# Regulatory mechanisms of non-coding RNAs during zebrafish embryogenesis

CHIRAG NEPAL

Dissertation for the degree philosophiae doctor (PhD)
At the University of Bergen

2013

## Scientific environment

This work has been performed in

**Computational Biology Unit (CBU)**

**Uni Computing**

and

SARS International Centre for Marine Molecular Biology

Chirag Nepal is affiliated with

**Department of Informatics**

Good science is all about following the data as it shows up letting yourself be proven wrong, and letting everything change while you’re working on it.

Rebecca Skloot

# Acknowledgements

I would sincerely thank everybody who have supported and helped me during my doctoral studies in the last $5^{1/2}$ years. Here is a non-exhaustive list of those fantastic people.

First of all, I would like to thank my supervisor Boris Lenhard for taking me as your Ph. D student. Thank you for introducing me into the world of non-coding RNAs and providing exciting topics and data to work on. Your guidance has been highly invaluable. Your deep and sound knowledge on broad range of topics, dedication, and enthusiasm towards work have been a great source of inspiration to me. Above all, thank you for teaching me that Ph.D. is all about making a candidate to be able to create his own hypothesis and work towards proving it. Thanks to Inge Jonassen for agreeing to be my co-supervisor and always being friendly and helpful. I would like to thank Vidar Martin Steen for offering a job before the completion of Ph.D..

I would like to thank our past and present members of Lenhard group, Pär, David, Christopher, Reidar, Altuna, Jan Christian, Supat, Yogita, Gemma, Chandu, Vanja, Xianjun and Ying, as well as short term members: Nathan, Sara, Melis, Joao, and Vedran for creating a nice and friendly environment in the group. Thanks for all those simulating discussions on both scientific and non-scientific topics, your valuable feedbacks and social nights out. Special thanks to Christopher and Nathan for providing feedback on the first draft of my thesis. I would like to thank our collaborators, Ferenc Mueller, Piero Carninci and Yavor Hadzheiv, and others who were involved in ZEPROME consortium. Thanks in particular to Ferenc Mueller, for countless numbers of stimulating discussions over skype and your continuous help.

Thanks also to everyone else who made CBU a pleasant and stimulating research environment and a great place to work. Thanks also to all the group members of Martin's lab for providing a wonderful environment for working. I would also like to thank all my friends that I meet then here in Bergen and Fantoft, with whom I have spend great $5^{1/2}$ years of my life.  I would like to include all of your names, but the list became exhaustive.

Lastly, I wish to thank my to parents and sisters for their continuous love and support during this time.

## List of publications included in the thesis

**Paper I**

Dynamic regulation of coding and non-coding transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis

**Chirag Nepal**[*], Yavor Hadzhiev[*], Christopher Previti[*], Vanja Haberle, Nan Li, Hazuki Takahashi, Ana Maria S. Suzuki, Ying Sheng, Rehab F. Abdelhamid, Santosh Anand, Jochen Gehrig, Altuna Akalin, Christel E.M. Kockx, Antoine A.J. van der Sloot, Wilfred F.J. van IJcken, Olivier Armant, Sepand Rastegar, Uwe Strähle, Elia Stupka, Piero Carninci, Boris Lenhard, Ferenc Müller

Manuscript submitted

* Contributed equally

**Paper II**

Transcriptional, post-transcriptional and chromatin associated regulation of pri-miRNAs, pre-miRNAs and moRNAs in zebrafish development

**Chirag Nepal**, Yavor Hadzhiev, Vidar M Steen, Piero Carninci, Ferenc Müller, Boris Lenhard

**Paper III**

Genome wide characterization of snoRNAs in zebrafish reveals their co-transcription with coding and long non-coding host genes by non-canonical transcription initiation

**Chirag Nepal**, Estefanía Tarifeño-Saldivia, Wei Deng, Pål Sætrom, Christopher Previti, Bernard Peers, Ferenc Müller, Boris Lenhard

# List of publications not included in the thesis

**Paper I**

**The genome sequence of Atlantic cod reveals a unique immune system.**

Bastiaan Star, Alexander J. Nederbragt, Sissel Jentoft, Unni Grimholt, Martin Malmstrøm, Tone F. Gregers, Trine B. Rounge, Jonas Paulsen, Monica H. Solbakken, Animesh Sharma, Ola F. Wetten, Anders Lanzen, Roger Winer, James Knight, Jan-Hinnerk Vogel, Bronwen Aken, Øivind Andersen, Karin Lagesen, Ave Tooming-Klunderud, Rolf B. Edvardsen, Kirubakaran G. Tina, Mari Espelund, **Chirag Nepal**, Christopher Previti, Bard Ove Karlsen, Truls Moum, Morten Skage, Paul R. Berg, Tor Gjøen, Heiner Kuhl, Jim Thorsen, Ketil Malde, Richard Reinhardt, Lei Du, Steinar D. Johansen, Steve Searle, Sigbjørn Lien, Frank Nilsen, Inge Jonassen, Stig W. Omholt, Nils Chr. Stenseth & Kjetill S. Jakobsen.

Nature 477, 207-210, 08 Sep 2011

**Paper II**

**Two independent transcription initiation codes overlap on vertebrate core promoter**

Vanja Haberle, Nan Li, Christopher Previti, **Chirag Nepal**, Jochen Gehrig, Xianjun Dong, Altuna Akalin, Yavor Hadzhiev, Wilfred van IJcken, Uwe Strähle, Piero Carninci, Ferenc Müller, Boris Lenhard

Manuscript submitted

# Table of Contents

## List of abbreviations

| | |
|---|---|
| CAGE: | Cap analysis of gene regulation |
| ENCODE: | Encyclopedia of DNA Elements |
| FANTOM: | Functional Annotation of Mouse |
| MBT: | Mid-blastula transition |
| MZT: | Maternal-zygotic transition |
| HCNE: | Highly conserved non-coding elements |
| H3K4me3: | Trimethylation of histone 3 on lysine 4 |
| ncRNAs: | non-coding RNAs |
| lncRNAs: | long non-coding RNAs |
| miRNAs: | microRNAs |
| PRC2: | Polycomb Repressive Complex 2 |
| snoRNAs: | small nucleolar RNAs |
| TSSs: | Transcription start sites |
| UTRs: | Untranslated region |

# Abstract

For many years, RNAs were thought to be intermediate products between DNA and protein. The discovery of RNA interference (RNAi), a regulatory process that uses small non-coding RNAs to regulate gene expression at the post-transcriptional level, changed our view about RNAs. However, the discovery of microRNAs was the realization of RNAs as the regulatory elements. In recent years, many high-throughput sequencing studies have identified hundreds to thousands of various kinds of non-coding RNAs. The existence and biological relevance of these non-coding RNAs detected in large-scale analysis of human tissues have not yet been characterized in a vertebrate animal in vivo. To gain insight into the existence and biological relevance of these non-coding RNAs in vertebrate animal in vivo, we have set out to generate the first global description of TSS usage during key stages of vertebrate embryonic development at single nucleotide resolution. We have coupled CAGE maps to protein-coding and non-coding transcripts by RNA sequencing (providing a quantitative description of TSS usage on a genome scale) and anchored to posttranslational histone modifications (H3K4me3) by ChIP sequencing.

We reveal an extraordinary dynamics of promoter usage that takes place during development of the vertebrate embryo. We showed that the onset of transcription and subsequent differentiation of the embryo is characterized by the developmentally regulated appearance of 5'-ends of intragenic RNAs on many genes, and of an entire hitherto unknown layer of RNA species overlapping known genes and having specific signatures occurring in exons, introns and 3'-UTRs of developmentally active genes. We characterize the pervasive production of intragenic processed RNAs including exonic and intron-5' end specific RNAs and provide the first indication for the biological processes in which they may function. Notably, intron 5' end associated non-coding RNAs are active zygotically and restricted to genes that encode RNA processing and the splicing proteins in both fish and human. We demonstrated evidence that exonic RNAs are produced by a non-canonical posttranscriptional mechanism independent of the gene 5' end. We show the initiation landscape and developmental dynamics of lincRNAs; we show the evolutionary conserved process of developmentally regulated posttranscriptional processing of lincRNAs into intragenic RNAs, which demonstrate the utility of zebrafish in studying mammalian lincRNA processing.

The main aim of this work was to provide a (currently non-existent) annotation of

miRNA promoters and characterize their common characteristics features at transcription, post transcription and chromatin level. We describe the first genome-wide identification of miRNA promoters in zebrafish active during the early embryonic developmental stages. We identified a small number of maternally transcribed miRNAs, one MBT specific miRNA and the majority that are zygotically transcribed. We report the first evidence of moRNAs in zebrafish and pufferfish that were previously reported in human and Ciona intestinalis. We show evidence for unexpected enrichment of pre-miRNA sites with promoter-associated histone modification marks (H3K4me3 and H2A.Z) suggesting chromatin regulation and potential involvement of transcription machinery in pre-miRNA processing, suggesting co-transcriptional splicing of pre-miRNAs and pri-miRNA.

We have provided a catalogue of intermediate-sized non-coding RNAs in zebrafish, by making RNA library enriched for intermediate-sized (50-500 nt) non-coding RNAs, collected from zebrafish larvae (5-7 days post fertilization). In particular, we validated most annotated snoRNAs and identified few hundreds of novel snoRNAs making the most comprehensive annotations of zebrafish snoRNAs. Host genes for most snoRNAs showed no evidence for independent transcription of snoRNAs, suggesting they are co-transcribed by host genes. Interestingly, host (coding and non-coding) genes require non-canonical transcription initiation machinery, as indicated by TCT initiation signals, that is associated with translation machinery. 5'-end of many snoRNAs overlaps with CAGE 5'-ends, suggesting either they are capped or undergo post-transcriptional modification, which is also evolutionary conserved in human snoRNAs. Small RNAs derived from snoRNAs are generated from most snoRNAs and provide first evidence of sd-snoRNAs produced in oocytes, suggesting their potential importance during early embryogenesis.

# 1. Introduction

Non-coding RNAs (ncRNAs) are mature products of genes that are transcribed but not translated into proteins. The size of these non-coding RNAs can range from as small as 18-22 nucleotides (nt) to tens of kilobases (KBs). Non-coding RNA genes can be found within introns of protein-coding genes, proximally to the promoters of such genes, or in intergenic regions as defined with reference to protein-coding genes. In general, the functional specificity for many non-coding RNAs is conferred by a secondary structure or a small sequence that binds to its target through complementary base pairing in the 3' untranslated region (UTR). Here, the RNA transcripts themselves are the functional end products rather than an intermediate RNA. While we can describe these non-coding RNAs as "regulatory RNAs", it is far from certain if all transcribed RNAs are functional and we are yet to understand all their regulatory roles. The genome-wide discovery of thousands of such regulatory RNAs in mammals, vertebrates, and plants has provided new insights into their contribution to gene regulation, as well as the forms in which genetic information is interpreted.

The central dogma of molecular biology states the direction of flow of genetic information: DNA is transcribed into messenger RNAs, which serves as the template for protein synthesis (Brenner et al. 1961; Jacob and Monod 1961). Brenner, Jacob, and Meselson confirmed this model by isolating the unstable RNA carrier of information, distinct from ribosomal (rRNA) and transfer RNAs (tRNAs), and disproved the hypothesis stating that each gene has a unique ribosome responsible for synthesizing its protein product. RNA was thereafter recognized as the information carrier for protein construction from the genes to the rest of the cell (mRNA): it presented the correct amino acid to the growing protein (tRNA) and facilitated the creation of new proteins at the ribosome (rRNA). The concept of functional non-coding RNA dates back to the initial days, as it was shown that some RNAs are transcribed but not translated; however, it was thought that their function was limited to coordinating genes and protein production, and that they did not have regulatory roles on their own.

For the next two decades, our knowledge of non-coding RNAs was primarily limited to those involved in protein synthesis. In the early 1980s, different classes of functional RNAs were discovered, which led to the realization that non-coding RNAs have an important role in gene regulation. Various classes of non-coding RNAs were discovered during this time, such as RNase P, a ribozyme, required for the maturation of tRNAs (Stark et al. 1978); the ‘U’ RNAs, which assist in splicing mRNAs; and small nucleolar RNAs, which guide modifications of other RNAs (reviewed in (Zieve 1981; Matera et al. 2007). However, the real surprise was the discovery of microRNAs (miRNAs) (Ambros 1989), which were able to regulate genes using a highly specialized and efficient cellular mechanism. This discovery turned out to be the beginning of the realization that non-coding RNAs are major players in gene regulatory networks. The first miRNA gene (lin-4), discovered in *Caenorhabditis elegans*, was identified as a regulator of developmental genes lin-14 and lin-28 (Ambros 1989). The mechanism by which miRNAs were found to control the gene expression of lin-14 was via the gene’s 3’ untranslated region (UTR) (Wightman et al. 1991). An effort to clone the lin-4 gene identified the lin-4 gene product as a 21 nt transcript with complementarity to the lin-14 3’ UTR (Lee et al. 1993), which was apparently the first microRNA. Initially, lin-4 gene was thought to be an exception, as no other miRNAs with similar functionality were found in the years to follow. It was not until 2000 that a second miRNA gene, let-7, was identified in *C. elegans*, with a similar functional mechanism (Reinhart et al. 2000). Analysis of sequence conservation revealed that these two miRNAs were highly conserved, and detectably transcribed in other species ranging from *C. elegans* to *D. melanogaster* and humans (Pasquinelli et al. 2000). These findings suggested that miRNAs were not just a nematode peculiarity, but rather a RNA type widespread across all species. Since then, thousands of microRNA genes have been found in organisms including vertebrates, invertebrates and plants (Lim et al. 2003) (Reinhart et al. 2002; Jones-Rhoades and Bartel 2004).

Genome-wide prediction of miRNAs based on sequence homology, secondary hairpin structure and evolutionary conservation suggested thousands (~15,000) of genomic segments that were predicted to form stem loops (Lim et al. 2003). However, Lim et al had predicted only 188 candidates to be true miRNA candidates, and suggested maximum cap of about 255 miRNAs in human genome. To provide the scientific community with high-confidence miRNAs based on experimental evidence, a dedicated database called miRBase was established (Kozomara and Griffiths-Jones 2011) and many of miRNAs stored there have been implicated in the regulation of basic biological functions (reviewed in (Bartel 2004)).

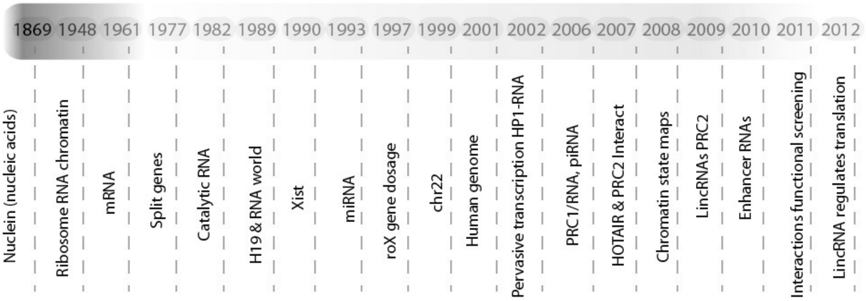

Figure 1. Timeline of the discovery of functional RNAs. The figure has been adapted from a recent review (Rinn and Chang 2012)

Subsequently, studies inspired by the discovery of miRNAs were able to identify novel small RNA classes using high-throughput sequencing technologies (Ruby et al. 2006). Among these classes of non-coding RNAs, two of the most well characterized and studied were *short interfering RNAs* (siRNAs) and *piwi-interacting RNAs* (piRNAs). Short interfering RNA are 21-25 nt RNAs usually derived from exogenous RNA and are believed to be part of a defense system against foreign RNA (Meister and Tuschl 2004). Piwi-interacting RNAs is a novel class of RNA that were first identified in germlines and are associated with Piwi-subclade member of the Argonaute protein family (Girard et al. 2006; Grivna et al. 2006).

Genome-wide evidence of novel classes of small RNAs with developmentally regulated patterns across various cell lines and tissues asserted the importance of small RNAs and indicated a much more important regulatory role for them than previously anticipated. One of the dominant classes of small RNAs identified was in promoter regions (Kapranov et al. 2007; Seila et al. 2008; Taft et al. 2009a), termed promoter-associated RNAs (pasRNAs) or transcription-initiation RNAs (tiRNAs). But evidence of genomic regions from which small RNAs are produced was not limited to just promoters, but also included exons, introns, exon-exons junctions (Carninci et al. 2006; Fejes-Toth et al. 2009; Mercer et al. 2010), splice-sites (Taft et al. 2010; Valen et al. 2011), 3'UTR (Mercer et al. 2011) and intergenic regions. The detailed biogenesis, mechanism and functional aspect of these RNAs has been reviewed (Kim et al. 2009).

Unlike the small non-coding RNAs described above, two long non-coding RNAs (lncRNAs), Xist and Air were identified long before miRNAs were identified, as shown in Figure 1. At the time, studies based on DNA microarrays had revealed that most of the transcribed regions in the genome did not code for proteins (Kapranov et al. 2002; Rinn et al. 2003; Bertone et al. 2004; Kampa et al. 2004). As the function of these transcribed regions was not evident, it was assumed that these transcripts were just a by-product of transcription, rather than a functional product. A large-scale effort taken by the FANTOM (Functional Annotation of Mouse) consortium to sequence the full-length cDNAs of both mouse and human revealed genome-wide evidence of transcribed RNAs, the majority of which did not code for proteins (Carninci et al. 2005; Katayama et al. 2005). While it was evident that the mammalian genome is pervasively transcribed, forming a complex interlaced architecture (Katayama et al. 2005; Engstrom et al. 2006), skepticism still remained about the functionality of these non-coding RNAs (Willingham et al. 2005). The ENCODE (Encyclopedia of DNA Elements) pilot project suggested at least 74% of the assessed region (1% of genome) were transcribed, as assessed by two or more different technologies (Birney et al. 2007). This number was later increased to 80% (Bernstein et al. 2012). A detailed analysis on sequence conservation revealed only small stretches of highly conserved non-coding RNA elements (Pang et al. 2006). On the other hand, even poorly conserved non-coding RNAs possess an imprint of purifying selection on their promoter and primary sequence (Ponjavic et al. 2007).

