programs such as ncFANs (Liao et al. 2011) to obtain the IncRNA expression data by a simple data retrieval procedure. Both databases also provide online tool options that allow expression analysis or target gene search without downloading the dataset.

Recently, microarrays with probes specified for IncRNA detection have become available commercially. These include Affymetrix Genechip 2.0 ST series and Arraystar Mouse Stringent LncRNA microarray, which are based on updated IncRNA genomic annotation resources such as RefSeq, ENCODE, UCSC Genes, NRED, NONCODE, and rRNAdb. These platforms allow direct measurement of multiple known IncRNA candidates. Wu’s group identified 8265 IncRNAs expressed in testis from either 6-day- or 8-week-old mouse, where 3025 showed difference in expression (Sun et al. 2013). At the same time, Yan’s group used a similar approach to investigate six time points in mouse testis development: embryonic day (E)12.5, E15.5, postnatal day 7 (PND 7), 14, 21, and adult (Bao et al. 2013). They identified over 1000 down- and over 1500 up-regulated IncRNAs at each time point when compared with the previous time point. Most IncRNAs displayed the highest expression after PND 14, suggesting that IncRNA expression is most active in the meiosis and spermiogenesis phases. Also, both studies detected a significant portion of differentially expressed IncRNAs overlapping with small RNAs, suggesting possible precursor or regulator relationships. Furthermore, both studies observed a close association between IncRNAs and epigenetic regulation. Wu’s group found IncRNA expression is correlated with epigenetic
modification marks H3K4me3 and H3K27me3 (Sun et al. 2013). Yan's group further showed by cross-link RNA IP experiment that 11 IncRNAs interact with one or multiple factors of chromatin-remodeling complexes, such as EZH2 and LSD1 (Bao et al. 2013). Taken together, these data suggest that epigenetic regulation by IncRNAs may be involved in chromatin remodeling during spermatogenesis.

Serial analysis of gene expression
The serial analysis of gene expression (SAGE) is a quantitative sequencing technique to assess the abundance of individual transcript in cells before the advent of next-generation sequencing (NGS). Compared with microarrays, it is an open system and the results can be reanalyzed with updated genomic annotations. It also provides transcript orientation information and allows scalable library construction.

In a typical SAGE protocol (Velculescu et al. 1995), all transcripts with poly-A tails are pulled-down and the abundance of a 14-bp tag sequence starting from the 3′-most restriction site of a tagging enzyme, typically BsmFI, which recognize a tagging sequence of CATG, are determined. After mapping the tag sequences with sequencing mapping data such as SAGEmap (http://www.ncbi.nlm.nih.gov/SAGE/), the expression level of each gene can be calculated based on the number of unique tags present in the library. This technique allowed researchers to obtain the expression information of all poly-adenylated, coding, and noncoding transcripts.

Mori's team (Yao et al. 2004) and Forejt's team (Divina et al. 2005), each generated a SAGE library of mouse testes to study reproductive aging and chromosome clustering of genes, while Chan's group (Wu et al. 2004) built a cell-type specific SAGE database on type A spermatogonia, pachytene spermatocytes, and round spermatids of mouse. These data provided unbiased transcriptome information for coding and noncoding gene expression analyses. Non-annotated transcripts with poly-A signals will also be detected by this sensitive method. However, again, this method is weak in distinguishing isoforms of genes. Moreover, the resolution of SAGE is limited by the small number of base of a tag, which often led to multiple alignments for the same SAGE tag.

RNASeq
With the availability of NGS, transcriptome can be analyzed by a technique known as RNASeq. It provides advantages on speed, sensitivity, resolution, and coverage compared with the Sanger sequencing method. Driven by the falling cost of NGS, RNASeq has become a popular approach in transcriptome analysis and in IncRNA identification. Thousands of IncRNAs were identified in recent human and mouse genome projects (Derrien et al. 2012). Each run of RNASeq generates billions of reads, with typical lengths, of 30–400 bp, depending on the technology used. The number of reads mapped to a location directly reflects the expression level and splicing status of transcripts, facilitating intersample comparison.

