# Chromatin and transcriptome-based integrative approaches to profile functional long noncoding RNAs

## *- A computational approach*

Santhilal Subhash

UNIVERSITY OF GOTHENBURG

Department of Medical Biochemistry and Cell Biology,
Institute of Biomedicine at Sahlgrenska Academy
University of Gothenburg

Gothenburg, Sweden, 2020

Cover illustration: Long noncoding RNA molecule (center) as a central player in mammalian development (top), regulating cancer cell hallmarks (left), and modulating chromatin structures (right).

by **Santhilal Subhash**

**Chromatin and transcriptome-based integrative approaches to profile functional long noncoding RNAs – A computational approach**

© Santhilal Subhash 2020

santhilal.subhash@gu.se

ISBN 978-91-8009-062-9 (PRINT)
ISBN 978-91-8009-063-6 (PDF)

Printed in Borås, Sweden 2020
Printed by Stema Specialtryck AB

"Without your involvement you cannot succeed. With your involvement you cannot fail." – **Dr. A.P.J. Abdul Kalam**

**This thesis is dedicated to my family**

# Chromatin and transcriptome-based integrative approaches to profile functional long noncoding RNAs

## - A computational approach

**Santhilal Subhash**

Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg
Gothenburg, Sweden

## ABSTRACT

One of the major hallmarks of cancer is aberrant or uncontrollable cell division, which occurs due to a defective cell cycle process. During the synthesis phase (S-phase) of the cell cycle, before cell division or mitosis phase, the DNA in the cell makes a new copy to pass on genetic information to the daughter cells. Therefore, S-phase is one of the crucial steps for a successful cell division to occur. The DNA in the nucleus is wrapped around a set of proteins called histones, forming nucleosomes, and multiple nucleosomes give rise to the higher-order chromatin structure. This well-established chromatin structure determines which portion of DNA or gene gets activated or suppressed by switching between open or closed chromatin states. Tri- or di-methylation of lysine 4 from histone 3 (H3K4me2/3) leads to open chromatin, which in turn promotes active gene transcription. The product of gene transcription is either protein-coding mRNA that translates into protein for its function or noncoding RNA, which do not code for any protein and function as RNA. However, the human genome project has identified that protein-coding genes only constitute 2% of the genome, and the vast majority of it is noncoding. Unlike protein-coding genes, the significance of RNAs transcribed from the noncoding genome is not well-established. Apart from housekeeping noncoding RNAs (rRNA, tRNA, snRNA, and snoRNA) and microRNAs (miRNAs), most functional noncoding RNAs fall into the long noncoding RNA (lncRNA) category.

In this thesis, we implemented comprehensive computational approaches to identify functionally relevant lncRNAs by analyzing chromatin and transcriptome-based sequencing datasets. In the first study **(paper I)**, using a transcriptome approach, we profiled lncRNAs associated with the S-phase

stage of the cell cycle. We demonstrated the oncogenic properties of various S-phase associated lncRNAs in multiple cancers.

Earlier, studies proposed that chromatin-associated RNAs, with the help of chromatin-modifying enzymes, determines the active/open or close chromatin status to promote or suppress gene transcription. Hence, in the second study **(paper II)**, we used chromatin-based approaches to propose a possible mechanism through which the active chromatin-associated lncRNAs may function. We show that active chromatin-associated lncRNAs regulate their partner genes *in-cis* by recruiting the WDR5 chromatin modifier to establish an open chromatin structure at the partner protein-coding gene promoters.

In our third study **(paper III)**, we integrated both transcriptome and chromatin-based approaches to find early development-associated lncRNAs. Here, we focused on tracing the molecular footprints of sperm lncRNAs throughout the stages of organismal development. For this purpose, we integrated datasets from gametes, preimplantation and post-implantation stages of an embryo. Interestingly, we observed distinct chromatin structures in the sperm. Also, sperm lncRNAs were active during the onset of zygotic genome activation in the preimplantation stages and in cancers. In summary, this study reveals a unique set of sperm-specific lncRNAs that are temporally activated during preimplantation stages and also aberrantly expressed in multiple cancers.

Overall, the present thesis provides an extensive catalogue of functionally relevant lncRNAs that can take part in cell cycle regulation, cancer, chromatin modulation, and organism development. Our studies can serve as a comprehensive resource for future investigations on lncRNAs.