However, it was not until the identification of HOTAIR (Rinn et al. 2007) that a potential mode of regulation of such non-coding RNA might be in regulating the epigenetic landscape by modifying chromatin structures. HOTAIR is located in HOXC locus and represses transcription of HOXD locus in trans by interacting with the Polycomb Repressive Complex 2 (PRC2) and required for PRC2 occupancy and trimethylation of lysine-27 on histone 3 of HOXD locus. Soon, more publications detailed the functional roles of lncRNA and its association with various chromatin structure (Khalil et al. 2009; Gupta et al. 2010; Mondal et al. 2010). As of now, lncRNAs are known to form ribonucleoprotein complexes with various chromatin regulators and then target their enzymatic activities to appropriate locations in the genome (Rinn and Chang 2012).

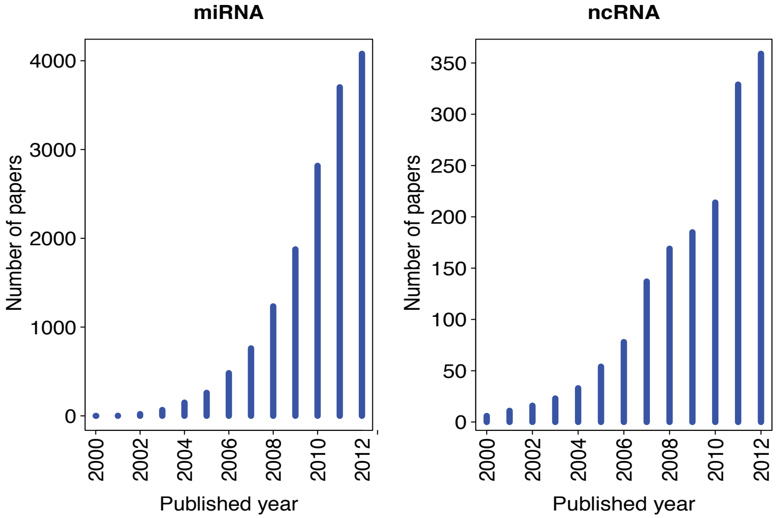

Figure 2: The rise in number of papers related to miRNAs and non-coding RNAs. The number indicates the number of published papers (which are PubMed indexed) per year with keywords ("miRNA" or "microRNA" or "micro RNA") and ("non-coding RNA" or "ncRNA" or "non-protein-coding RNA" or "lncRNA" or "lincRNA" or "long non-coding RNAs"). The data for this figure were extracted from PubMed during the end of October 2012.

All these studies revealed that the genome is pervasively transcribed, with only a fraction (2%) of transcripts representing coding exons and another 20% consisting of other gene components such as the 5' and 3' UTRs as well as the introns. While this may be partially accounted for by unannotated protein-coding genes, the vast majority of these transcripts are indeed non-coding (Carninci et al. 2005; Katayama et al. 2005; Birney et al. 2007; Bernstein et al. 2012). Given the recent discovery of new types, the list of known non-coding RNAs as well as our understanding of their importance is likely to continue to grow. Extrapolating their regulatory roles for the majority of these transcripts remains challenging and requires more detailed analysis on each specific locus.

Non-coding functional elements such as *cis*-regulatory modules (promoters and enhancers) and non-coding RNAs have been the center of recent attention of recent genomics research. The functional analysis and annotation of these features is expected to shape genomics research in the forthcoming decade. The rise in the number of research papers related to miRNAs and non-coding RNAs published per year is a testimony to the realization that ncRNAs are important functional elements (Figure 2). Most of the studies using high-throughput data to study miRNAs, non-coding RNAs and other regulatory elements are based on *in vitro* cell culture work. In order to take advantage of the zebrafish (*Danio rerio*) model organism, the zebrafish transcriptome and promoterome (ZEPROME) consortium was formed with the aim of elucidating the developmental transcriptional regulation codes of both coding and non-coding RNA in the context of a developing vertebrate embryo. Our work is pioneering in the mapping of functional elements of transcription initiation in the complexity of the vertebrate animal during development and yielded the most comprehensive and highest resolution genome-wide map of coding and non-coding RNA promoter usage throughout embryonic development. Furthermore, this dataset is the first of its kind for any animal model that covers all key stages from fertilization, through maternal zygotic transition to body patterning and organogenesis. It represents an important genomic annotation resource for the discovery of novel features of non-coding RNAs and *cis*-regulatory codes present in vertebrates. Since this thesis deals only with non-coding RNAs, I will mainly focus on various types of non-coding RNAs that we have identified in this study.

## 1.1. Genomics era

Our current understanding of RNA biology is the result of a series of landmark events. One of the most important was the completion of the first draft of the human genome (Lander et al. 2001; Venter et al. 2001), which was a giant leap in the field of genomics. The availability of the human genome coupled with advances in massively parallel sequencing technologies is considered the beginning of modern-day genomics era. One of the major outcomes of the human genome project was to reveal that only ~2% of the genome is composed of protein-coding genes, and that the estimated gene count was about ~21500. In contrast two previous (only a year earlier) large scale studies based on expressed sequence tag (EST) data, had estimated the number of coding genes in human genome to be between 35,000 (Ewing and Green 2000) and 120,000 (Liang et al. 2000).

Already around the time of the completion of the draft human genome, some studies suggested that most of the human genome was transcribed (Wong et al. 2001), a notion that was later supported by large scale consortium studies such as ENCODE (Birney et al. 2007; Bernstein et al. 2012) and FANTOM (Carninci et al. 2005). Such large-scale studies have now been extended to other species - i.e. to *Drosophila* and *C. elegans* through modENCODE (Gerstein et al. 2010; Roy et al. 2010) project that further supports the notion of genome being pervasively transcribed is an evolutionarily widespread phenomenon. The ENCODE pilot project used different technologies to investigate the transcriptional landscape on 1% of the genome (44 genomic loci) across various cell lines, revealed a staggering 74% of the nucleotide positions examined were biochemically active. The outcome of the ENCODE pilot project was a testimony to other prior work that had reported a plethora of developmentally regulated transcribed RNAs (Birney et al. 2007). The recently concluded second (whole-genome) phase of the ENCODE project, which had made a courageous effort to catalogue the transcriptional landscape on a genome-wide scale across various human cell lines, reported at least of 80% of the bases in human genome are transcriptionally active (Bernstein et al. 2012). The functional importance for many of these non-coding RNA is still unclear, and the identification and functional characterization of these non-coding RNAs is currently among the most important and interesting challenges in genomics.

## 1.2 Conserved non-coding elements

After the completion of the human genome, the availability of the draft genomes of mouse (Waterston et al. 2002), rat (Gibbs et al. 2004) and many other vertebrates allowed a comparison of multiple genomes to identify regions that are evolutionary conserved. Comparative genomics allowed us to assess the rate of purifying selection acting on the different segments of genes (promoter, exons, introns, UTRs), and intergenic regions, which revealed an unexpectedly high degree of conservation outside the coding regions (Bejerano et al. 2004; Boffelli et al. 2004; Sandelin et al. 2004). These regions were described as ultraconserved region (Bejerano et al. 2004; Sandelin et al. 2004), conserved non-coding elements (Woolfe et al. 2005), or highly conserved non-coding elements (HCNEs)(Engstrom et al. 2008). The use of different thresholds in terms of minimum number of conserved bases and percentage identity along the conserved bases, and different species used for comparison with human sequence, resulted in different estimates regarding the number of HCNEs (Bejerano et al. 2004; Sabarinadh et al. 2004; Sandelin et al. 2004). Nevertheless, what these elements have in common is that they tend to cluster around developmental regulators regions, including in gene deserts, and that those clusters that can span several megabases each (Sandelin et al. 2004; Woolfe et al. 2005). Many HCNEs were identified to be functional enhancers able to drive the expression of (nearby or distal) target genes, in both vertebrates (de la Calle-Mustienes et al. 2005; Pennacchio et al. 2006; Kikuta et al. 2007) and invertebrates (Glazov et al. 2005; Papatsenko et al. 2006; Engstrom et al. 2007). Despite many HCNEs being capable of driving the target gene's expression in a reporter gene assay, almost an equal number of examined HCNEs were unsuccessful (Pennacchio et al. 2006). Genome-wide analysis of enhancer studies using ChIP-seq revealed tissue specific (Visel et al. 2009b) and stage-specific (Bogdanovic et al. 2012) enhancers. The temporal and spatial expression patterns of enhancers partly explain the inability of some HCNEs to function in a particular transgenic assay, which may be limited with respect to the developmental time points it covers. The data from the ENCODE and most recent FANTOM projects will shed light in choosing appropriate tissues or developmental time points. Recent studies have reported highly conserved non-coding RNAs are altered in human cancer (Calin et al. 2007), emphasizing the importance of these HCNEs in the regulation of human health.

One obvious question is, why are these non-coding elements so conserved and what functions (if any) are encoded in them? Two percent of the human genome codes for proteins and the remaining 98% comprise intron (intragenic) and intergenic regions. Before the completion of the human genome these intragenic and intergenic sequences were considered junk or selfish DNA (Ohno 1972; Orgel and Crick 1980), and were thought be genetically inert. The comparisons of multiple genomes showed that most conserved sequences were not coding sequences but rather non-coding sequences (Lindblad-Toh et al. 2011). Though highly conserved sequences are non-coding sequences, in general, the overall conservation of coding sequences is much higher than non-coding sequences. Later it was shown that the relative amount of non-coding sequence increases with complexity (Taft et al. 2007). Human (and other mammals) have higher ratio of non-coding sequences compared to coding sequences, while the number of coding genes remains similar. It has been speculated that the complexity in higher organisms may be inferred through these non-coding sequences. Highly expressed genes in the nervous system have large intronic sequences indicating the complexity of these brain specific genes might have acquired through gain of these non-coding sequences (Taft et al. 2007). As the genome sequence of more (distantly related) species are now available, a closer inspection revealed at least of 5% of human genome is under strong purifying selection (Lindblad-Toh et al. 2011). For the majority of these regions (excluding coding genes) under purifying selection, functions are not yet annotated.

## 1.3 Conserved RNA structures

The function of many non-coding RNAs is mediated through their secondary and tertiary structures. The availability of multiple genomes allowed researchers to identify conserved secondary structures, an indication of putative functional RNAs. Most known house keeping RNA types have conserved structures despite relatively low sequence conservation. The rationale behind the approach to identify these structured RNAs was: given any RNA sequence, how likely is it to have higher conservation, both at sequence and structure level, than one would expect by chance, and how stable would the secondary structures be? Conserved secondary structure is an act of purifying selection on the functional RNA that allows changes at the sequence level as long as the secondary structure and a small number of key residues are preserved. RNA sequences are often highly variable while maintaining structural conservation, often resulting in a substitution pattern. One of the first tools to predict genome-wide non-coding RNAs, qrna, is based on a stochastic context free grammar (SCFG) method to assess the probability that a pair of aligned sequences evolves under a constraint for preserving a secondary structure (Rivas and Eddy 2001). RNAs that are under long-time selection pressure to maintain secondary structure can be expected to have sequences more resilient to mutation (van Nimwegen et al. 1999). This in turn correlates with increased thermodynamic stability of the fold. It has been observed that functional RNAs are more stable than the structures formed by randomized sequences (Washietl and Hofacker 2004; Clote et al. 2005).

To accomplish this on a genome-wide scale, two different prediction tools, RNAz (Washietl et al. 2005) and Evofold (Pedersen et al. 2006), based on different approaches, were used to predict evolutionarily conserved secondary structures in human genome. RNAz calculates the probability that a multi-sequence alignment represents a conserved structured RNA by predicting the thermodynamic stability of a shuffled alignment. It measures thermodynamic stability of individual sequences and a structure conservation index obtained by comparing the folding energies of the individual sequences and the energy of the predicted consensus folding. Evofold uses a probabilistic model of RNA sequence and structure evolution, called phylogenetic-SCFG with structural and non-structural model, to evaluate how well a substitution pattern in an alignment matches a secondary structures annotation. It predicts the structure only if the segment of the alignment is better described by the structural model than the nonstructural model. Despite the different approach in their underlying mechanisms, both tools predicted tens of thousand of structured RNAs, located in all segment of genes and in intergenic regions. A collaborative effort between both groups, as a part of ENCODE pilot project, revealed strikingly low (around 25 %) overlap between prediction from both tools (Washietl et al. 2007). However, experimental validation of non-overlapping candidates confirmed the existence of 25% of the examined candidates from both tools. Similar results were obtained in *C. elegans* (Lu et al. 2011).

## 1.4 High-throughput sequencing technologies

In recent years traditional Sanger sequencing has largely been replaced by second-generation (high-throughout) sequencing. Among the best-known next-generation sequencing technologies are Illumina (Solexa) sequencing, Roche 454 pyrosequencing (Margulies et al. 2005), and Applied BioSystems SOLiD sequencing. The main advantage of next-generation sequencing is its ability to generate millions to billion of sequence tags per run by multiplexing the sequencing process, significantly lowering costs relative to standard dye-terminator methods. Each sequencing technology has its own advantages and limitations. The preference for any technologies is entirely dependent upon the resources (time, money and coverage) and the nature of the project. Along with the next-generation sequencing, many different techniques (for RNA and DNA sample preparation) have been developed, coupling the previously existing technologies to high-throughput genome-wide sequencing: RNA analysis (RNA-seq) (Nagalakshmi et al. 2008), histone ChIP-Seq (Barski et al. 2007), CAGE (Shiraki et al. 2003), 3P-Seq (Jan et al. 2011) and ribosomal profiling (Ingolia et al. 2009). I shall only discuss CAGE, RNA-seq, histone ChIP-Seq and small RNA sequencing as methods used to generate the data analyzed in this thesis.

### 1.4.1 Cap analysis of gene expression (CAGE)

Cap analysis of gene expression (CAGE) allows the quantification of gene expression and transcriptional profiling of transcription start sites (TSSs) usage by sequencing DNA tags from the initial 20 to 27 nucleotides of 5’ end of mature mRNAs selected by the presence of 5’ cap (Shiraki et al. 2003). CAGE libraries are constructed from full-length cDNAs selected through biotinylated 5’ cap. The capping procedure ensures the stability to the RNA and the CAGE protocol is therefore able to differentiate coding mRNAs and functional PolII-transcribed non-coding RNAs from the rest of cellular RNA and incomplete transcripts. Cap trapping is followed by sequencing the 5’ end of RNA and results in millions of sequence tags that can be mapped specifically to the genome.

CAGE tags are generally small, so to avoid ambiguous mapping only those tags mapping uniquely to the genome are analyzed further, except in special cases of analyses of transcriptional initiation from repetitive elements (Faulkner et al. 2009). The CAGE technology has a known bias, where an additional G is often added in the first base of tags. To overcome this problem, if the first base of CAGE tag starts with “G” and does not map to the genome, the first base is chopped off and remapped. CAGE tags with identical 5’ start sites are grouped into CAGE-tag starting sites (CTSSs) (Figure 3), whereas CTSSs that overlap on the same strand were merged to form transcript clusters (TCs) (Carninci et al. 2006). The extensive use of CAGE was first made in the FANTOM3 projects (Carninci et al. 2005; Carninci et al. 2006).

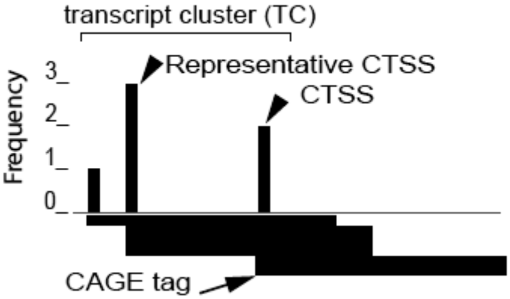

Figure 3. Schematic representation of clustering of CAGE tag starts sites. CAGE tags overlapping within 20 bases apart are clustered to form transcript cluster. 5’-ends of each CAGE tags are represented by vertical bars. Numbers on the y-axis represent number of unique CAGE tags mapping at that position which is used the expression level. Distance between the 5’-end of first Ctss and the last Ctss within an overlapping region determines the width of transcript cluster.

One of the major findings of FANTOM3 was that the vertebrate genome was a "transcriptional forest" giving rise to interlaced transcripts (Carninci et al. 2005; Engstrom et al. 2006). The CAGE method could for the first time characterize the vertebrate promoterome and TSS usage and dynamics at single nucleotide resolution. The analysis of TSS distribution made it possible to separate the vertebrate core promoters into two major classes of sharp and broad promoters (Figure **4),** where broad promoters were further characterized into multimodal or broad with dominant peaks. One of the most important observations was the lack of TATA box in the majority of gene promoters, the element that was initially considered as the core, seeding element of transcription initiation. Sharp promoters are predominantly associated with TATA promoters and found mostly on developmental regulator genes. On the other hand, broad promoters are mostly associated with CpG islands and housekeeping genes (Akalin et al. 2009).

Earlier studies from the FANTOM consortium had mostly focused on promoter usage and transcriptional dynamics of core promoters in various tissues from human and mouse. In other words, they were mostly focused on transcriptional initiation, even though even the first sets of CAGE data had already contained evidence for the (then unexplained) post-transcriptional processing and the associated RNA products, which is manifested in the form of CAGE tags being produced off internal coding exons or 3' UTRs (termed "exon painting"). However, the ENCODE pilot project took these observation to the next level of understanding, revealing that the exonic CAGE tags are generated by post-transcriptional recapping events, which was also evident in exon junctions (Fejes-Toth et al. 2009). In addition, they showed that the 5' end of CAGE tags coincide with the 5' end of small RNAs (obtained from small RNA sequencing), confirming that post-transcriptional recapping events lead to the recapping of small RNA fragments. Both CAGE tags and small RNAs were enriched using the CAGE procedure and RNA purification, which favors the capped RNA. Moreover, these exonic related tags were enriched only in certain set of genes and underrepresented in intron and intergenic regions. A subsequent study showed that these post-transcriptional events were conserved between human and mouse (Mercer et al. 2010).