There are a number of recent RNAseq studies on testis and sperm development, aiming to investigate the global transcriptomic dynamics in male germ cell development. Laiho et al. (2013) used RNA from whole testis on PNDs 7, 14, 17, 21, and 28 for analyzing transcriptomic changes in the first wave of spermatogenesis. By comparing two adjacent time points, they discovered 17–46 down-regulated and 16–605 up-regulated IncRNAs in each transition. However, the functions of those IncRNAs were completely unknown. With a similar approach, Gong et al. (2013) performed RNASeq on testis samples of 6-day-, 4-week-, and 10-week-old mice. Their main emphasis was on global protein-coding genes, revealing that the expressions of many stage-specific marker genes and genes for specific developmental processes, such as self-renewal and differentiation, largely coincided with the RNASeq or microarray results.

Schmid et al. (2013) focused on the splicing landscape during the meiotic phase of mouse spermatogenesis, by comparing the transcriptome of testis from PNDs 6 to 21, when the first-round meiosis begins. They identified 150 splicing switches which affected the isoform formation of protein-coding genes, and a number of exons highly expressed only during meiosis, but not at earlier stage or in other tissues. Although not reported, it is reasonable to expect that IncRNAs may also be regulated by similar alternative splicing mechanisms at a specific stage, and this would be an interesting topic for further study.

To discover the potential regulation of a novel epigenetic modification in spermatogenesis, Gan et al. (2013) performed a combined analysis on RNASeq and 5-hydroxymethylated cytosine (5hmC) profiling by chemical labeling and sequencing, showing that 5hmC status in different genomic regions changes continuously during eight selected stages of mouse spermatogenesis. They concluded some coding genes accommodated with abruptly up-regulated 5hmC signal were highly expressed in the late stages, suggesting that 5hmC may be involved in gene activation. This mechanism is also expected in the regulation of IncRNAs.

The RNASeq studies not only provided high-resolution transcriptome data, but also led to novel insights for IncRNAs identification in spermatogenesis. These high-throughput data are available for public access. Therefore, the data can be reanalyzed and mapped to various reference annotation databases such as RefSeq, Ensembl, Noncode, Fantom, and fRNAdb for IncRNAs prediction.

Given the abundance of expression evidence, the results from different studies can be combined to improve the prediction confidence and compliment the limitations of a specific platform. For instance, stage- and
process-specific lncRNA transcripts could be retrieved from multiple expression platforms by applying signal cutoffs and Venn diagram approach. This approach is also essential to allow de novo lncRNA prediction when the lncRNA annotation is not known. The completeness of a lncRNA transcript could be predicted with supporting regulatory signals, such as presence of RNA polymerase II, CpG island, CAGE tags, ChIP-Seq mapping data on various transcription factors in the promoter region, and polyA signals in the 3'-UTR.

**Functional lncRNAs in spermatogenesis**

Despite many potential lncRNAs identified in male germ cell development, only a few have been functionally annotated and characterized. Some were identified by the high-throughput approach described above, while others were identified by the de novo annotation approach. The expression patterns of these novel lncRNA candidates in spermatogenesis are illustrated (Fig. 3). Reviewing the discoveries and regulatory roles of these candidates may bring insights for future lncRNA identification.

**Meiotic recombination hot spot locus (Mrhl)**

Meiotic recombination hot spot locus (Mrhl) RNA was identified when Nishant et al. (2004) were focusing on a 17.2 kb fragment of a novel meiotic recombination hot spots in mouse chromosome 8. Mrhl is a 2.4-kb mono-exonic lncRNA located in the nucleus. Subsequent study suggested that Mrhl RNA regulates spermatogenesis potentially by two molecular mechanisms. The first mechanism was the cleavage of Mrhl by Drosha to form an 80-nucleotide RNA intermediate (Ganesan & Rao 2008). Evidence indicated that both RNAs were localized in nucleoli of the GC1 spermatogonia cell line, suggesting their possible interactions with chromatin (Ganesan & Rao 2008). Besides, Mrhl could inhibit the Wnt signaling pathway through interacting with p68 (Arun et al. 2012). However, the functional effects on spermatogenesis of this lncRNA remain elusive, and may not be easily deciphered without using a knockout animal.

**HongrES2**

HongrES2 is a 1588-bp lncRNA and expressed specifically in the cauda region of epididymis (Ni et al. 2011). It was detected at a constant level in rats aged from 30 days, when the first wave of spermatogenesis had just completed, to 450 days. This expression pattern perfectly coincided with its reported function in regulating sperm maturation.