**Keywords**: Computational biology, Long noncoding RNAs, Chromatin, Histone, Cell Cycle, S-phase, Cancer, Epigenetics, ChIP-seq, RNA-seq, ChRIP-seq, WDR5, H3K4me2, DNA methylation, Histone modifications.

ISBN 978-91-8009-062-9 (PRINT)
ISBN 978-91-8009-063-6 (PDF)

# SAMMANFATTNING PÅ SVENSKA

Ett av de viktigaste kännetecknen för cancer är avvikande eller okontrollerbar celldelning, vilket sker på grund av en defekt cellcykelprocess. Under syntesfasen (S-fas) av cellcykeln, innan celldelning eller mitosfas, gör cellen en ny kopia av DNAt, för att förmedla genetisk information till dottercellerna. Därför är S-fasen ett av de avgörande stegen för att en framgångsrik celldelning ska ske. DNAt i kärnan lindas runt en uppsättning proteiner som kallas histoner och bildar nukleosomer och flera nukleosomer ger upphov till kromatinstrukturen av högre ordning. Denna väletablerade kromatinstruktur avgör vilken del av DNAt eller gener som aktiveras eller undertrycks, genom att växla mellan öppet eller slutet kromatintillstånd. Tri- eller di-metylering av lysin 4 på histon 3 (H3K4me2 / 3) leder till öppet kromatin, vilket i sin tur främjar aktiv gentranskription. Produkten av gentranskription är antingen proteinkodande mRNA som översätts till protein för dess funktion eller icke-kodande RNA, som inte kodar för något protein och har sin funktion som RNA. Emellertid har humant genomprojekt identifierat att proteinkodande gener endast utgör 2% av genomet, och den stora majoriteten av det är icke proteinkodande. Till skillnad från proteinkodande gener är betydelsen av RNA som transkriberas från det icke-kodande genomet, inte väl etablerad. Förutom cellens vanliga underhåll av icke-kodande RNA (rRNA, tRNA, snRNA och snoRNA) och mikroRNA (miRNA) faller de flesta funktionella icke-kodande RNA i kategorin långa icke-kodande RNA (lncRNA).

I denna avhandling implementerade vi omfattande beräkningsmetoder för att identifiera funktionellt relevanta lncRNA genom att analysera kromatin- och transkriptombaserade sekvenseringsdatamängder. I den första studien (**artikel I**), med hjälp av ett transkriptom-tillvägagångssätt, profilerade vi lncRNAs associerade med S-fas-steget i cellcykeln. Vi demonstrerade de onkogena egenskaperna hos olika S-fasassocierade lncRNA i flera cancerformer.

Tidigare studier föreslog att kromatin-associerade RNA, med hjälp av kromatinmodifierande enzymer, bestämmer aktiv / öppen eller sluten kromatinstatus för att främja eller tysta gentranskription. Därför använde vi i den andra studien (**artikel II**) kromatinbaserade metoder för att föreslå en möjlig mekanism genom vilken de aktiva kromatin-associerade lncRNA kan fungera. Vi visar att aktiva kromatin-associerade lncRNA reglerar sina partnergener i cis genom att rekrytera WDR5-kromatinmodifieraren för att skapa en öppen kromatinstruktur hos partnerproteinkodande genpromotorer.

I vår tredje studie (**artikel III**) integrerade vi både transkriptom- och kromatinbaserade metoder för att hitta tidiga utvecklingsassocierade lncRNA. Här fokuserade vi på att spåra molekylära fotavtryck från spermier lncRNA genom stadierna av organismens utveckling. För detta ändamål integrerade vi datamängder från könsceller, preimplantation och post-implantationsstadier i ett embryo. Intressant nog observerade vi distinkta kromatinstrukturer i spermierna. Spermiers lncRNAs var också aktiva under början av zygotisk genomaktivering i preimplantationsstegen och cancer. Sammantaget avslöjar denna studie en unik uppsättning spermiespecifika lncRNA som aktiveras temporärt under preimplantationsstadier och också har avvikande uttryckt i flera cancerformer.

Sammantaget tillhandahåller den aktuella avhandlingen en omfattande katalog med funktionellt relevanta lncRNA som kan delta i cellcykelreglering, cancer, kromatin-modulering och en organisms utveckling. Våra studier kan fungera som en omfattande resurs för framtida undersökningar av lncRNA.

# LIST OF PAPERS INCLUDED IN THIS THESIS

This thesis is based on the following papers, referred to in the text by their roman numerals.