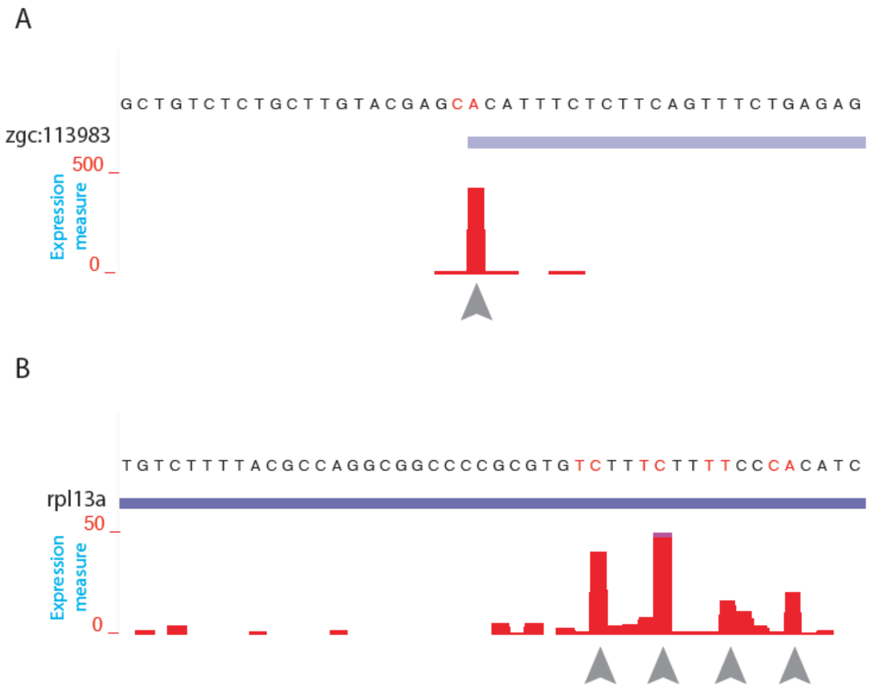

Figure 4. Schematic representation of promoter types. Vertical red bars represent 5' nucleotide of CAGE tags representing transcription start sites. The 5' nucleotide of CAGE tags and the immediate upstream nucleotide define the initiation sequence, which are colored in red. Height of the vertical bars indicates the expression level. Promoters are classified into sharp and broad on the width of the transcript clusters. (A) Single (or sharp) peak have TSSs dedicated from few bases. Arrowhead refers to the CAGE tag that is used dominantly. (B) Broad promoters have transcription starting from wider range. Arrowheads refer to the CAGE tags that are preferentially used during transcription.

### 1.4.2 RNA Sequencing with high-throughput technologies (RNA-seq)

RNA-seq aims to give a quantitative measure of the transcribed regions, where complementary DNA fragments are sequenced. The high-throughput sequencing technologies generate millions of sequenced reads (fragmented sequence), which are mapped to the genome, to build and quantify the transcriptome landscape. The ability of high-throughput sequencing technologies to generate expressed sequence at a very high coverage (which is subjected to the user's choice) at base-level resolution has made RNA-seq technology a popular and widely used tool for transcriptional profiling (reviewed in (Wang et al. 2009)). In addition to the higher coverage at the base-level resolution, the use of RNA-seq technology ensures the sequencing of the whole transcriptome, directly from the cDNA, which has been instrumental in finding novel transcribed regions. The approach of sequencing the whole transcriptome has been revolutionary, as most of the previous method on custom made arrays was limited to the predefined segments of the genome. The other advantage of RNA-seq is its low noise signal, which is helpful in detecting lowly expressed genes at higher confidence, than would have been possible with previous microarray technologies.

RNA-seq was first used in the yeast genome and showed that 74% of the non-repetitive genome was transcribed (Nagalakshmi et al. 2008). Using RNA-seq, Nagalakshmi et al were able to validate most of the previously annotated gene models, and to identify many more novel transcripts from regions previously thought to be inactive. Soon after, RNASeq technology was adopted by the whole community and has been instrumental in finding non-coding RNAs (Khalil et al. 2009; Cabili et al. 2011; Pauli et al. 2012), alternative splicing variants (Sultan et al. 2008) and gene alleles (Skelly et al. 2011).

The high volume of sequencing reads creates bioinformatics challenges. The ability of sequencing machines to produce large amounts of data, with increasing depth due to decreasing sequencing cost, needs to develop better tools for an efficient way of archiving, storing and retrieval and mapping. Mapping million of reads at high accuracy rate and at a relatively short time is indeed challenging, for which several short reads mapping tools such as SOAP (Li et al. 2009), MAQ (Li et al. 2008) and Bowtie (Langmead et al. 2009) have been developed. Initial versions of all these tools suffered from two main problems - dealing with multimapping reads (small reads are bound to map multiple times in the genome) and inability to efficiently map the reads spanning across the exon-exon junctions. The problem of multi mapping reads has since been resolved to some extent with better mapping algorithms and longer reads. The problems with mapping across exon-exon junctions have been addressed with new tools such as Tophat (Trapnell et al. 2009). On top of that, the advances in high-throughput sequencing platforms lead to the generation of longer reads, sequencing of pair-end reads instead of single end, and the strand-specific sequencing protocols, which all make it easier for the mapping tools to map the reads to the genome.

### 1.4.3 Histone modification and ChIP sequencing

Eukaryotic DNA is wrapped around histone proteins, forming a higher order chromatin structure consisting of repeating nucleosomes. Histones are alkaline proteins found in the cell nuclei, where histones H2A, H2B, H3 and H4 function as core histones, and H1 and H5 are linker histones. A nucleosome is a stretch of DNA (~147 bp) wrapped around two of each of core histones to form a histone octamer. Adjacent nucleosomes are connected by internucleosomal stretches of DNA known as linker DNA. The core histones and linker histones are subjected to a large number of post-translational modifications such as methylation, acetylation, phosphorylation and ubiquitination. The modification itself always appears in the naming convention, i.e. H3K4me4 indicates the tri-methylation of histone 3 at lysine 4.

Histone modification signals can be captured by chromatin immunoprecipitation (ChIP), in which a specific antibody is used to enrich DNA fragments from modified sites. ChiP is a method used to determine the genomic location for DNA-binding proteins (histones or transcription factors). Several ChIP-based techniques, including ChIP-chip, ChIP-PET and ChIP-SAGE, have been developed for the study of histone modification or transcription factors binding in large genomic regions (Impey et al. 2004; Wei et al. 2006). With recent advances in sequencing technologies, ChIP-Seq has become the preferred method to identify the genome-wide binding of TFs or modified histones (Barski et al. 2007; Johnson et al. 2007; Mikkelsen et al. 2007). The basic idea of the standard (single-read) ChIP-Seq is to read the sequence of one end of a ChIP-enriched DNA fragment followed by mapping the resulting short sequencing reads to the genome assembly. Millions of tags sequenced from a ChIP library are mapped and form a genome-wide profile in which ChIP fragment counts are overrepresented at sites where a particular histone modification is present or a transcription factor binds.

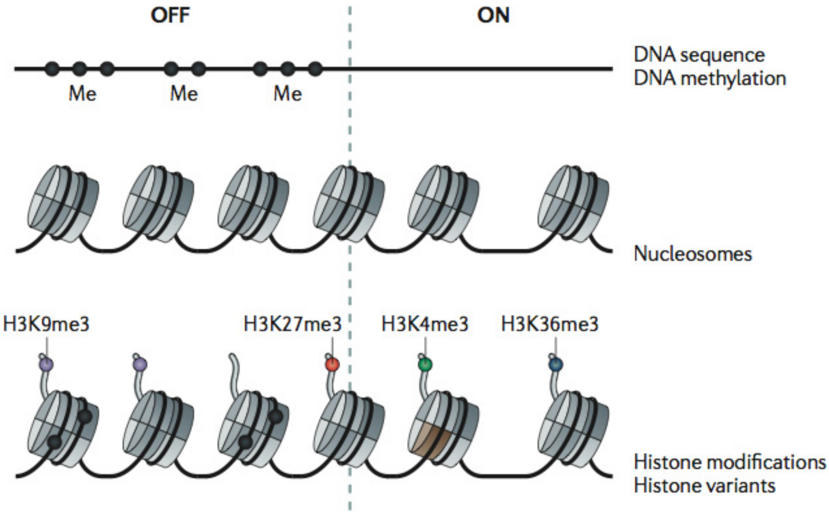

Figure 5. Layers of chromatin organization. DNA is methylated (Me) on cytosine bases in specific contexts and is packaged into nucleosomes, which vary in histone composition and histone modifications. The figure is taken from (Zhou et al.).

Histone proteins are subjected to different forms of post-transcriptional modification, and are associated with distinct cis-regulatory elements.  For example, trimethylation of histone 3 at lysine 4 (H3K4me3) is preferentially associated with promoters of active genes (Bernstein et al. 2002; Santos-Rosa et al. 2002; Bernstein et al. 2005; Barski et al. 2007). Monomethylation of histone 3 at lysine 4 is a modification that is mostly associated with active and poised enhancers and elements found at insulator elements concordant with binding by CTCF (Barski et al. 2007; Heintzman et al. 2007; Akkers et al. 2009), but only to a lesser extent with the promoter region (Barski et al. 2007; Heintzman et al. 2007). H3K27ac is mostly associated with active regulatory elements. Different forms of epigenetic modifications are critical factors influencing gene expression and genome function, and one of the emerging theme is that epigenetic mechanisms of gene expression are controlled by non-coding RNAs (Saxena and Carninci 2011), with many other unknown mechanism.

In addition to the presence of any individual histone mark linked to the activity of various functional elements, the combinatorial patterns of different histones mark can be used for functional annotations indicating distinct biological roles. Large clusters of HCNEs encode developmentally important transcription factors (TFs) genes, consisting of large regions H3K4me27 harboring smaller regions of H3K4me3, often found at silent genes that are poised for activation (Bernstein et al. 2006). The genomic regions marked by H3K4me3 at their promoter and trimethylation of lysine 36 of histone H3 (H3K36me3) along the transcribed region has been used to identify lncRNAs (Guttman et al. 2009). The combinatorial pattern of 51 distinct chromatin states revealed a diverse class of epigenetic functions at different genomic loci, such as promoter regions, intergenic regions and repeat associated regions (Ernst and Kellis 2010). Furthermore, nine chromatin marks from nine different cell types were systematically mapped to characterize cell-type specific regulatory elements, activators, repressors and their functional interaction (Ernst et al. 2011). In addition, it was shown that bivalent and monovalent domains might poise embryonic genes for activation and that the chromatin profile associated with pluripotency is established during the maternal–zygotic transition (Vastenhouw et al. 2010).

## 1.5 Regulatory non-coding RNAs

From being just an intermediate product, RNA has become a central player in gene regulation. High-throughput sequencing technologies have facilitated identification of tens of thousands of non-coding transcripts, many of which are categorized into distinct RNA classes. However, one of the biggest challenges lies in differentiating functional RNAs from the pool of thousands of transcribed RNAs that might be products of RNA degradation. What differentiates regulatory RNAs from the pool of pervasively transcribed RNAs is an open question, which I have tried to address to some extent in papers I, II and III. Non-coding RNAs, with or without functional annotations, are broadly classified into two classes, namely small non-coding RNAs and long non-coding RNAs. No exact fundamental differences in terms of their biological relevance are known to differentiate small non-coding RNAs from lncRNAs. One general approach adopted by the scientific community is the arbitrary length cutoff of 200 nt for separating them (Mercer et al. 2009). Within the two broad classes, small RNAs are further characterized depending upon the sequence, structure and functional similarity. Although no specific sub-classes of lncRNAs have yet been characterized, they are generally described as single/multi exonic, or intergenic/intronic. Despite the arbitrary length cutoff of 200 nt, most studies focused on small RNAs that are in the range of 18-30 nt, which includes miRNAs (Bartel 2004), piwi-RNAs (Houwing et al. 2007), splice-site associated RNAs (Taft et al. 2010; Valen et al. 2011), transcription initiation RNAs (Taft et al. 2009a). On the other hand, functionally annotated long non-coding RNAs are generally few kilobases long. This two end of the spectrum leaves a void for intermediate-sized non-coding RNAs, that generally falls between 50-500 nt. Many of the previously known house keeping RNAs, such as tRNAs, rRNAs, snRNAs, snoRNAs and many other with unknown classes fall under these category. Recent studies have identified many such intermediated-sized non-coding RNAs in *C.elegans* (Deng et al. 2006), human (Yan et al. 2011), chicken (Zhang et al. 2009), *Oryza sativa* (Liu et al. 2012) and we have extended it in zebrafish (Paper-III). The various classes of ncRNAs are described in the following section.

## 1.5.1 Small non-coding RNAs

As the initial large-scale studies in the past decade were primarily focused on small RNAs, partly inspired by miRNA, many different types of small RNAs were discovered. Two novel classes of small RNAs that were studied extensively were siRNA, piwiRNAs, while a lot of expressed RNAs were left without any classification. Many regulatory RNAs are trans-acting elements encoded at different genomic loci than their target mRNAs and function through imperfect base pairing to their targets.

*MicroRNAs* are 22 nucleotide long RNA molecules, found in both plants and animals. They regulate their target genes by base pairing RNAs, which mainly act to downregulate gene expression (Bartel 2004) The primary transcript of a miRNA (pri-miRNA) is transcribed and processed into a short stem-loop structure called a pre-miRNA and finally into a functional miRNA by the RNA-induced silencing complex (RISC) (Bartel 2004; Denli et al. 2004; Lee et al. 2004). A miRNA is integrated into the RISC complex and controls the expression of target mRNAs by base pairing. The exact mechanism (either by inhibiting the translation or degrading the target mRNAs) through which miRNAs regulate the target genes has always been debated (Fabian et al. 2010; Djuranovic et al. 2011). However, two recent studies showed that miRNA regulate target genes first through inhibition of translation followed by mRNA degradation, in zebrafish (Bazzini et al. 2012) and *Drosophila* (Djuranovic et al. 2012). MicroRNAs have been found to regulate genes involved in diverse biological process and implicated in many human diseases (reviewed in (Zhong et al. 2012) (Calin and Croce 2006)

Small-interfering RNAs (siRNAs) are 21-25 nt RNAs usually derived from exogenous RNAs and are believed to be part of a defense system against foreign RNA (reviewed in (Meister and Tuschl 2004)). When foreign RNA enters the cell it is randomly cleaved into double stranded fragments by the RNA endonuclease Dicer. These fragments are recognized by the protein complex RISC (RNA-induced-silencing-complex) which separates the two strands and enables base-pairing of one strand to target RNA (other copies of the same original foreign RNA in the cell), which is subsequently cleaved and degraded.

*Piwi-interacting RNAs* are another small class of RNAs, which are different from miRNAs in terms of size (26-31 nt), general lack of sequence conservation or precise secondary structure and increased complexity. Despite the difference in the sequence, structure and biogenesis, they were likewise found to be extensively involved in the regulation of gene expression, by modulating either mRNA transcription, stability or translation (reviewed in (Erdmann et al. 2001; Storz et al. 2005)). They were first identified in germlines and are associated with Piwi-subclade member of the Argonaute protein family. (Girard et al. 2006; Grivna et al. 2006).

In addition, various new classes of small RNAs have been identified and annotated through high-throughput sequencing. One of the most prominent class of small RNAs are around the TSSs of coding genes, often called as promoter associated small RNAs (Kapranov et al. 2007; Seila et al. 2008) or transcription initiation RNAs (Taft et al. 2009a). They are generally small in size, appear mostly downstream of TSS and transcribed only from a subset of genes. PROMoter uPstream Transcripts (PROMPTs) were identified as a new class of human RNAs, which have varying length and produced only upstream of promoter of active coding genes (Preker et al. 2011). On analyzing the nuclear and cellular component of the cell, a new class of small RNAs transcribed from intron-exon junctions were identified in subset of genes, and termed as splice-spite associated RNAs (Taft et al. 2010; Valen et al. 2011). Evidence of independent 3'UTR transcripts giving rise to various RNAs has already been documented (Mercer et al. 2011).

### 1.5.2 Intermediate-sized non-coding RNAs

Small nuclear RNA (snRNA) is a class of small RNA molecules that are mostly found within the nucleus of eukaryotic cells. The snRNAs are involved in various functions, e.g. mRNA splicing, regulation of transcription factors and maintenance of telomeres. There are 5 major spliceosomal snRNAs (U1, U2, U4, U5 and U6), which are responsible for most of the mRNA splicing at canonical sites. The minor spliceosomal RNAs (U11, U12, U4atac and U6atac) are mostly responsible for splicing of U-12 introns by noncanonical splicing. These minor spliceosomal RNAs constitute less than 2% of the splicing of mRNA, as U12 introns constitute less than 2% of introns. Spliceosomal snRNAs typically occur in multiple copies in the genome of all higher eukaryotes. The spliceosomal snRNA of the same type are mostly found in a tight cluster.

Two distinct classes of small nucleolar RNAs (snoRNAs), box H/ACA snoRNA and box C/D snoRNA, are completely different in their function and structure. The box C/D snoRNAs and box H/ACA snoRNAs guide 2'-O-ribose methylation and pseudouridylation modifications of target RNAs respectively (Bachellerie et al. 2002) . The biological role of snoRNAs is not limited to rRNA modifications, as they have complementary sites in other RNAs, including snRNA and mRNA (Bachellerie et al. 2002; Henras et al. 2004; Kiss et al. 2004; Meier 2005). snoRNAs with no identified complementary sites (yet unidentified targets) are termed orphan-snoRNAs (Huttenhofer et al. 2002) . snoRNAs are generally transcribed from introns of protein-coding genes or non-coding genes (Kiss 2002), with an increasing number of snoRNAs found in intergenic regions independently transcribed by RNA polymerase II (Tycowski et al. 2004). Some C/D snoRNAs were shown to be independently transcribed by RNA polymerase II in human (Tycowski et al. 2004) and in *Caenorhabditis elegans* (Deng et al. 2006).