The transcript is preferably localized in nucleus, where it is processed to a 23-bp miRNA-like small RNA, named as mil-HongrES2, which inhibits the expression of an epididymis-specific protein CES7 and hence its cholesterol esterase activity. Moreover, over-expression of mil-HongrES2 results in retarded sperm capacitation as indicated by the global tyrosine phosphorylation level (Ni et al. 2011), suggesting that an endogenous low level of the lncRNA expression is a gatekeeper for normal sperm maturation process in epididymis.

**Testis-specific X-linked (Tsx)**

Testis-specific X-linked (Tsax) was thought to be a protein-coding gene, but has recently been shown to be a lncRNA (Anguera et al. 2011). The same group also found that Tsx is specifically expressed in pachytene spermatocytes, but not in spermatogonia or round spermatids, suggesting a regulatory role in germline meiotic division. TUNEL assay analysis showed that Tsx knockout mice had an abnormally high proportion of pachytene spermatocytes that had undergone apoptosis, suggesting that these lncRNAs are somehow critical for progression of meiosis. Detailed mechanism remains open for investigation.

**Dmrt1-related gene (Dmr)**

Dmrt1-related gene (Dmr) is another interesting example of a testis-specific functional lncRNA. It was discovered unintentionally when Zhang et al. (2010) tried to clone the Dmrt1 gene, and found a novel RNA isoform of unexpected size and sequence on the 3' half. Genetic mapping showed this mysterious 3'-half was actually coded by a gene, named therefore Dmr, in
Long noncoding RNAs in spermatogenesis

R139

chromosome 5, while Dmrt1 is a gene in chromosome 19. In other words, Dmr is able to form a trans-splicing RNA isoform with Dmrt1. The formation of this chimeric mRNA disrupted the coding region of Dmrt1 and replaced the 3′-UTR of Dmrt1, both of which resulted in down-regulation of normal Dmrt1 protein. Since Dmrt1 protein is an essential transcription factor to promote spermatogonial development by up-regulating Sohlh1 and also prevents premature meiosis in spermatogonia by repressing Stra8 (Ottolenghi et al. 2000, Agbor et al. 2013), Dmr suppression on Dmrt1 is also possibly involved in the switching between mitosis and meiosis of the development germ cell.

Spga-IncRNAs

To demonstrate the idea of integrating high-throughput data on IncRNA identification, we used previously published SAGE data (Chan et al. 2006, Lee et al. 2006, 2009) and whole-genome tiling microarray data (Lee et al. 2012a). Both platforms were previously used to study the transcriptomes of three critical germ cell stages, namely type A spermatogonia, pachytene spermatocytes, and round spermatids, where the initial focus was on coding genes with differential expression.

With this approach, we identified 50, 35, and 24 single-exonic IncRNAs with annotation specifically expressed in type A spermatogonia, pachytene spermatocyte, and round spermatids respectively (Lee et al. 2012b). Further validated by qPCR, northern blot, and RACEs, two spermatogonia-specific IncRNA candidates, known as Spga-IncRNA1 and 2, demonstrated significant differentiation inhibitory effects in in vitro model, suggesting they may be important in maintaining stemness of spermatogonia.

Conclusion and perspectives

Spermatogenesis is a complex and highly regulated biological process. Based on our current view in ncRNA biology, we suspect that the current understanding of IncRNA regulations in spermatogenesis is very incomplete. Although advances in experimental methods and public IncRNA annotation resources allow rapid discovery of potential IncRNA candidates, the functional annotation of IncRNAs in spermatogenesis remains incomplete. The generation of various transcriptome data will definitely provide better expression evidence for IncRNA prediction and annotation. While the use of public IncRNA annotation could be a straightforward approach for rapid IncRNA identification in spermatogenesis, there is still a subset of transcripts that are not covered by the public annotations. One possibility is that these transcripts may be cell type or process specific. Therefore, applying de novo IncRNA prediction would be a better way to identify and characterize these important novel candidates. We anticipate IncRNAs will be an attractive direction in male germ cell research, and more functional IncRNAs will be characterized when more high-throughput data, refined bioinformatics pipelines, and regulatory evidence from ENCODE, ontology databases, and protein–RNA binding models become available.

Declaration of interest

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

Funding

This work was supported by funds from the Hong Kong Research Grant Council (grant numbers 469711 and 468312) and CUHK Direct Grants (MD10783 and MD11449), Faculty of Medicine, The Chinese University of Hong Kong.

References


Received 18 November 2013
First decision 19 December 2013
Revised manuscript received 14 January 2014
Accepted 21 January 2014