**Paper I.** Ali MM*, Akhade VS*, Kosalai ST*, **Subhash S***, Statello L, Meryet-Figuiere M, Abrahamsson J, Mondal T, Kanduri C. PAN-cancer analysis of S-phase enriched lncRNAs identifies oncogenic drivers and biomarkers. Nature communications. 2018;9(1):1-20. doi: 10.1038/s41467-018-03265-1. PMID: 29491376; PMCID: PMC5830406. (**Co-first authors*)

**Paper II.** **Subhash S***, Mishra K*, Akhade VS, Kanduri M, Mondal T, Kanduri C. H3K4me2 and WDR5 enriched chromatin interacting long non-coding RNAs maintain transcriptionally competent chromatin at divergent transcriptional units. Nucleic acids research. 2018;46(18):9384-400. doi: 10.1093/nar/gky635. PMID: 30010961; PMCID: PMC6182144. (**Co-first authors*)

**Paper III.** **Subhash S**, Kanduri M, Kanduri C. Sperm Originated Chromatin Imprints and LincRNAs in Organismal Development and Cancer. iScience. 2020;23(6):101165. doi: 10.1016/j.isci.2020.101165. PMID: 32485645; PMCID: PMC7262563.

# PAPERS NOT INCLUDED IN THIS THESIS

1. **Subhash S**, Kalmbach N, Wegner F, Petri S, Glomb T, Dittrich-Breiholz O, Huang C, Bali KK, Kunz WS, Samii A, Bertalanffy H, Kanduri C, Kar S. Transcriptome-wide Profiling of Cerebral Cavernous Malformations Patients Reveal Important Long noncoding RNA molecular signatures. Scientific Reports. 2019;9(1):1-13. doi: 10.1038/s41598-019-54845-0. PMID: 31796831; PMCID: PMC6890746.
2. Sukonina V, Ma H, Zhang W, Bartesaghi S, **Subhash S**, Heglind M, Foyn H, Betz MJ, Nilsson D, Lidell ME, Naumann J, Haufs-Brusberg S, Palmgren H, Mondal T, Beg M, Jedrychowski MP, Taskén K, Pfeifer A, Peng XR, Kanduri C, Enerbäck S. FOXK1 and FOXK2 regulate aerobic glycolysis. Nature. 2019. doi: 10.1038/s41586-019-0900-5. PMID: 30700909.
3. Brown P, **RELISH consortium***, Zhou Y. Large expert-curated database for benchmarking document similarity detection in biomedical literature search. Database: The Journal of Biological Databases and Curation. 2019;2019(2019). doi: 10.1093/database/baz085. (**Member of RELISH consortium*)
4. Mondal T, Juvvuna PK, Kirkeby A, Mitra S, Kosalai ST, Traxler L, Hertwig F, Wernig-Zorc S, Miranda C, Deland L, Volland R, Bartenhagen C, Bartsch D, Bandaru S, Engesser A, **Subhash S**, Martinsson T, Carén H, Akyürek LM, Kurian L, Kanduri M, Huarte M, Kogner P, Fischer M, Kanduri C. Sense-antisense lncRNA pair encoded by locus 6p22. 3 determines neuroblastoma susceptibility via the USP36-CHD7-SOX9 regulatory axis. Cancer Cell. 2018;33(3):417-34. e7. doi: 10.1016/j.ccell.2018.01.020. PMID: 29533783.
5. **Subhash S,** Kanduri C. GeneSCF: a real-time based functional enrichment tool with support for multiple organisms. BMC bioinformatics. 2016;17(1):365. doi: 10.1186/s12859-016-1250-z. PMID: 27618934; PMCID: PMC5020511.
6. **Subhash S**, Andersson PO, Kosalai ST, Kanduri C, Kanduri M. Global DNA methylation profiling reveals new insights

into epigenetically deregulated protein coding and long noncoding RNAs in CLL. Clinical epigenetics. 2016;8(1):106. doi: 10.1186/s13148-016-0274-6. PMID: 27777635; PMCID: PMC5062931.

7. Mondal T, **Subhash S**, Vaid R, Enroth S, Uday S, Reinius B, Mitra S, Mohammed A, James AR, Hoberg E, Moustakas A, Gyllensten U, Jones SJM, Gustafsson CM, Sims AH, Westerlund F, Gorab E, Kanduri C. MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA–DNA triplex structures. Nature communications. 2015;6:7743. doi: 10.1038/ncomms8743. PMID: 31754097; PMCID: PMC6872786.