In addition to well-annotated intermediate-sized non-coding RNAs, recent studies dedicated to intermediate-sized non-coding RNAs have identified from hundreds to a thousands of such novel non-coding RNAs in multiple species across metazoans (Deng et al. 2006; Zhang et al. 2009; Wang et al. 2011; Yan et al. 2011) and plants (Liu et al. 2012). Irrespective of the species studied, three features were associated with many novel intermediate-sized non-coding RNAs. Majority of the intermediate-sized non-coding RNAs could neither be annotated into known RNA classes nor be categorized into novel RNA classes, suggesting large number of such intermediate-sized non-coding RNAs with diverse secondary structures. Secondly, many intermediate-sized non-coding RNAs are less conserved, mostly limited to other species within the clade, giving rise to clade specific non-coding RNAs, while others lack total conservation, giving rise to species specific non-coding RNAs. Thirdly, many intermediate-sized non-coding RNAs exhibit transient or tissue specific expressions.

### 1.5.3 Long non-coding RNAs

The word “long” describes the nature of RNA transcripts that are generally longer than 200 nt (an arbitrary threshold). Long non-coding RNAs are often abbreviated as lncRNAs (long non-coding RNA) or lincRNAs (long interspersed non-coding RNA); from here on I will refer to them as lncRNA. FANTOM3 analysis of full-length transcripts identified around 35,000 transcripts that were 5’ capped, poly-adenylated, spliced, and had short or no open reading frame (ORF) (Carninci et al. 2005), showing first genome-wide evidence of lncRNAs. Later on various technologies such as histone profiling (Guttman et al. 2009), RNA-seq (Khalil et al. 2009; Orom et al. 2010) and manual curation from available EST were successfully used to predict lncRNA. Later on similar approaches were used to annotate lncRNA in various species including mammals (Guttman et al. 2010), vertebrates (Pauli et al. 2012), *Drosophila* (Young et al. 2012) , *C.elegans* (Nam and Bartel 2012), and plants (Ding et al. 2012). lncRNAs can overlap coding exons (either in sense or antisense orientation), lie proximal to promoter regions or in intergenic regions forming a complex interlaced architecture with coding and non-coding genes (Engstrom et al. 2006). lncRNAs typically have relatively low conservation, with occasional short stretches of highly conserved sequence (Pang et al. 2006). As a testimony to above statement, a recent study has shown such short stretches of conserved sequences between orthologous human and zebrafish lncRNAs had conserved functionality (Ulitsky et al. 2011). Despite the accelerated evolution at the sequence level, lncRNAs and proximal/overlapping coding genes retain their synteny across various species (Ponjavic et al. 2009; Cabili et al. 2011; Ulitsky et al. 2011), which can be argued for their functionality.

Unlike small RNAs, where all members of one class of RNA have similar function, lncRNAs have so far defied similar classification, exhibiting diverse functional roles ranging from imprinting (Braidotti et al. 2004), epigenetic regulation (Rinn et al. 2007), splicing (Beltran et al. 2008), enhancer (Orom et al. 2010), repressor (Yochum et al. 2007) among others. At the time of writing (late 2012), the number of functionally characterized lncRNA stands close to 200 (Amaral et al. 2011), which is likely to represent a just a tip of the iceberg. In addition to their regulatory roles, recent studies on dysregulation and mutagenesis of lncRNAs have been linked to various human diseases  (as reviewed in (Wapinski and Chang 2011) ).

## 1.5.4 Derived small RNAs

In search for novel classes of small regulatory RNAs, many researchers have identified the traces of post-transcriptional event and other regulatory element encoded within the well-annotated small or intermediate-sized non-coding RNAs. One of the first studies to identify small RNA encoded within a snoRNA was first reported by Meister and colleague (Ender et al. 2008), where small RNAs are produced by miRNA like processing. Re-analyses of previously published small RNA datasets, in both plants and animals, revealed that many snoRNAs have small RNAs produced from their ends, often termed as small-RNA derived snoRNAs (sd-snoRNAs) (Taft et al. 2009b). However, C/D box snoRNAs and H/ACA box have preferential position for sd-snoRNAs production, where C/D snoRNAs have more sd-snoRNAs at their 5'-end, while H/ACA have more sd-snoRNAs at their 3'-end. This phenomenon of production of small RNAs from snoRNAs ends is evolutionarily conserved indicating there might be an interplay between RNA silencing and snoRNA-mediated RNA processing (Taft et al. 2009b).

Similar mechanism of generation of small RNAs within pre-miRNA was reported on *Ciona intestinalis* (Shi et al. 2009). Small RNAs generated were predominantly ~20 nt long and found in both 5' and 3' arm of pre-miRNAs and are called as miRNA-offset RNAs (moRNAs) (Figure 6). Shi et.al had reported that moRNAs were expressed during early embryogenesis of *Ciona intestinalis,* generally expressed at low level though some of them exceeded the expression of mature miRNAs. Exact biogenesis of moRNAs production is unknown, however it has been speculated to be associated with Drosha processing (Shi et al. 2009). Detailed re-analyses of various small RNA datasets from different human tissues and cell lines revealed an evidence of moRNAs in human (Langenberger et al. 2009). The authors were able to show that many moRNAs were preferentially produced in same miRNAs between human and *Ciona intestinalis,* and significantly overrepresented in oldest animal miRNAs where half of them originated already in ancestral bilaterian. Analysis of nuclear and cytoplasmic sub-cellular localization of RNAs revealed moRNAs were enriched in nucleus, while miRNAs are enriched in cytoplasm (Taft et al. 2010).

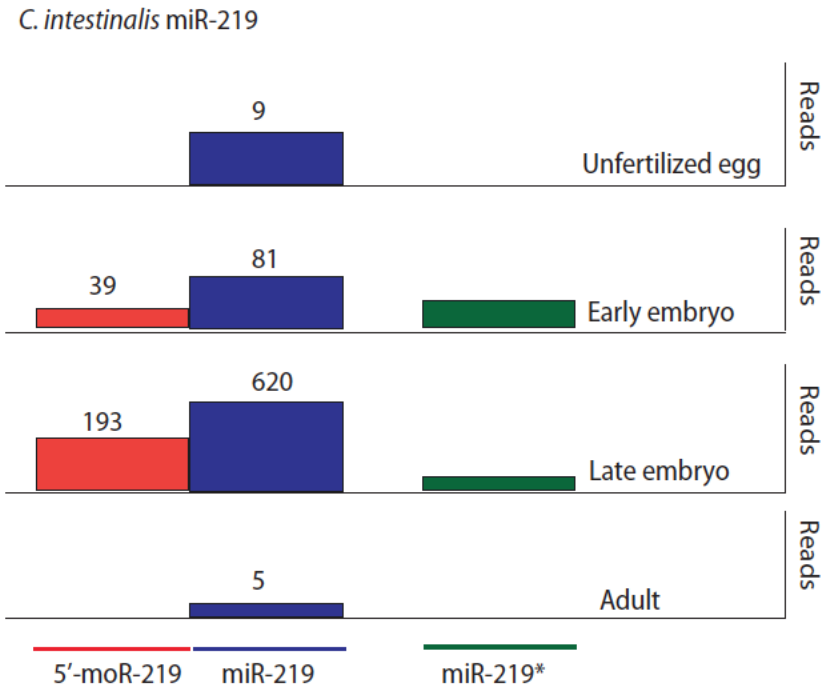

Figure 6. Schematic representation of small RNA reads mapped to pre-miRNA (miR-219) locus in *C.intestinalis*, adapted from (Shi et al. 2009). (A) Reads mapping to miR-219 locus at four developmental time points. Number on top of histograms represents the number of mapped reads (number * $10^3$) at each stage, and color coded histogram represents (miRNA:blue, miRNA*:burgundy, 5'-moRNAs:yellow).

After the identification of small RNAs produced from snoRNAs and miRNAs, evidence of such RNAs were examined in other non-coding RNAs. Detailed re-examination of high-throughput data reveal high abundance of small RNA fragments, derived from mature tRNAs (Cole et al. 2009). They were previously discarded as degraded product, but the authors were able to show the processing of small RNAs derived from tRNA was dependent on Dicer (Cole et al. 2009). Recently, it has been shown that many housekeeping RNAs (tRNAs, rRNAs, snRNAs) undergo asymmetric terminal processing, preferably at 5', producing small RNAs of mostly 18 -22 nt (Li et al. 2012). It is currently unknown, why all these non-coding RNAs produce such small RNAs from their 5'-end, 3'-end or both ends, while it has been speculated that these small RNAs might distinguish these constitutively expressed

RNAs from the pool of other random degrading RNAs (Li et al. 2012). The production of derived small RNAs is not limited to house keeping RNAs or intermediate-sized RNAs, it is even observed in lncRNAs. MALAT1, a well-studied lncRNA, have clusters of small RNAs, not limited to its 5'-end or 3'-end, but spans across the whole transcript (Guffanti et al. 2009). The phenomenon of production of small RNAs across the whole transcript is evolutionary conserved in human, mouse (Mercer et al. 2010) and zebrafish (Pauli et al. 2012). Genome-wide analysis has shown that a subset of lncRNAs are processed to produce small RNAs (Jalali et al. 2012).

## 1.6 Promoters of non-coding RNA genes

RNA polymerase and various transcription factors (TFs) typically bind to the region upstream of gene 5'-end, known as promoter region. Promoter regions contain *cis*-regulatory elements, such as initiator element (Inr) at the TSS, TATA box at 28-34 bp upstream of the TSS, a downstream promoter element (DPE) about 30 bp downstream of TSS (Ohler et al. 2002; Carninci et al. 2006; Ponjavic et al. 2006; Sandelin et al. 2007; Xi et al. 2007; Lenhard et al. 2012). In eukaryotes, three different kinds of polymerase are responsible for transcription initiation. RNA polymerase I transcribes genes encoding all kinds of ribosomal RNA (excluding 5S rRNA) (Russell and Zomerdijk 2006). RNA polymerase II is responsible for synthesis of mRNA, and most small nuclear RNA and miRNAs (Kornberg 1999; Sims et al. 2004). Transcription of 5S rRNA, tRNA and U6 snRNA are initiated by RNA polymerase III, though a few exceptional miRNAs (Borchert et al. 2006), snoRNAs (Dieci et al. 2007) and antisense RNAs (Pagano et al. 2007) are transcribed by RNA polymerase III.

As Pol III transcribes most small RNAs, such as tRNA, U6 snRNA, RNA Pol III was originally assumed to be responsible for miRNA transcription, too. However, in the year 2004, most miRNAs were shown to be transcribed by RNA pol II (Lee et al. 2004). Primary transcripts of miRNAs (pri-miRNAs) contain cap structures as well as poly (A) tails, a hallmark of Pol II transcripts. Prior work from other groups provided additional evidence that miRNA genes are transcribed by Pol II (Lee et al. 2004). Since some of pri-miRNA did not contain 5' cap or poly(A) tail, the authors suggested that other RNA polymerase might also be responsible for miRNA gene transcription. Later in 2006, the first evidence for human miRNA genes transcribed by RNA pol III was published (Borchert et al. 2006).

In relation to protein-coding genes, the majority of miRNA genes are located in intergenic regions (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001; Mourelatos et al. 2002) as independent transcription units (Lee et al. 2002). Most of the other miRNA genes are found in annotated intronic regions of protein-coding genes, some of which may be transcribed as part of the annotated genes (Rodriguez et al. 2004). In animals, miRNAs are transcribed as long primary transcripts (pri-miRNAs), which are cropped into the hairpin-shaped pre-miRNAs by nuclear RNase III Drosha (Lee et al, 2003; Kim, 2004). This cleavage event predetermines mature miRNA sequence and generates optimal substrate for the subsequent events (Lee et al, 2003; Lund et al, 2004).

Our knowledge on miRNA biogenesis has been significantly advanced in recent years. However, little is still known about transcription of miRNA genes although it is likely to be the key regulatory step in miRNA biogenesis. To understand the mechanism of miRNA gene regulation, the basic machinery for miRNA transcription needs first to be identified.

## *1.7 Functional non-coding RNAs*

The rationale behind the functionality of non-coding RNA genes is the ability to produce a RNA product that has an effector function rather than one that serves as an information intermediary in protein synthesis.

### 1.7.1. non-coding RNAs transcribed from enhancers

*Cis*-regulatory elements located away from proximal promoter region that can enhance the transcriptional level of gene are termed enhancers, and play just as important function in gene regulatory network as proximal promoters. Previous studies have shown that enhancers can be located at very long distances from proximal promoter of their target gene, or in introns, overlapping the exons, or beyond neighboring genes (Lettice et al. 2003; Visel et al. 2009a; Dong et al. 2010). Many regulatory regions that function as enhancers are HCNEs, which regulate the expression of surrounding genes (including both target and bystander genes) and maintain the synteny across large evolutionary distances (Engstrom et al. 2007; Kikuta et al. 2007; Maeso et al. 2012). While conserved non-coding elements are able to function as enhancers, non-coding RNAs have generally been perceived as negative regulators, probably due to historical viewpoint. The first two identified lncRNAs, Xist and Air, were involved in genomic imprinting, while miRNAs regulated their target genes either through inhibition of translation or mRNA degradation. Some studies, however, had shown many highly expressed genes were enriched with small RNAs around promoter regions (Taft et al. 2009a; Preker et al. 2011). The hypothesis about exact mechanisms are still unknown and debated if these small RNAs are produced due to polymerase stalling or if they are independently transcribed units that can enhance the expression of nearby gene. However, in the year 2010, two independent studies showed that both small RNAs (Kim et al. 2010) and lncRNAs (Orom et al. 2010) had enhancer-like function. Depletion of a number of lncRNAs led to decreased expression of the neighboring coding genes (Orom et al. 2010). The functional aspect of these non-coding RNAs acting as enhancers, was surprising, as previously known non-coding RNAs had repressing roles. Soon after, another study identified chromatin associated RNAs were able to fine tune the gene expression of neighboring gene by modulating the chromatin structure in cis (Mondal et al. 2010).

### 1.7.2. non-coding RNAs as regulators of embryonic development

MicroRNAs have turned out to be master regulators involved in various biological processes, as mentioned in section **1.4.1**, including embryogenesis. During the early stages of fertilization and cell division, RNA deposited by the mother drives transcription. As maternally inherited RNAs are present to begin with, they can be regulated only at post-transcriptional level. It was first shown from Schier's lab, where mir-430 family is expressed before the onset of the zygotic genome, and accelerates the degradation of maternally inherited mRNAs (Giraldez et al. 2006). Similar mechanisms of miRNA-mediated regulation during embryogenesis were also observed in the frog *Xenopus laevis* (Lund et al. 2009) and *Drosophila* (Bushati et al. 2008). Lund et al. showed that miR-427 mediates the rapid deadenylation of maternally inherited mRNA that follows right after MBT of embryogenesis. However, miRNA that performs this function is different in zebrafish (miR-430) compared to *Xenopus* (mir-427), but they share similar seed site (reviewed in (Svoboda and Flemr 2010)). Reprogramming of an oocyte into pluripotent blastomeres, often referred as Mid blastula transition (MBT) can be considered as analogous to differentiation and pluripotency of embryonic stem cells (ESC) into differentiated cells. The transcription factors SOX2, NANOG and Oct2 form the core component for the transcriptional control of ESC renewal and pluripotency (Boyer et al. 2005). A recent study has shown that miRNA (miR-302/367) -mediated reprogramming of human and mouse somatic cells to pluripotency was two order of magnitude more efficient than standard Oct4/Sox2/Myc-mediated methods (Anokye-Danso et al. 2011), and miR-302/367 also have similar seed sites with miR-430.

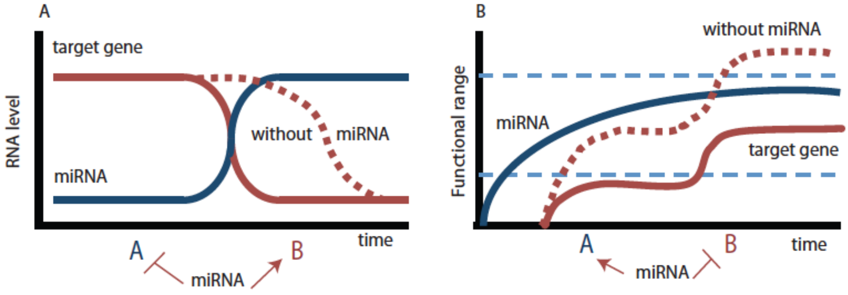

Figure 7. Reciprocal roles of miRNAs during early development in plants vs. mammals. The figure has been adapted from the review paper (Bazzini and Giraldez 2011). (A) Schematic representation of miRNA during early embryogenesis in animals, which helps in clearance of maternal mRNAs. (B) miRNAs prevent premature expression of target mRNA. High levels of mRNAs are reached rapidly in absence of miRNAs.

Ever since its discovery, miRNA has always surprised us with new functional roles. Contrary to our previous understanding, a recent paper from Bartel's lab showed the exact opposite function of miRNA during plant embryogenesis (Nodine and Bartel 2010). Nodine et.al showed that mir-156 preemptively represses the genes that function later in the development to prevent developmental transition (Figure 7). Though this function has been observed only in plants, it remains to be seen, if the functionality is conserved in the animal kingdom. Considering the similarities in biogenesis and targeting of miRNAs between plants and animals, it might be just a matter of time to uncover orthologous functionality in animals.

Unlike miRNA, the possible roles of lncRNAs as a regulator in the context of embryonic differentiation and pluripotency have just started to be uncovered. Sheik Mohamed et al showed the implications of lncRNAs in the modulation of mouse ESc pluripotency by genome-wide transcriptional mapping of key Esc transcription factors Oct4 and Nanog, and indicated the possible role of lncRNA in controlling the pluripotent state (Sheik Mohamed et al. 2010). Later, many lncRNAs were shown to act as regulators of differentiation and pluripotency state in mouse ESC differentiation (Guttman et al. 2011). Guttman et.al showed that knocking down these lncRNAs affected gene expression, mostly in trans, or cause it to exit from the pluripotent state, similar to the effect of knocking down major transcription factor regulators. Since then, many papers have highlighted the roles of lncRNAs as regulators in various processes (Hu et al. 2011; Kretz et al. 2012). Two recent papers annotating zebrafish lncRNAs, across various stages during embryogenesis (from oocytes to an adult), identified around 550 (Ulitsky et al. 2011) and 1100 (Pauli et al. 2012) lncRNAs. Most of these lncRNA exhibit the temporal and spatial expression pattern, indicating their possible functional roles. Ulitsky et.al showed the evidence of two lncRNAs regulating the brain morphogenesis on zebrafish, the functionality of which is conserved to human.

## 1.8 Annotation of non-coding RNAs

Identifying the transcripts that are functional and further characterizing into sub classes is still challenging. Ideal(istical)ly, annotation refers to the identification of common properties (high resemblance within the class) in a set of transcripts that distinguishes them from the rest of transcripts. However, so far, only a few classes of non-coding RNAs have been annotated, which indicates difficulties regarding annotation. The complexity of non-coding RNAs annotation has been evident in the ENCODE project, which was unable to annotate the transcripts into various families of non-coding RNAs despite cataloguing the transcriptome across various cell lines and tissues. Far from annotation, even an approximation of non-coding RNAs is lacking in many species. This is in stark contrast to the situation with protein-coding genes in eukaryotic genomes, most of which are annotated and a substantial number have been functionally characterized (excluding their alternative isoforms).

Two broad classes of (short and long) non-coding RNAs have different approaches for annotation. One common step is to filter out spurious transcripts (very low coverage), although low expression level is an intrinsic property of many non-coding RNAs. Despite low expression, if non-coding RNAs exhibit developmentally regulated patterns across multiple tissues and developmental time points, it is likely to be a functional non-coding RNA. One could also exclude RNA fragments overlapping coding genes on sense strand, as this might be a degradation product.

However, many exonic RNA fragments show developmentally regulated patterns independent of coding genes they overlap, indicating that one has to be cautious while dealing with them but not discard them indiscriminately.

Annotations of small RNAs are typically based on the sequence and structure homology they share with the other members within each class, as reviewed in (Griffiths-Jones 2007; Forrest et al. 2009). Most of the known “housekeeping” RNAs, such as tRNAs, rRNAs and snoRNAs, have high sequence similarity within the group, and are evolutionarily conserved which makes detection in other species straightforward. In addition to the high sequence similarity, they maintain conserved secondary structure, so the combination of the two can be used to detect them more reliably and often distinguish active genes from numerous pseudogenes. Initially, the annotation of miRNAs was done based on its characteristics secondary hairpin structure, and evolutionary signatures were further used to filter the false positive predictions. However, it has since been established that miRNAs cover a broad spectrum of levels of evolutionary conservation, from those shared across phyla to those that are specific for recent lineages. High-throughput sequencing data along with bioinformatics approaches are used to annotate miRNAs. Evolutionary conservation is even less useful for the detection of other types of recently discovered small RNAs (e.g. tiRNAs, pasRNAs, splice-site RNAs), most of which do not possess any sort of sequence or structure similarity among them. However, they do possess certain characteristics, such as distance from reference TSSs or splice junction, preference for a particular nucleotide at 5’-end, that can be used to classify one type of RNAs from another. Remaining candidates can be filtered on the basis of coding potential (Kong et al. 2007) or codon substitution frequency (Lin et al. 2011), that can differentiate if a transcribed sequence is a coding or a non-coding transcript.

The rationale behind the annotations of lncRNAs is entirely based on the evidence that a given transcript does not encode for a long ORF, and hence might not encode for protein. In short, to annotate lncRNAs, transcripts that share homology with known protein sequences from the protein database are filtered out by blast (tblastx, blastp). The remaining transcripts are further filtered out based on the length of ORF. General criteria used for ORF length cutoff is 300 nt, which is empirically based on the fact that most annotated coding genes have ORFs longer than 300 nt (Dinger et al. 2008), although some exceptions exist, where functionally annotated non-coding RNAs have ORF longer than 300 nt. Finally, various tools, such as CRITICA (Badger and Olsen 1999), coding potential calculator (Kong et al. 2007), and Portait (Arrial et al. 2009), are used to filter remaining transcripts with evidence of protein-coding, leaving probable lncRNA candidates.

One of the first studies showing the evidence of genome-wide lncRNA sequenced full-length cDNA (Carninci et al. 2005). The approach used to annotate was entirely based on the stringent threshold of 100 aa (amino acids) of ORF. A later study had shown that indeed most coding genes have longer than 100 aa (Dinger et al. 2008). While the subsequent studies used various high-throughput sequencing technologies including chromatin profiling, RNA-Sequencing, and ribosomal profiling (Ingolia et al. 2011) to annotate lncRNAs. The use of chromatin maps, i.e. for H3K4me3 (maps promoter regions) and H3K36me3 (maps the gene body) was very successful in identifying lncRNAs, (Guttman et al. 2009). However, the use of chromatin signature alone was limited, as it was unable to reconstruct the precise gene structure. Subsequently, the combinations of RNA-Seq and chromatin maps have facilitated lncRNAs annotation (Khalil et al. 2009; Guttman et al. 2010). At present, the use of RNA-Seq alone or together with chromatin marks is a standard approach to identify lncRNAs. Various sequencing technologies have been used to identify lncRNAs, however the overlap between lncRNAs reconstructed from different technologies with the manually curated lncRNA (partial lncRNA set) from GENCODE was surprisingly low (Orom et al. 2010). Small overlap among lncRNA transcripts annotated from different sequencing technologies probably reflects the sensitivity of different sequencing technologies in identifying transcribed regions at low level. Hence, using various sequencing technologies in combination and possible manual curation can strengthen lncRNAs annotation.

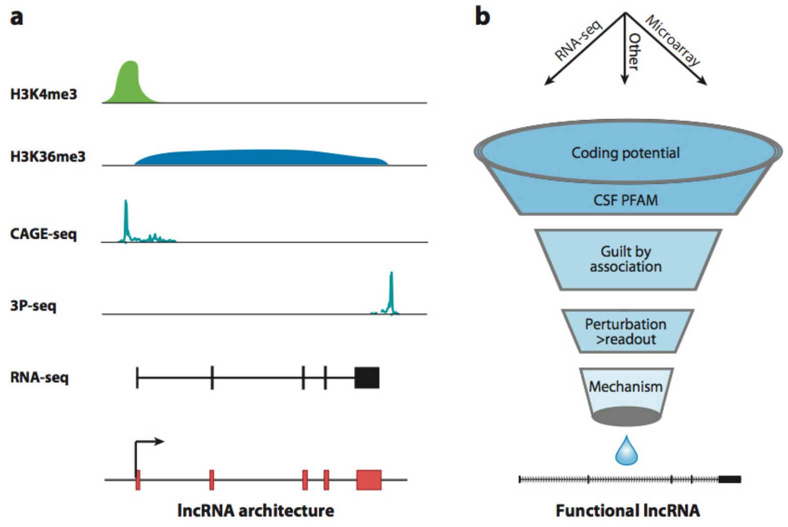

Figure 8: Identification, annotation and functional characterization of lncRNA. The figure is taken from (Rinn and Chang 2012). (a) Different high-throughout sequencing technologies can be used to identify putative lncRNAs. Combining different sequencing technologies can be used to filter unlikely candidates. (b) Various filtering steps used to identify the functional lncRNA.

Two recent studies on zebrafish lncRNA (Ulitsky et al. 2011; Pauli et al. 2012) were performed using similar sequencing technologies (both RNA-Sequencing and histone profiling) including some overlapping developmental stages. On comparing the annotated lncRNA showed an overlap of merely around 5%, which reflects that the methods used to annotate lncRNAs are still not standardized, Hence, there is a need for a more robust method. The problem often lies in the (varying) threshold used to differentiate coding vs. non-coding transcripts; even worse, there is no guarantee that the annotated lncRNA is indeed not translated. The only way to resolve this would be to verify experimentally, if each of these lncRNA can be translated, which can be done by ribosomal profiling and tandem mass spectrometry (Banfai et al. 2012). Ribosomal profiling gave an indication that many lncRNAs were indeed translated into small peptides and might not be a genuine lncRNA (Ingolia et al. 2011). However, the recent ENCODE paper revealed that most of manually annotated GENCODE lncRNAs (Derrien et al. 2012) were indeed true lncRNAs as they were rarely translated (Banfai et al. 2012).

# 2 Present investigation

The main purpose of this thesis was to characterize the transcriptional (both coding and non-coding) landscape of zebrafish during embryogenesis. When the project started, the zebrafish protein-coding gene annotation was incomplete, while non-coding annotation was almost non-existent (limited to a handful of “housekeeping” RNAs and miRNAs). However, during this time, there has been a great improvement in the genome assembly (from Zv6 -> Zv9, thanks to the efforts at Wellcome Trust Sanger Institute) and the current Ensembl gene build has incorporated RNA-Seq transcripts from various developmental stages and tissues to annotate both coding and non-coding transcripts (Collins et al. 2012). In addition, efforts by individual labs have helped in further characterizing the zebrafish embryonic transcriptome (Aanes et al. 2011), lncRNAs (Ulitsky et al. 2011; Pauli et al. 2012) and miRNAs (Wei et al. 2012). Despite this recent work focused on zebrafish embryogenesis, our data provides a unique resource for the core promoter (both coding and non-coding transcripts) repertoire and its dynamic usage during embryogenesis at the single-nucleotide resolution.

High-throughput sequencing technologies allowed the detection of a plethora of non-coding RNAs (Kapranov et al. 2007), and their importance has been well recognized in the scientific community  (Figure 2). However, the existence and biological relevance of these non-coding RNAs detected in large-scale analysis of human tissues have not been comprehensively applied to a vertebrate animal model *in vivo*. We choose zebrafish as model organism, as experiments can be conducted *in vivo*, moreover zebrafish embryonic developmental stages allow researchers to study core promoter usage and their dynamics under changing conditions, such as in a developing vertebrate embryo, which is analogous to the differentiation of pluripotent embryonic stem cells to differentiated cells.  The ontogeny of the zebrafish embryo, like that of other anamniotes, is characterized by a dramatic transitions with global changes in transcriptional activities during the mid-blastula transition (MBT) (Schier 2007). Before the MBT, the pluripotent cell mass evolving from the fertilized egg is characterized by transcriptional inactivity. During MBT, dramatic upregulation of the zygotic genome occurs in parallel with maternal mRNA degradation (Mathavan et al. 2005), providing the necessary transcriptome changes to drive specification and determination of cell fates during specialization and differentiation.

We used CAGE technology as the primary sequencing technology for this study. CAGE has enabled us to perform an improved annotation and description of core promoters on a genomic scale, revealing intricate details about TSS usage and dynamics at single nucleotide resolution. CAGE technology provided the opportunity to classify several non-coding RNAs categories detected in human and other genomes, and suggest involvement of post-transcriptional processing in their generation (Taft et al. 2009a; Mercer et al. 2010). We have coupled CAGE maps to protein-coding and non-coding transcripts by RNA sequencing, providing a quantitative description of TSS usage on a genome scale. Finally, these maps were compared with post-translational histone modifications (H3K4me3 and H2AZ) by ChIP sequencing to correlate the CAGE data with chromatin structure, which would further help in differentiating the embedded post-transcriptional landscape within the transcriptional landscape.

Here, we have set out to generate the first global description of TSS usage during key stages of vertebrate embryonic development at single nucleotide resolution. Twelve developmental time points from zebrafish embryogenesis were selected, representing the critical phases of vertebrate ontogeny: they spanned maternal to zygotic transition at MBT and the subsequent stages of differentiation leading to formation of the body plan and organ systems. CAGE gave rise to improved annotation and description of core promoters on a genomic scale, revealing intricate details about TSS usage and dynamics at single nucleotide resolution (Carninci et al. 2006). Next we focused on identification and characterization of miRNA promoters, which due to various reasons, have not been analyzed in zebrafish before. These reasons stem from the fact that pri-miRNAs are almost never identified as full-length products because they are quickly processed into pre-miRNAs. We combined high-throughput transcriptional profiling (CAGE-Seq) and chromatin profiling (H3K4me3 and H2A.Z) along various developmental time points and characterized transcriptional, post-transcriptional, and chromatin regulation of pri-miRNAs and pre-miRNAs. Next, we generated an intermediate-sized non-coding RNAs, to validate previously annotated house keeping RNAs and identified a thousand of novel non-coding RNAs. Most annotated non-coding RNAs were snoRNAs, and hence we focused on the characterization of the transcriptional initiation mechanisms of snoRNAs host genes. We showed they are co-transcribed with their host genes, which are regulated by noncanonical transcription, specialized to translation machinery.

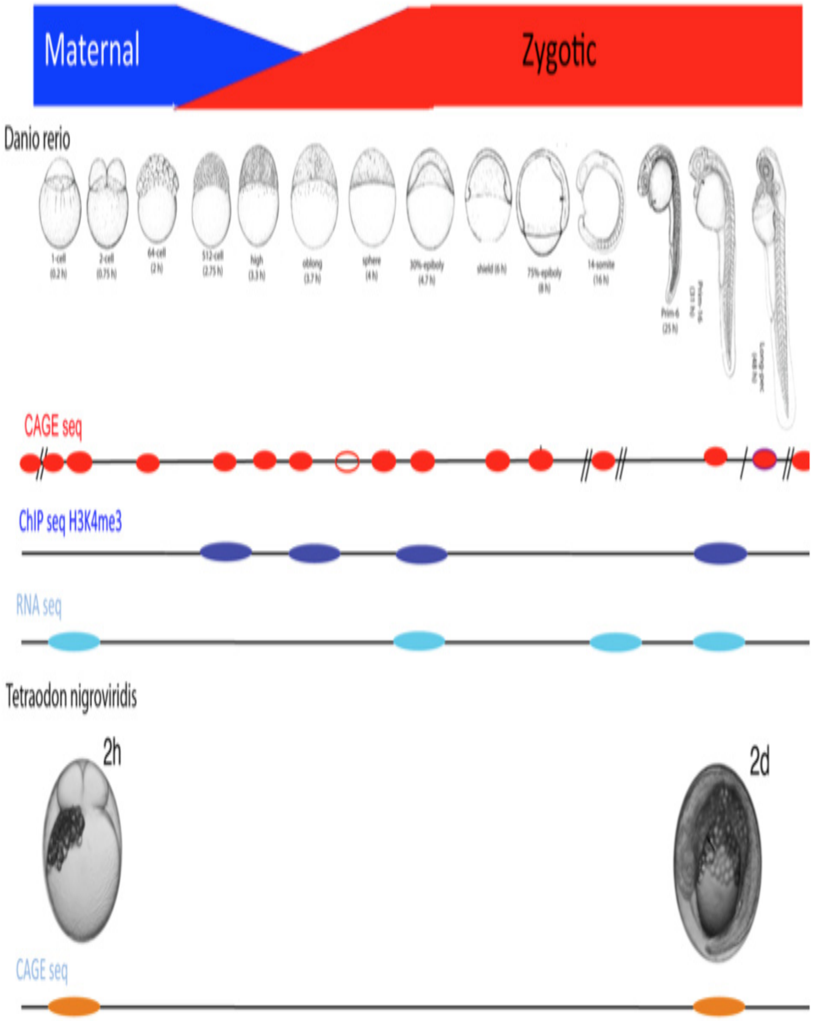

Figure 9. Schematic representation of developmental time points and the corresponding sequencing technologies used in the study. Three different technologies, CAGE-seq, RNA-seq and ChIP-seq, were used to characterize the early embryonic transcriptome of zebrafish. Two stages, corresponding to maternal and zygotic stages, were selected from pufferfish (*Tetraodon nigroviridis*) to analyze their evolutionary conservation of promoter usage.

## 2.1 Paper I

**Dynamic regulation of coding and non-coding transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis**

### 2.1.1 Context

Precise spatial and temporal control of the transcription of protein coding and non-coding genes is a fundamental process underlying development and differentiation of multicellular organisms. The core promoter, which is a relatively short stretch of sequence around the transcription start site, contains regulatory information for the recruitment of transcription initiation factors. Core promoters interact with gene regulatory elements and vary in different physiological and developmental contexts. Accurate promoter predictions based on mapping TSSs during development are needed to decipher the complex interplay between DNA sequence determinants for transcription initiation and epigenetic regulation on core promoters. Despite the existence of a number of alternative core promoter motifs (Juven-Gershon and Kadonaga 2010), a global code for core promoters is still elusive. The lack of TSS data so far has restricted the study of developmental regulatory mechanisms of transcription initiation in vertebrates due to the unreliable TSS position detection based on cDNA/EST and RNA-Seq data, and scarcity of available datasets.

High-throughput sequencing technologies allowed the detection of a plethora of non-coding RNAs (Jacquier 2009; Pauli et al. 2011). CAGE gave rise to improved annotation and description of core promoters on a genomic scale (Carninci et al. 2006; Kodzius et al. 2006) and provided the opportunity to classify several non-coding RNA categories detected in human and other genomes and suggest involvement of posttranscriptional processing in their generation (Hoskins et al. 2011). Despite progress in our understanding of promoters, which is mostly based on cell and tissue culture experiments, we lack a genome-scale data of core promoter usage and the dynamics of it under changing conditions in a developing vertebrate embryo. The early ontogeny of the zebrafish, like other anamniotes, is characterised by a dramatic transitions with global changes in transcriptional activities during the MBT (Kane and Kimmel 1993; Schier 2007).

The existence and biological relevance of these non-coding RNAs detected in large-scale analysis of human tissues have not yet been characterized in a vertebrate animal *in vivo*. To gain insight into the existence and biological relevance of these non-coding RNAs in vertebrate animal *in vivo,* we have set out to generate the first global description of TSS usage during key stages of vertebrate embryonic development at single nucleotide resolution. We have coupled CAGE maps to protein-coding and non-coding transcripts by RNA sequencing, providing a quantitative description of TSS usage on a genome scale. Finally, these maps were anchored to posttranslational histone modifications (H3K4me3) by ChIP sequencing to correlate the CAGE data with chromatin structure across critical phases of vertebrate ontogeny, among them the maternal to zygotic transition at MBT and the subsequent stages of differentiation leading to formation of the body plan and organ systems.

### 2.1.2 Results

We reveal an extraordinary dynamics of promoter usage that takes place during development of the vertebrate embryo. We show that the onset of transcription and subsequent differentiation of the embryo is characterized by the developmentally regulated appearance of 5'-ends of intragenic RNAs on many genes, and of an entire hitherto unknown layer of RNA species overlapping known genes and having specific signatures occurring in exons, introns and 3'-UTRs of developmentally active genes. We showed MZT is manifested by initiation of pervasive transcripts of non-coding intragenic RNAs, and demonstrated existence of intronic and exonic hitherto underappreciated alternative transcripts. We provide several lines of independent evidence that the intragenic CAG tags indicate posttranscriptional products rather than *de novo* transcribed RNAs. We provide insights into evolutionary conserved features of core promoters and describe a novel vertebrate specific initiator sequence shared by a subset of genes also in human, showing that our zebrafish dataset has the potential to reveal promoter features shared by all vertebrates. The key findings of this work are summarized below:

- We describe core promoter dynamics on a genome-wide scale throughout embryo development, demonstrating the developmental diversity of transcription initiation regulation mechanisms and promoter types.
- We characterize the pervasive production of intragenic processed RNAs including exonic and intron-5' end specific RNAs and provide the first indication for

the biological processes in which they may function. Notably, intron 5’ end associated non-coding RNAs are active zygotically and restricted to genes that encode RNA processing and the splicing proteins in both fish and human.

- We demonstrate by four lines of evidence that exonic RNAs are produced by a non-canonical posttranscriptional mechanism independent of the gene 5’ end. We used reporter assays to show that the sequences associated with 5’ end of exonic RNAs cannot function as promoters, and are hence likely to be associated with post-transcriptional processing.
- We show the initiation landscape and developmental dynamics of lincRNAs; we show the evolutionary conserved process of developmentally regulated posttranscriptional processing of lincRNAs into intragenic RNAs, which demonstrate the utility of zebrafish in studying mammalian lincRNA processing.
- We generated the promoterome of a small genome species, pufferfish *Tetraodon nigroviridis*, and exploited the power of comparative genomics to identify novel features of core promoters. Thus we discovered a novel type of transcription initiation signal (AA initiator) conserved with human and used by a subset of vesicle and membrane transport-associated genes. This initiator indicates the existence of a non-canonical initiation mechanism in vertebrates.

### 2.1.3 Ideas for future work

As part of ZEPROME consortium, we have conducted a concerted effort to provide the zebrafish scientific community with high quality data, from three different genome-wide assays, which can be used to study vertebrate embryogenesis. We have reported a novel usage of transcription initiation codes on vertebrate core promoters, on a subset of maternally inherited genes that have different promoter usage during maternal and zygotic initiation, within the same core promoter (Haberle et. al, manuscript under review). In addition, we have various interesting observations that can be examined further to decipher the roles of previously unappreciated intragenic RNAs during embryogenesis. One of the major works is to verify the existence of these intragenic and intergenic RNAs by independent approach and to show their size and dynamics during development. This can be followed by *in situ* hybridization, to show their subcellular localization, which might lead us to more insight into further studies to associate them with functions.

Two distinct features of 5’-intronic CAGE tags (transcribed from the first base of an intron) make them an interesting class of RNAs for further studies. These 5’-end intronic RNAs are encoded on a subset of genes that are enriched for splicing regulators themselves. We also identified 5'-end intronic RNAs on a subset of genes in human, that are enriched for splicing regulators, which suggest the production of 5'-end intronic RNAs associated with evolutionary conserved function. Temporal dynamics of the host genes and 5'-end intronic CAGE tags reveal interesting patterns. The host genes are maternally transcribed, while transcription initiation of 5'-end intronic RNAs coincides with the onset of zygotic genome. This suggest 5'-end intronic RNAs can be specifically associated with regulation of splicing. After initiation of zygotic genome, cells starts dividing, and during that time 5'-intronic RNAs are transcribed, very likely to regulate the splicing of de novo transcribed genes. To study the functional roles of 5-'-end intronic tags, splicing of the genes encoding 5'-intronic tags can be blocked by morpholino, and see its affect during embryonic development. However, one needs to be cautious, as interference with splicing might also lead to phenotypes due to loss of function of full-length RNAs, which are not easy to separate with current existing technologies. Alternatively, a knock-in with GFP tag in intron could be useful to detect them by ISH in the embryo and in subcellular compartment.

We also confirmed the 5' end and promoter regions of several hundreds of lncRNAs that were recently reported by two other groups (Ulitsky et al. 2011; Pauli et al. 2012). Ulitsky et al showed two lncRNAs candidates that have crucial roles required for proper embryonic development, and hence showing as zebrafish as a vertebrate model organism for studies of lncRNAs in vivo. We have identified many non-coding transcripts with transient expression, specific to maternal, MZT or zygotic stages, and such transient non-coding RNAs specific to maternal and MZT specific are the best candidates for functional studies. Large-scale studies, based on knockout of candidate genes will help in elucidating the functional roles of lncRNAs.

## 2.2 Paper II

**Transcriptional, post-transcriptional and chromatin associated regulation of pri-miRNAs, pre-miRNAs and moRNAs in zebrafish development**

### 2.2.1 Context

The main aim of this work was to provide a (currently non-existent) annotation of miRNA promoters and characterize their common characteristics features at transcription, post-transcription and chromatin level. Identifying the precise TSS and its dynamic usage during embryogenesis at single base resolution along with its epigenetic state will aid in better understanding of miRNA biogenesis and transcriptional regulation. So far, large-scale studies have unraveled the functional roles and biogenesis of miRNA (mostly on pre-miRNA) during zebrafish embryogenesis (Chen et al. 2005; Watanabe et al. 2005). However, only limited efforts have been made to try to characterize miRNA promoters (He et al. 2011). Large-scale profiling of small RNAs has revealed developmental dynamics during embryogenesis (Chen et al. 2005; Watanabe et al. 2005). Relatively low expression and rapid processing of pri-miRNAs had hindered the studies the pri-miRNA and hence were limited to a handful of examples (Lee et al. 2004; Liu et al. 2007; Woods et al. 2007). However, with the recent advances in high-throughput sequencing technologies that allow reliable detection of transcription start site regions, detection of promoter regions of miRNA genes has also become possible on a genome-wide scale (Marson et al. 2008).

### 2.2.2 Results

By coupling high-throughput sequencing with on transcriptional (CAGE) and chromatin profiling (H3K4me3 and H2A.Z), we systematically characterized miRNA promoters during zebrafish embryogenesis. To this end, we identified the TSSs for a total of **154** distinct miRNAs representing **87** distinct miRNA families. To the best of our knowledge, this is the first large-scale study to characterize zebrafish miRNA promoters, and reveal temporal dynamics of pri-miRNAs during early embryogenesis. Alignment of H3K4me3 and H2A.Z with respect to pri-miRNAs and pre-miRNAs showed an enriched H3K4me3 and H2A.Z modified nucleosomes downstream of TSSs, providing an additional layer of evidence supporting pri-miRNAs. We identified CAGE evidence for RNA species within pre-miRNAs, mostly towards 3'-end of pre-miRNAs, similar to what has been reported as moRNAs (Shi et al. 2009), and hence report the first evidence of moRNAs in zebrafish. We further showed the production of moRNAs towards 3'-end of pre-miRNAs is conserved in *Tetraodon nigroviridis* (a genomics model vertebrate teleost with a compact genome). The key findings of this work are summarized below:

- We describe the first genome-wide identification of miRNA promoters in zebrafish active during the early embryonic developmental stages. We identified a small number of maternally transcribed miRNAs, one MBT-specific miRNA and the majority that are zygotically transcribed.
- Sequence characteristics of pre-miRNAs reveal high CG content centered on pri-miRNA TSSs, often overlapping with CpG Island and TATA motifs, and are evolutionary conserved.
- We report the first evidence of moRNAs in zebrafish and pufferfish (*Tetraodon nigroviridis*), that were previously reported in human and *Ciona intestinalis* (Shi et al. 2009).
- We show evidence for unexpected enrichment of pre-miRNA sites with promoter-associated histone modification marks (H3K4me3 and H2A.Z) suggesting chromatin regulation and potential involvement of transcription machinery in pre-miRNA processing, suggesting co-transcriptional splicing of pre-miRNAs and pri-miRNA.

## 2.2.3 Ideas for future work

MicroRNAs play important roles during embryogenesis, either in accelerating the clearance of maternal mRNA (Giraldez et al. 2006) or for proper development of zygotic genome (Tang et al. 2007). These miRNAs are transcribed during early onset of the zygotic genome, and no maternally inherited miRNAs were previously known, at least in zebrafish. The fact that we identified nine miRNAs as maternally transcribed, their importance during embryogenesis remains to be seen. Of these nine maternal miRNAs, 5 miRNAs from the miR-17 cluster are the most interesting candidates for further analysis. The miR-17 cluster has seed sites similar to miR-430 or miR-302 in human (embryonic stem cell regulator), as reviewed in (Svoboda and Flemr 2010). Many transcripts upregulated in MZ-dicer mutants but not targeted by miR-430 (Giraldez et al. 2006), are apparently miR-17 target genes (Nepal. et.al, manuscript under preparation). A recently study showed that the miR-17 cluster is already expressed at 2hpf, and has speculated on its importance during embryogenesis (Bazzini et al. 2012). However, a functional study by knocking down the miR-17 cluster would be necessary to characterize its roles and phenotypic effect during embryogenesis. Considering the similarities in seed sites between the miR-430 and miR-17 clusters, along with their expression in early embryonic stages, it remains interesting to see, if they work coordinately or in tandem.

Although first identified in the year 2009 in *Ciona intestinalis* (Shi et al. 2009), moRNAs are evolutionary conserved in human (Langenberger et al. 2009), and now we report moRNAs in zebrafish. We identified moRNAs are produced exactly in same pre-miRNAs between human, *C.intestinalis* and zebrafish, which indicates the functional (yet unknown) significance of moRNAs. However, some common differences have been observed regarding the position of moRNAs, where it was identified on both arm of pre-miRNA on *C. elegans* (Shi et al. 2009), while 5’ end of pre-miRNA is preferred in human (Langenberger et al. 2009; Taft et al. 2010) and 3’-end is preferred in zebrafish, irrespective of location of mature miRNA. The fact that moRNAs are developmentally regulated during embryogenesis suggests potential roles for moRNAs in gene regulation. To establish whether they are themselves functional one needs to perform a site-directed mutagenesis experiment. The CAGE technology provides information about the exact nucleotide that is cleaved. If mutating that particular nucleotide changes the miRNA processing, then it could be concluded that its potential role is associated with miRNA processing machinery. Another approach would be to address the mechanism of pre-miRNA processing, on a genome-wide scale, by carrying out loss of function analysis with miRNA biogenesis enzymes such as dicer for example by high-throughput sequencing of dicer mutant (on human cells and or zebrafish).

One of the most intriguing differences observed between *C. intestinalis* and zebrafish moRNAs was in their temporal expression patterns. In *C. intestinalis*, it was reported that moRNAs were expressed very early during embryogenesis (Shi et al. 2009), while we identified no traces of moRNAs during early embryogenesis in zebrafish. Although we have identified some miRNAs that are maternally inherited, none of these miRNAs have their moRNA expressed until the onset of zygotic genome activation. If these moRNAs are associated with dicer processing, it raises the question if the maternally inherited miRNAs are functional or not until the onset of zygotic genome activation.

## 2.3 Paper III

**Genome-wide characterization of snoRNAs in zebrafish reveals their co-transcription with coding and long non-coding host genes by non-canonical transcription initiation**

## 2.3.1 Context

Recent high-throughput sequencing studies on zebrafish embryonic transcriptome have identified a thousand of lncRNAs (Ulitsky et al. 2011; Pauli et al. 2012), and many novel coding genes (Aanes et al. 2011), miRNAs and other small non-coding RNAs (Wei et al. 2012). As mentioned in Section 1.5.2, the focus of two extremes in size of RNA genes generally makes such intermediate-sized non-coding RNAs (is-ncRNAs) underrepresented in most studies. The range includes well-annotated RNAs such as snRNAs, tRNAs, rRNAs and snoRNAs. Moreover, previous studies have identified hundreds of clade specific intermediate-sized novel is-ncRNAs in various species, (Deng et al. 2006; Zhang et al. 2009; Wang et al. 2011; Yan et al. 2011). Hence, the main aim of this work was to provide a catalogue of intermediate-sized non-coding RNAs in zebrafish, and analyze their characteristics features and modes of transcriptional regulation during early embryogenesis. Since snoRNAs were over-represented in the dataset, we were interested on charactering their transcriptional and chromatin profiling, with respect to their host genes, during early embryogenesis.

### 2.3.2 Results

We have provided a catalogue of intermediate-sized non-coding RNAs in zebrafish, by making RNA library enriched for intermediate-sized (50-500 nt) non-coding RNAs, collected from zebrafish larvae (5-7 days post fertilization (dpf)), and performed 454 pyrosequencing. This allowed us to verify a large portion of previously predicted non-coding RNAs and identify a thousand novel intermediate-sized non-coding transcripts. In particular, we validated most annotated snoRNAs and identified many novel snoRNAs making the most comprehensive annotations of zebrafish snoRNAs. We identified host genes for most snoRNAs and showed no evidence for independent transcription of snoRNAs, suggesting they are co-transcribed by host genes. Interestingly, host (coding and non-coding) genes require non-canonical transcription initiation machinery, as indicated by TCT initiation signals, that is associated with translation machinery. Comparative genomics suggest orthologous human snoRNA host genes also have conserved transcription initiation machinery, suggesting an evolutionary conserved mechanism. The key findings of this work are summarized below:

- We identified a thousand of novel intermediate-sized non-coding RNAs. In particular, we validated most annotated snoRNAs and identified 190 novel putative snoRNAs, making the most comprehensive annotations of zebrafish snoRNAs.
- We identified host genes for most (>85%) snoRNAs. Temporal dynamics of host genes revealed that many host genes are maternally transcribed and remain transcriptionally active during maternal-zygotic transition, suggesting their important roles during early embryogenesis.
- Transcriptional and epigenetic profiling showed no evidence for independent transcription of snoRNAs suggesting that they are co-transcribed with their host genes, which was further supported by sequence composition around TSSs.
- We showed indication that host (coding and non-coding) genes require non-canonical transcription initiation machinery that is associated with translation machinery.
- We showed that 5'-end of many snoRNAs overlaps with CAGE 5'-ends, suggesting either they are capped or undergo post-transcriptional modification, which is also evolutionary conserved in human snoRNAs.
- We further showed that small RNAs derived from snoRNAs (sd-snoRNAs) are generated from most snoRNAs and provide first evidence of sd-snoRNAs produced in oocytes, suggesting their potential importance during early embryogenesis.

### 2.3.3 Ideas for future work

The functional roles of many snoRNAs during early embryogenesis are largely uncharacterized. Considering many snoRNA host genes (both coding and non-coding) snoRNAs are transcribed maternally, but their importance during early embryogenesis remains largely unknown. However, a recent study has shown that disrupting the splicing of host (coding) genes or inhibiting snoRNA precursor processing led to severe morphological defects and embryonic lethality in zebrafish (Higa-Nakamine et al. 2012). Many snoRNAs are encoded in introns of lncRNA, but it remains to be seen how the disruption of lncRNAs or snoRNAs encoded in them affect development of zebrafish embryogenesis.

Do lncRNAs encoding snoRNAs have their own function or simply function as vehicles for transcription regulation in the generation of functional snoRNAs encoded in them? Many coding genes encoding snoRNAs are ribosomal protein genes, and have their independent function in addition to providing transcription initiation to snoRNAs. However, functional annotations of lncRNAs encoding snoRNAs are largely undocumented. We showed that host genes (both coding and non-coding) encoding snoRNAs tend to have TCT initiators, that is associated with translational machinery (Parry et al. 2010). Small nucleolar RNAs have roles in modifying rRNAs and are co-transcribed from genes associated with translation machinery, suggesting transcription of snoRNAs and translational machinery can be co-regulated process. As snoRNAs encoded in lncRNAs also have similar TCT motifs, it suggest for selection pressure for snoRNAs being the functions that require non-canonical translation associated transcription initiation mechanism. Triple nucleotide substitution mutations from -4 to +11 relative to C+1 start site, resulted in substantial reduction in transcription (Parry et al. 2010). It would be interesting to analyze, if poly-pyrimidine initiator is converted into a typical canonical promoter, how does the snoRNA biogenesis and translation machinery gets affected. It is also important to understand the composition of the transcriptional machinery involved in recognizing TCT initiator, which is expected to be different from the canonical transcription factor II D (TFIID). This in turn will help in understanding why translation associated genes have a distinct initiation system.

# 3. Discussion and perspective

### 3.1 What is non-coding RNA?

This dissertation addresses an open problem of how non-coding RNA genes are identified in genomic sequences and how they are regulated during embryogenesis. Our views and perceptions about non-coding RNAs have significantly changed over the last decade. We have moved on from the initial days, when, evolutionary signatures on genomic sequences were used to predict non-coding RNAs, which today is accomplished through extensive high-throughput sequencing. High-throughput sequencing also generates large amounts of pervasive transcripts. The amount of the functional RNAs within these pervasive transcripts is debated, while some have been skeptics and termed as a dark matter (van Bakel et al. 2010). While the recently concluded ENCODE project has revealed 80% of the genome contains elements linked to biochemical function (Bernstein et al. 2012) and regulatory elements (*cis*-regulatory elements, enhancers, promoters and non-coding RNAs) might constitute only about 20% of the genome. After the completion of the ENCODE project, some scholars have suggested a rethink on the definition of the “gene” itself. A precise definition of non-coding RNAs might still be challenging at this moment. However, in its simplicity, any transcribed RNAs that can regulate other genes (both coding or non-coding) or have their own regulatory functions can be termed non-coding RNAs.

Functional association of non-coding transcripts is a daunting task, due to their relative low expression (difficult to distinguish from noise), cell/tissue specific expression (necessary to profile across multiple samples), low sequence conservation and temporal expression (require transcriptional profiling across multiple time points). Analyses of non-coding transcripts in a few tissues and developmental time-points, and predicting them as a transcriptional noise might not be appropriate, as the assessed tissues or developmental time point might have been inappropriate. Large-scale studies, such as ENOCDE and modENCODE, analyzing developmental dynamics of transcripts across multiple tissues and time points, need to be analyzed exhaustively for its functionality, before these transcripts are called as junk. Here, we have presented additional evidence to show how many of these pervasive transcripts, located within gene body, are developmentally regulated, during zebrafish embryogenesis, suggesting gene-architecture is highly interlaced giving rise to multiple transcripts of various size and functions, providing additional evidence to what ENCODE project has recently concluded.

Whether non-coding RNAs have to be highly expressed and constitutively expressed to be functional is an open question in the field. Due to the very nature of non-coding RNAs, i.e. low and transient expression, some scholars believe it could be spurious transcripts, reflecting either an unstable transcripts or sequencing artifacts. While many others believe these transcripts to be a functional RNA, and I personally support the later notion.  Rather than thinking gene regulation as an on/off process, genomes are pervasively transcribed; that regulatory non-coding RNAs acting in *cis* complement transcription factors in gene regulation; that genes that are not expressed (which in the conventional sense are ‘‘off’’) are often associated with engaged RNA polymerase II, producing short non-coding transcripts at their promoters; and that genes are differentially marked to respond to a particular developmental program long before they are actually expressed.

## 3.2 Differentiating a classical promoter from non promoter class

We revealed an extraordinary dynamics in promoter usage (Paper-I), characterized by an entire hitherto of unknown layer of RNA species occurring in exons, introns and 3’-UTRs of developmentally active genes. Genome-wide analyses have revealed classical promoters are enriched for Py/Pu [C/T][A/G] as their initiation sequence at [-1+1] position (Carninci et al. 2006). Consistent with previous observation, most coding and non-coding genes have similar initiation motifs, shown in Paper I. We also identified recently reported TCT motif (Parry et al. 2010), on a subset of genes associated with translational machinery. In addition, we identified a novel AA initiator that is conserved on an orthologous subset of genes in zebrafish and human, and hence identifying a novel noncanonical initiation motif associated with promoters of coding genes and lncRNAs.

To differentiate pervasive transcripts from classical promoters, we compared the sequence and epigenetic marks around their 5’-ends and showed striking differences between them. We argue that they can be used to differentiate a genuine promoter from a non-promoter. Pervasive transcripts generated from exons have highly enriched GG as an initiation signals, which is also enriched on sequence reads mapping at splicing junction suggesting these pervasive transcripts must be post-transcriptionally generated. However, intronic tags have three different types of initiation signals, based on their position with reference to the host intron. As expected, 5’ and 3’ intronic tags have GT and AG initiation signals, while intra intronic tags are slightly enriched for Py/Pu, suggesting that can be associated with promoters of intronic miRNAs or other non-coding RNA promoters. Furthermore, we showed that internal CAGE signals and associated transcripts lack histone modification marks enriched for promoters and lack CpG islands, suggesting the areas around pervasive signals are different from classical promoters. Despite their developmentally regulated expression pattern, such pervasive transcripts did not work as promoter in a reporter construct, suggesting that they are likely to be associated with post-transcriptional processing. Hence, we have provided multiple lines of evidence to differentiate pervasive transcripts from classical promoters.

## 3.3 Functional roles of non-coding RNA during embryogenesis

Various non-coding RNAs have emerged as key players in wide range of biological processes. Nevertheless, our understanding of non-coding RNA during embryogenesis was limited to a few miRNAs (Giraldez et al. 2006; Tang et al. 2007), and functional studies of lncRNAs were never done before in animal model *in vivo* until recently. Two zebrafish lncRNAs, *cryano* and *megamind*, were successfully demonstrated to regulate the development of brain (Ulitsky et al. 2011). In paper I, we provide evidence for many intragenic and intergenic transcripts that have transient expression and are developmentally regulated during embryogenesis. The fact that many intragenic transcripts are maternally inherited and remain active even after the onset of zygotic genome activation emphasizes their importance during embryogenesis. Intragenic transcripts are located in the set of non-overlapping genes and are developmentally regulated, and yet their functional elucidation remains elusive without knock-down studies. Isolating intragenic transcripts without interfering the host transcripts is challenging, and it remains close to impossible, with the current art of technology, to isolate such intragenic transcripts on a genome-wide scale. However, functional studies on intergenic transcripts, giving rise to lncRNAs, can be carried out through morpholino knockdowns to block the transcript by targeting the conserved sequence or spliced sequence, as performed by (Ulitsky et al. 2011) on zebrafish lncRNAs. Experiments directed towards functional studies need large resources (in terms of manpower, finance and time) and hence such candidates should be chosen with caution.

### 3.4 Annotation difficulties of non-coding RNAs

The plethora of data generated by high-throughput sequencing technologies provides us an opportunity to characterize various classes of non-coding RNAs. Both the annotation and characterization (based on certain similar characteristics) of non-coding RNAs are hindered, possibly due to its diverse characteristics or due to incomplete understanding of its nature. However, a lot of progress has been made in last few years. Annotation of small RNAs is done with higher precision, with various available tools that exploit the sequence and structure information to predict them. lncRNA annotation is still problematic and erroneous. A recent ENCODE paper suggests most lncRNAs are inefficiently spliced (Tilgner et al. 2012). The annotations of lncRNAs are predominantly based on the assembled transcripts from RNA-Seq reads. The tools that assemble transcripts, mostly relies on canonical splice sites (GT-AG) to predict intron-exon junctions. As lncRNAs are inefficiently spliced (Tilgner et al. 2012), intron-exon junctions can differ upon coverage. The lack of proper canonical splice sites with variable coverage across intron-exon junctions makes it difficult to annotate intron-exon junction with high precision. This uncertainty in the prediction of intron-exon junction can influence the way cDNA is computationally assembled (as inefficient splicing can insert few bases from the intron). The annotations of lncRNAs are mostly relied on its inability to code for long ORF (<100 aa). While the inclusion of a couple of bases from intron or exclusion of couple of bases from exon can introduce a stop reading frame. The inclusion of undesired stop reading frames gives the impression of short ORF and hence likely a lncRNA candidate. Since there might be no immediate solution, but one needs to be cautious and the computational annotation needs to be manually curated, similar to the GENCODE annotation from the ENCODE project (Derrien et al. 2012; Harrow et al. 2012).

### *3.5 Important resource for the zebrafish community*

The zebrafish is one of the most widely used vertebrate model organisms, attracting scholars ranging from basic biologists to clinicians in >1000 laboratories worldwide. Our high quality data regarding the transcriptional start sites will serve as a huge resource for studying gene regulation in development and the dynamics of promoter use across vertebrates. In addition to the high resolution of the data, as well as the diversity of novel findings on core promoters, this work may become a classic resource for studying transcriptional control in vertebrate development. The resource data generated in this study provides a range of practical applications and benefits. Understanding core promoter regulation is central to the informed choice of core promoter for transgene assays designed either to control cell type-specific activities (fluorescence reporter labeling) or to detect and functionally characterize *cis*-regulatory modules (e.g. enhancer trapping and enhancer tests). Widespread usage of alternative promoters during development suggests pervasive variability of genes 5'-UTR sequences, with implications for translation start site selection during development. The identification and description of the developmental dynamics of evolutionary conserved classes of novel ncRNAs will aid in exploiting the zebrafish model in the search for the function and genetic signatures of a variety of non-coding RNAs, which are proposed to be important regulators of development and are likely sources of mutations associated with human disease. Furthermore, the correct identification of core promoters will be critical for finding non-coding mutations that affect development and may lead to phenotypes suitable for modeling disease. This variability should be taken into account in techniques widely used in disease modeling with zebrafish (Eisen and Smith 2008), such as the design of translation blocking knock-down reagents (e.g. translational start site targeting morpholino antisense oligonucleotides) or the generation of site-specific mutations. It provides key 5'-end data needed for core promoter design in transgenic applications and for translation blocking morpholino knockdown – widely used by the zebrafish community. We would like to note that this work will impact not only fish developmental biology, but much broader fields. Zebrafish is used in numerous assays, including those for testing mammalian *cis*-regulatory elements and gene function. Moreover, the uncovered developmental dynamics of several classes of non-coding RNAs will be useful for a range of scholars in the fields of genomics, transcription biology, and developmental genetics. In conclusion, the high-resolution transcription initiation resource presented here provides foundation for the analysis of transcription initiation complexes on core promoters during development and the elucidation of developmental codes of transcription initiation in vertebrates.

**References**

Aanes H, Winata CL, Lin CH, Chen JP, Srinivasan KG, Lee SG, Lim AY, Hajan HS, Collas P, Bourque G et al. 2011. Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. *Genome Res* **21**(8): 1328-1338.

Akalin A, Fredman D, Arner E, Dong X, Bryne JC, Suzuki H, Daub CO, Hayashizaki Y, Lenhard B. 2009. Transcriptional features of genomic regulatory blocks. *Genome Biol* **10**(4): R38.

Akkers RC, van Heeringen SJ, Jacobi UG, Janssen-Megens EM, Francoijs KJ, Stunnenberg HG, Veenstra GJ. 2009. A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. *Dev Cell* **17**(3): 425-434.

Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. 2011. lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res* **39**(Database issue): D146-151.

Ambros V. 1989. A hierarchy of regulatory genes controls a larva-to-adult developmental switch in C. elegans. *Cell* **57**(1): 49-57.

Anokye-Danso F, Trivedi CM, Juhr D, Gupta M, Cui Z, Tian Y, Zhang Y, Yang W, Gruber PJ, Epstein JA et al. 2011. Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. *Cell Stem Cell* **8**(4): 376-388.

Arrial RT, Togawa RC, Brigido Mde M. 2009. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. *BMC Bioinformatics* **10**: 239.

Bachellerie JP, Cavaille J, Huttenhofer A. 2002. The expanding snoRNA world. *Biochimie* **84**(8): 775-790.

Badger JH, Olsen GJ. 1999. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* **16**(4): 512-524.

Banfai B, Jia H, Khatun J, Wood E, Risk B, Gundling WE, Jr., Kundaje A, Gunawardena HP, Yu Y, Xie L et al. 2012. Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res* **22**(9): 1646-1657.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**(4): 823-837.

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**(2): 281-297.

Bazzini AA, Giraldez AJ. 2011. MicroRNAs sculpt gene expression in embryonic development: new insights from plants. *Dev Cell* **20**(1): 3-4.

Bazzini AA, Lee MT, Giraldez AJ. 2012. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science* **336**(6078): 233-237.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* **304**(5675): 1321-1325.

Beltran M, Puig I, Pena C, Garcia JM, Alvarez AB, Pena R, Bonilla F, de Herreros AG. 2008. A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev* **22**(6): 756-769.

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Bernstein BE, Humphrey EL, Erlich RL, Schneider R, Bouman P, Liu JS, Kouzarides T, Schreiber SL. 2002. Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc Natl Acad Sci U S A* **99**(13): 8695-8700.

Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ, 3rd, Gingeras TR et al. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**(2): 169-181.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**(2): 315-326.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**(5705): 2242-2246.

Birney E Stamatoyannopoulos JA Dutta A Guigo R Gingeras TR Margulies EH Weng Z Snyder M Dermitzakis ET Thurman RE et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146): 799-816.

Boffelli D, Nobrega MA, Rubin EM. 2004. Comparative genomics at the vertebrate extremes. *Nat Rev Genet* **5**(6): 456-465.

Bogdanovic O, Fernandez-Minan A, Tena JJ, de la Calle-Mustienes E, Hidalgo C, van Kruysbergen I, van Heeringen SJ, Veenstra GJ, Gomez-Skarmeta JL. 2012. Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Res.*

Borchert GM, Lanier W, Davidson BL. 2006. RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol* **13**(12): 1097-1101.

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**(6): 947-956.

Braidotti G, Baubec T, Pauler F, Seidl C, Smrzka O, Stricker S, Yotova I, Barlow DP. 2004. The Air noncoding RNA: an imprinted cis-silencing transcript. *Cold Spring Harb Symp Quant Biol* **69**: 55-66.

Brenner S, Jacob F, Meselson M. 1961. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190**: 576-581.

Bushati N, Stark A, Brennecke J, Cohen SM. 2008. Temporal reciprocity of miRNAs and their targets during the maternal-to-zygotic transition in Drosophila. *Curr Biol* **18**(7): 501-506.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding

RNAs reveals global properties and specific subclasses. *Genes Dev* **25**(18): 1915-1927.

Calin GA, Croce CM. 2006. MicroRNA signatures in human cancers. *Nat Rev Cancer* **6**(11): 857-866.

Calin GA, Liu CG, Ferracin M, Hyslop T, Spizzo R, Sevignani C, Fabbri M, Cimmino A, Lee EJ, Wojcik SE et al. 2007. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* **12**(3): 215-229.

Carninci P Kasukawa T Katayama S Gough J Frith MC Maeda N Oyama R Ravasi T Lenhard B Wells C et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309**(5740): 1559-1563.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**(6): 626-635.

Chen PY, Manninga H, Slanchev K, Chien MC, Russo JJ, Ju JY, Sheridan R, John B, Marks DS, Gaidatzis D et al. 2005. The developmental miRNA profiles of zebrafish as determined by small RNA cloning. *Gene Dev* **19**(11): 1288-1293.

Clote P, Ferre F, Kranakis E, Krizanc D. 2005. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *Rna-a Publication of the Rna Society* **11**(5): 578-591.

Cole C, Sobala A, Lu C, Thatcher SR, Bowman A, Brown JW, Green PJ, Barton GJ, Hutvagner G. 2009. Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA* **15**(12): 2147-2160.

Collins JE, White S, Searle SM, Stemple DL. 2012. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. *Genome Res*.

de la Calle-Mustienes E, Feijoo CG, Manzanares M, Tena JJ, Rodriguez-Seguel E, Letizia A, Allende ML, Gomez-Skarmeta JL. 2005. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res* **15**(8): 1061-1072.

Deng W, Zhu X, Skogerbo G, Zhao Y, Fu Z, Wang Y, He H, Cai L, Sun H, Liu C et al. 2006. Organization of the Caenorhabditis elegans small non-coding transcriptome: genomic features, biogenesis, and expression. *Genome Res* **16**(1): 20-29.

Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ. 2004. Processing of primary microRNAs by the Microprocessor complex. *Nature* **432**(7014): 231-235.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* **22**(9): 1775-1789.

Dieci G, Fiorino G, Castelnuovo M, Teichmann M, Pagano A. 2007. The expanding RNA polymerase III transcriptome. *Trends Genet* **23**(12): 614-622.

Ding J, Lu Q, Ouyang Y, Mao H, Zhang P, Yao J, Xu C, Li X, Xiao J, Zhang Q. 2012. A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. *Proc Natl Acad Sci U S A* **109**(7): 2654-2659.

Dinger ME, Pang KC, Mercer TR, Mattick JS. 2008. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol* **4**(11): e1000176.

Djuranovic S, Nahvi A, Green R. 2011. A parsimonious model for gene regulation by miRNAs. *Science* **331**(6017): 550-553.

-. 2012. miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science* **336**(6078): 237-240.

Dong X, Navratilova P, Fredman D, Drivenes O, Becker TS, Lenhard B. 2010. Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons. *Nucleic Acids Res* **38**(4): 1071-1085.

Eisen JS, Smith JC. 2008. Controlling morpholino experiments: don't stop making antisense. *Development* **135**(10): 1735-1743.

Ender C, Krek A, Friedlander MR, Beitzinger M, Weinmann L, Chen W, Pfeffer S, Rajewsky N, Meister G. 2008. A human snoRNA with microRNA-like functions. *Mol Cell* **32**(4): 519-528.

Engstrom PG, Fredman D, Lenhard B. 2008. Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome Biol* **9**(2): R34.

Engstrom PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res* **17**(12): 1898-1908.

Engstrom PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G, Brozzi A, Luzi L, Tan SL, Yang L et al. 2006. Complex Loci in human and mouse genomes. *PLoS Genet* **2**(4): e47.

Erdmann VA, Barciszewska MZ, Szymanski M, Hochberg A, de Groot N, Barciszewski J. 2001. The non-coding RNAs as riboregulators. *Nucleic Acids Res* **29**(1): 189-193.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**(8): 817-825.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**(7345): 43-49.

Ewing B, Green P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* **25**(2): 232-234.

Fabian MR, Sonenberg N, Filipowicz W. 2010. Regulation of mRNA translation and stability by microRNAs. *Annu Rev Biochem* **79**: 351-379.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**(5): 563-571.

Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon G, Philipp K, Sylvain F, Willingham AT, Duttagupta R, Dumais E et al. 2009. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**(7232): 1028-1032.

Forrest AR, Abdelhamid RF, Carninci P. 2009. Annotating non-coding transcription using functional genomics strategies. *Brief Funct Genomic Proteomic* **8**(6): 437-443.

Gerstein MB Lu ZJ Van Nostrand EL Cheng C Arshinoff BI Liu T Yip KY Robilotto R Rechtsteiner A Ikegami K et al. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**(6012): 1775-1787.

Gibbs RA Weinstock GM Metzker ML Muzny DM Sodergren EJ Scherer S Scott G Steffen D Worley KC Burch PE et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**(6982): 493-521.

Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ, Schier AF. 2006. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* **312**(5770): 75-79.

Girard A, Sachidanandam R, Hannon GJ, Carmell MA. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**(7099): 199-202.

Glazov EA, Pheasant M, McGraw EA, Bejerano G, Mattick JS. 2005. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res* **15**(6): 800-808.

Griffiths-Jones S. 2007. Annotating noncoding RNA genes. *Annu Rev Genomics Hum Genet* **8**: 279-298.

Grivna ST, Beyret E, Wang Z, Lin H. 2006. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev* **20**(13): 1709-1714.

Guffanti A, Iacono M, Pelucchi P, Kim N, Solda G, Croft LJ, Taft RJ, Rizzi E, Askarian-Amiri M, Bonnal RJ et al. 2009. A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. *BMC Genomics* **10**: 163.

Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL et al. 2010. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**(7291): 1071-1076.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**(7235): 223-227.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**(7364): 295-300.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**(5): 503-510.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22**(9): 1760-1774.

He X, Yan YL, DeLaurier A, Postlethwait JH. 2011. Observation of miRNA gene expression in zebrafish embryos by in situ hybridization to microRNA primary transcripts. *Zebrafish* **8**(1): 1-8.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**(3): 311-318.

Henras AK, Dez C, Henry Y. 2004. RNA structure and function in C/D and H/ACA s(no)RNPs. *Current Opinion in Structural Biology* **14**(3): 335-343.

Higa-Nakamine S, Suzuki T, Uechi T, Chakraborty A, Nakajima Y, Nakamura M, Hirano N, Suzuki T, Kenmochi N. 2012. Loss of ribosomal RNA modification causes developmental defects in zebrafish. *Nucleic Acids Res* **40**(1): 391-398.

Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH et al. 2011. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res* **21**(2): 182-192.

Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov DV, Blaser H, Raz E, Moens CB et al. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* **129**(1): 69-82.

Hu W, Yuan B, Flygare J, Lodish HF. 2011. Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation. *Genes Dev* **25**(24): 2573-2578.

Huttenhofer A, Brosius J, Bachellerie JP. 2002. RNomics: identification and function of small, non-messenger RNAs. *Current Opinion in Chemical Biology* **6**(6): 835-843.

Impey S, McCorkle SR, Cha-Molstad H, Dwyer JM, Yochum GS, Boss JM, McWeeney S, Dunn JJ, Mandel G, Goodman RH. 2004. Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell* **119**(7): 1041-1054.

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**(5924): 218-223.

Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**(4): 789-802.

Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**: 318-356.

Jacquier A. 2009. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet* **10**(12): 833-844.

Jalali S, Jayaraj GG, Scaria V. 2012. Integrative transcriptome analysis suggest processing of a subset of long non-coding RNAs to small RNAs. *Biol Direct* **7**: 25.

Jan CH, Friedman RC, Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* **469**(7328): 97-101.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**(5830): 1497-1502.

Jones-Rhoades MW, Bartel DP. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell* **14**(6): 787-799.

Juven-Gershon T, Kadonaga JT. 2010. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol* **339**(2): 225-229.

Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G et al. 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* **14**(3): 331-342.

Kane DA, Kimmel CB. 1993. The zebrafish midblastula transition. *Development* **119**(2): 447-456.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**(5569): 916-919.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**(5830): 1484-1488.

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J et al. 2005. Antisense transcription in the mammalian transcriptome. *Science* **309**(5740): 1564-1566.

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**(28): 11667-11672.

Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engstrom PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K et al. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res* **17**(5): 545-555.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**(7295): 182-187.

Kim VN, Han J, Siomi MC. 2009. Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* **10**(2): 126-139.

Kiss AM, Jady BE, Bertrand E, Kiss T. 2004. Human box H/ACA pseudouridylation guide RNA machinery. *Molecular and Cellular Biology* **24**(13): 5797-5807.

Kiss T. 2002. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell* **109**(2): 145-148.

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M et al. 2006. CAGE: cap analysis of gene expression. *Nat Methods* **3**(3): 211-222.

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35**(Web Server issue): W345-349.

Kornberg RD. 1999. Eukaryotic transcriptional control. *Trends Cell Biol* **9**(12): M46-49.

Kozomara A, Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**(Database issue): D152-157.

Kretz M, Webster DE, Flockhart RJ, Lee CS, Zehnder A, Lopez-Pajares V, Qu K, Zheng GX, Chow J, Kim GE et al. 2012. Suppression of progenitor differentiation requires the long noncoding RNA ANCR. *Genes Dev* **26**(4): 338-343.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294**(5543): 853-858.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Langenberger D, Bermudez-Santana C, Hertel J, Hoffmann S, Khaitovich P, Stadler PF. 2009. Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics* **25**(18): 2298-2301.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3): R25.

Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**(5543): 858-862.

Lee RC, Ambros V. 2001. An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**(5543): 862-864.

Lee RC, Feinbaum RL, Ambros V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**(5): 843-854.

Lee Y, Jeon K, Lee JT, Kim S, Kim VN. 2002. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J* **21**(17): 4663-4670.

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN. 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* **23**(20): 4051-4060.

Lenhard B, Sandelin A, Carninci P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* **13**(4): 233-245.

Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**(14): 1725-1735.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**(11): 1851-1858.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**(15): 1966-1967.

Li Z, Ender C, Meister G, Moore PS, Chang Y, John B. 2012. Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs. *Nucleic Acids Res*.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. 2000. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet* **25**(2): 239-240.

Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. 2003. Vertebrate microRNA genes. *Science* **299**(5612): 1540.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**(13): i275-282.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**(7370): 476-482.

Liu N, Williams AH, Kim Y, McAnally J, Bezprozvannaya S, Sutherland LB, Richardson JA, Bassel-Duby R, Olson EN. 2007. An intragenic MEF2-dependent enhancer directs muscle-specific expression of microRNAs 1 and 133. *Proc Natl Acad Sci U S A* **104**(52): 20844-20849.

Liu TT, Zhu D, Chen W, Deng W, He H, He G, Bai B, Qi Y, Chen R, Deng XW. 2012. A Global Identification and Analysis of Small Nucleolar RNAs and Possible Intermediate-Sized Non-Coding RNAs in Oryza sativa. *Mol Plant*.

Lu ZJ, Yip KY, Wang G, Shou C, Hillier LW, Khurana E, Agarwal A, Auerbach R, Rozowsky J, Cheng C et al. 2011. Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res* **21**(2): 276-285.

Lund E, Liu M, Hartley RS, Sheets MD, Dahlberg JE. 2009. Deadenylation of maternal mRNAs mediated by miR-427 in Xenopus laevis embryos. *RNA* **15**(12): 2351-2363.

Maeso I, Irimia M, Tena JJ, Gonzalez-Perez E, Tran D, Ravi V, Venkatesh B, Campuzano S, Gomez-Skarmeta JL, Garcia-Fernandez J. 2012. An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. *Genome Res* **22**(4): 642-655.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057): 376-380.

Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J et al. 2008. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**(3): 521-533.

Matera AG, Terns RM, Terns MP. 2007. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol* **8**(3): 209-220.

Mathavan S, Lee SG, Mak A, Miller LD, Murthy KR, Govindarajan KR, Tong Y, Wu YL, Lam SH, Yang H et al. 2005. Transcriptome analysis of zebrafish embryogenesis using microarrays. *PLoS Genet* **1**(2): 260-276.

Meier UT. 2005. The many facets of H/ACA ribonucleoproteins. *Chromosoma* **114**(1): 1-14.

Meister G, Tuschl T. 2004. Mechanisms of gene silencing by double-stranded RNA. *Nature* **431**(7006): 343-349.

Mercer TR, Dinger ME, Bracken CP, Kolle G, Szubert JM, Korbie DJ, Askarian-Amiri ME, Gardiner BB, Goodall GJ, Grimmond SM et al. 2010. Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome. *Genome Res* **20**(12): 1639-1650.

Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**(3): 155-159.

Mercer TR, Wilhelm D, Dinger ME, Solda G, Korbie DJ, Glazov EA, Truong V, Schwenke M, Simons C, Matthaei KI et al. 2011. Expression of distinct RNAs from 3' untranslated regions. *Nucleic Acids Res* **39**(6): 2393-2403.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**(7153): 553-560.

Mondal T, Rasmussen M, Pandey GK, Isaksson A, Kanduri C. 2010. Characterization of the RNA content of chromatin. *Genome Res* **20**(7): 899-907.

Mourelatos Z, Dostie J, Paushkin S, Sharma A, Charroux B, Abel L, Rappsilber J, Mann M, Dreyfuss G. 2002. miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev* **16**(6): 720-728.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**(5881): 1344-1349.

Nam JW, Bartel D. 2012. Long non-coding RNAs in C. elegans. *Genome Res*.

Nodine MD, Bartel DP. 2010. MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis. *Genes Dev* **24**(23): 2678-2692.

Ohler U, Liao GC, Niemann H, Rubin GM. 2002. Computational analysis of core promoters in the Drosophila genome. *Genome Biol* **3**(12): RESEARCH0087.

Ohno S. 1972. So much "junk" DNA in our genome. *Brookhaven Symp Biol* **23**: 366-370.

Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* **284**(5757): 604-607.

Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q et al. 2010. Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**(1): 46-58.

Pagano A, Castelnuovo M, Tortelli F, Ferrari R, Dieci G, Cancedda R. 2007. New small nuclear RNA gene-like transcriptional units as sources of regulatory transcripts. *PLoS Genet* **3**(2): e1.

Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* **22**(1): 1-5.

Papatsenko D, Kislyuk A, Levine M, Dubchak I. 2006. Conservation patterns in different functional sequence categories of divergent Drosophila species. *Genomics* **88**(4): 431-442.

Parry TJ, Theisen JW, Hsu JY, Wang YL, Corcoran DL, Eustice M, Ohler U, Kadonaga JT. 2010. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev* **24**(18): 2013-2018.

Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Muller P et al. 2000. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**(6808): 86-89.

Pauli A, Rinn JL, Schier AF. 2011. Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet* **12**(2): 136-149.

Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A et al. 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* **22**(3): 577-591.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2**(4): e33.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**(7118): 499-502.

Ponjavic J, Lenhard B, Kai C, Kawai J, Carninci P, Hayashizaki Y, Sandelin A. 2006. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol* **7**(8): R78.

Ponjavic J, Oliver PL, Lunter G, Ponting CP. 2009. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet* **5**(8): e1000617.

Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17**(5): 556-565.

Preker P, Almvig K, Christensen MS, Valen E, Mapendano CK, Sandelin A, Jensen TH. 2011. PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic Acids Res* **39**(16): 7179-7193.

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**(6772): 901-906.

Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP. 2002. MicroRNAs in plants. *Genes Dev* **16**(13): 1616-1626.

Rinn JL, Chang HY. 2012. Genome Regulation by Long Noncoding RNAs. *Annu Rev Biochem* **81**: 145-166.

Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M et al. 2003. The transcriptional activity of human Chromosome 22. *Genes Dev* **17**(4): 529-540.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**(7): 1311-1323.

Rivas E, Eddy SR. 2001. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2**: 8.

Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Res* **14**(10A): 1902-1910.

Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF et al. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**(6012): 1787-1797.

Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell* **127**(6): 1193-1207.

Russell J, Zomerdijk JC. 2006. The RNA polymerase I transcription machinery. *Biochem Soc Symp*(73): 203-216.

Sabarinadh C, Subramanian S, Tripathi A, Mishra RK. 2004. Extreme conservation of noncoding DNA near HoxD complex of vertebrates. *BMC Genomics* **5**: 75.

Sandelin A, Bailey P, Bruce S, Engstrom PG, Klos JM, Wasserman WW, Ericson J, Lenhard B. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**(1): 99.

Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA. 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* **8**(6): 424-436.

Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NC, Schreiber SL, Mellor J, Kouzarides T. 2002. Active genes are tri-methylated at K4 of histone H3. *Nature* **419**(6905): 407-411.

Saxena A, Carninci P. 2011. Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. *Bioessays* **33**(11): 830-839.

Schier AF. 2007. The maternal-zygotic transition: death and birth of RNAs. *Science* **316**(5823): 406-407.

Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. 2008. Divergent transcription from active promoters. *Science* **322**(5909): 1849-1851.

Sheik Mohamed J, Gaughwin PM, Lim B, Robson P, Lipovich L. 2010. Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. *RNA* **16**(2): 324-337.

Shi W, Hendrix D, Levine M, Haley B. 2009. A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. *Nat Struct Mol Biol* **16**(2): 183-189.

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* **100**(26): 15776-15781.

Sims RJ, 3rd, Mandal SS, Reinberg D. 2004. Recent highlights of RNA-polymerase-II-mediated transcription. *Curr Opin Cell Biol* **16**(3): 263-271.

Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. 2011. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res* **21**(10): 1728-1737.

Stark BC, Kole R, Bowman EJ, Altman S. 1978. Ribonuclease P: an enzyme with an essential RNA component. *Proc Natl Acad Sci U S A* **75**(8): 3717-3721.

Storz G, Altuvia S, Wassarman KM. 2005. An abundance of RNA regulators. *Annu Rev Biochem* **74**: 199-217.

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**(5891): 956-960.

Svoboda P, Flemr M. 2010. The role of miRNAs and endogenous siRNAs in maternal-to-zygotic reprogramming and the establishment of pluripotency. *EMBO Rep* **11**(8): 590-597.

Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest AR, Grimmond SM, Schroder K et al. 2009a. Tiny RNAs associated with transcription start sites in animals. *Nat Genet* **41**(5): 572-578.

Taft RJ, Glazov EA, Lassmann T, Hayashizaki Y, Carninci P, Mattick JS. 2009b. Small RNAs derived from snoRNAs. *RNA* **15**(7): 1233-1240.

Taft RJ, Pheasant M, Mattick JS. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* **29**(3): 288-299.

Taft RJ, Simons C, Nahkuri S, Oey H, Korbie DJ, Mercer TR, Holst J, Ritchie W, Wong JJ, Rasko JE et al. 2010. Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nat Struct Mol Biol* **17**(8): 1030-1034.

Tang F, Kaneda M, O'Carroll D, Hajkova P, Barton SC, Sun YA, Lee C, Tarakhovsky A, Lao K, Surani MA. 2007. Maternal microRNAs are essential for mouse zygotic development. *Genes Dev* **21**(6): 644-648.

Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigo R. 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* **22**(9): 1616-1625.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9): 1105-1111.

Tycowski KT, Aab A, Steitz JA. 2004. Guide RNAs with 5 ' caps and novel box C/D snoRNA-like domains for modification of snRNAs in metazoa. *Current Biology* **14**(22): 1985-1995.

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**(7): 1537-1550.

Valen E, Preker P, Andersen PR, Zhao X, Chen Y, Ender C, Dueck A, Meister G, Sandelin A, Jensen TH. 2011. Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nat Struct Mol Biol* **18**(9): 1075-1082.

van Bakel H, Nislow C, Blencowe BJ, Hughes TR. 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol* **8**(5): e1000371.

van Nimwegen E, Crutchfield JP, Huynen M. 1999. Neutral evolution of mutational robustness. *Proc Natl Acad Sci U S A* **96**(17): 9716-9720.

Vastenhouw NL, Zhang Y, Woods IG, Imam F, Regev A, Liu XS, Rinn J, Schier AF. 2010. Chromatin signature of embryonic pluripotency is established during genome activation. *Nature* **464**(7290): 922-926.

Venter JC Adams MD Myers EW Li PW Mural RJ Sutton GG Smith HO Yandell M Evans CA Holt RA et al. 2001. The sequence of the human genome. *Science* **291**(5507): 1304-1351.

Visel A, Akiyama JA, Shoukry M, Afzal V, Rubin EM, Pennacchio LA. 2009a. Functional autonomy of distant-acting human enhancers. *Genomics* **93**(6): 509-513.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F et al. 2009b. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**(7231): 854-858.

Wang Y, Chen J, Wei G, He H, Zhu X, Xiao T, Yuan J, Dong B, He S, Skogerbo G et al. 2011. The Caenorhabditis elegans intermediate-size transcriptome shows high degree of stage-specific expression. *Nucleic Acids Res* **39**(12): 5203-5214.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**(1): 57-63.

Wapinski O, Chang HY. 2011. Long noncoding RNAs and human disease. *Trends Cell Biol* **21**(6): 354-361.

Washietl S, Hofacker IL. 2004. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *Journal of Molecular Biology* **342**(1): 19-30.

Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America* **102**(7): 2454-2459.

Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermuller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A et al. 2007. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res* **17**(6): 852-864.

Watanabe T, Takeda A, Mise K, Okuno T, Suzuki T, Minami N, Imai H. 2005. Stage-specific expression of microRNAs during Xenopus development. *FEBS Lett* **579**(2): 318-324.

Waterston RH Lindblad-Toh K Birney E Rogers J Abril JF Agarwal P Agarwala R Ainscough R Alexandersson M An P et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915): 520-562.

Wei C, Salichos L, Wittgrove CM, Rokas A, Patton JG. 2012. Transcriptome-wide analysis of small RNA expression in early zebrafish development. *RNA* **18**(5): 915-929.

Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, Shahab A, Yong HC, Fu Y, Weng Z et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124**(1): 207-219.

Wightman B, Burglin TR, Gatto J, Arasu P, Ruvkun G. 1991. Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development. *Genes Dev* **5**(10): 1813-1824.

Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, Aza-Blanc P, Hogenesch JB, Schultz PG. 2005. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**(5740): 1570-1573.

Wong GK, Passey DA, Yu J. 2001. Most of the human genome is transcribed. *Genome Res* **11**(12): 1975-1977.

Woods K, Thomson JM, Hammond SM. 2007. Direct regulation of an oncogenic micro-RNA cluster by E2F transcription factors. *J Biol Chem* **282**(4): 2130-2134.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**(1): e7.

Xi H, Yu Y, Fu Y, Foley J, Halees A, Weng Z. 2007. Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Res* **17**(6): 798-806.

Yan D, He D, He S, Chen X, Fan Z, Chen R. 2011. Identification and analysis of intermediate size noncoding RNAs in the human fetal brain. *PLoS One* **6**(7): e21652.

Yochum GS, Cleland R, McWeeney S, Goodman RH. 2007. An antisense transcript induced by Wnt/beta-catenin signaling decreases E2F4. *J Biol Chem* **282**(2): 871-878.

Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, Liu JL, Ponting CP. 2012. Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome. *Genome Biol Evol* **4**(4): 427-442.

Zhang Y, Wang J, Huang S, Zhu X, Liu J, Yang N, Song D, Wu R, Deng W, Skogerbo G et al. 2009. Systematic identification and characterization of chicken (Gallus gallus) ncRNAs. *Nucleic Acids Res* **37**(19): 6562-6574.

Zhong X, Coukos G, Zhang L. 2012. miRNAs in human cancer. *Methods Mol Biol* **822**: 295-306.

Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* **12**(1): 7-18.

Zieve GW. 1981. Two groups of small stable RNAs. *Cell* **25**(2): 296-297.