8. Pandey GK, Mitra S, **Subhash S**, Hertwig F, Kanduri M, Mishra K, Fransson S, Ganeshram A, Mondal T, Bandaru S, Östensson M, Akyürek LM, Abrahamsson J, Pfeifer S, Larsson E, Shi L, Peng Z, Fischer M, Martinsson T, Hedborg F, Kogner P, Kanduri C. The Risk-Associated Long Noncoding RNA NBAT-1 Controls Neuroblastoma Progression by Regulating Cell Proliferation and Neuronal Differentiation. Cancer Cell. 2014;26(5):722–37. doi: 10.1016/j.ccell.2014.09.014. PMID: 25517750.

9. Meryet-Figuiere M, Alaei-Mahabadi B, Ali MM, Mitra S, **Subhash S**, Pandey GK, Larsson E, Kanduri C. Temporal Separation of Replication and Transcription during S Phase Progression. Cell Cycle. 2014;13(20):3241-8. doi: 10.4161/15384101.2014.953876. PMID: 25485504; PMCID: PMC4615114.

# CONTENTS

# ABBREVIATIONS

| | |
|---|---|
| CpG | Cytosine-phosphate-Guanine |
| HGNC | HUGO Gene Nomenclature Committee |
| HUGO | Human Genome Organization |
| FDR | False Discovery Rate |
| PCR | Polymerase Chain Reaction |
| qRT-PCR | Quantitative Reverse Transcription PCR |
| CNV | Copy Number Variation |
| SNV | Single Nucleotide Variation |
| GFP | Green Fluorescent Protein |
| MLL | Mixed Lineage Leukemia |
| WDR5 | WD repeat-containing protein 5 |
| CTCF | CCCTC-binding factor |

# PREFACE

The main purpose of this thesis is to dissect the functions of long noncoding RNAs (lncRNAs) using various computational approaches. I used pragmatic approach by combining both quantitative and qualitative methods to achieve my research goal. This uses data, theoretical and methodological triangulation where the focus is to analyse data generated from different experiments, interpreting the results for formulated hypotheses and, using multiple and complementary approaches to solve the problems. The core objective of the thesis is developed by a notion that protein-coding genes are well-studied classes to address problems in molecular biology. The major portion of the human genome comprises of noncoding elements that do not code for any proteins, nevertheless they have a biological significance. Among these elements, long noncoding RNAs (lncRNAs), have been implicated in various developmental and disease functions. However, the mechanisms by which these RNAs function are still lacking. Therefore, it is important to explore possible mechanisms by which these sub-classes of RNAs function in various biological context. Unlike protein-coding genes, we had complete freedom to explore long noncoding RNAs from different perspectives without keeping any previous notion or assumptions. From previous studies we only know that these RNAs function directly in the form of transcribed RNAs without translating into protein. Also, we known that cell cycle regulation with defective cell division is an important characteristic feature in many cancers. In cancerous cells, the cell divides rapidly in an uncontrollable fashion. It is important to study molecules other than proteins that may have a role in this cell cycle process. We focused on finding lncRNA molecules that are involved in the cell cycle process which may have role in different cancers. Moreover, our lab showed one of the developmentally regulated long noncoding RNA molecule *KCNQ1OT1*, to be interacting with chromatin and establishing its functions in a lineage-dependent manner. To characterize different chromatin-associated long noncoding RNAs, this thesis has considered analysing data from various high-throughput approaches. This includes exploring the mechanisms of chromatin associated long noncoding RNAs from different chromatin sub-compartments. While focusing on cancer cells and chromatin, we came across the fact that some of the cellular and molecular features of cancer cells resemble cells from early embryos. This similarity between cells from early developing embryos and cancer cells fascinated us to trace the behaviour of lncRNA molecules in gametes, early embryos, matured healthy tissues and tissues from cancer patients. To this end, we found interesting lncRNA molecules in different biological aspects having diverse role.

# 1 INTRODUCTION

## 1.1 Cell – Structure and function

A cell contains two major structural components, such as nucleus and cytoplasm. The outer surface of the cell is called the cell membrane that encloses the cytoplasm and the nucleus. The cytoplasm consists of the Golgi apparatus, endoplasmic reticulum, centrioles, ribosomes, lysosomes, and mitochondrion. A nuclear membrane encloses the components of nucleus in the cell. These nuclear components consist of nucleolus and chromatin (**Figure 1**).

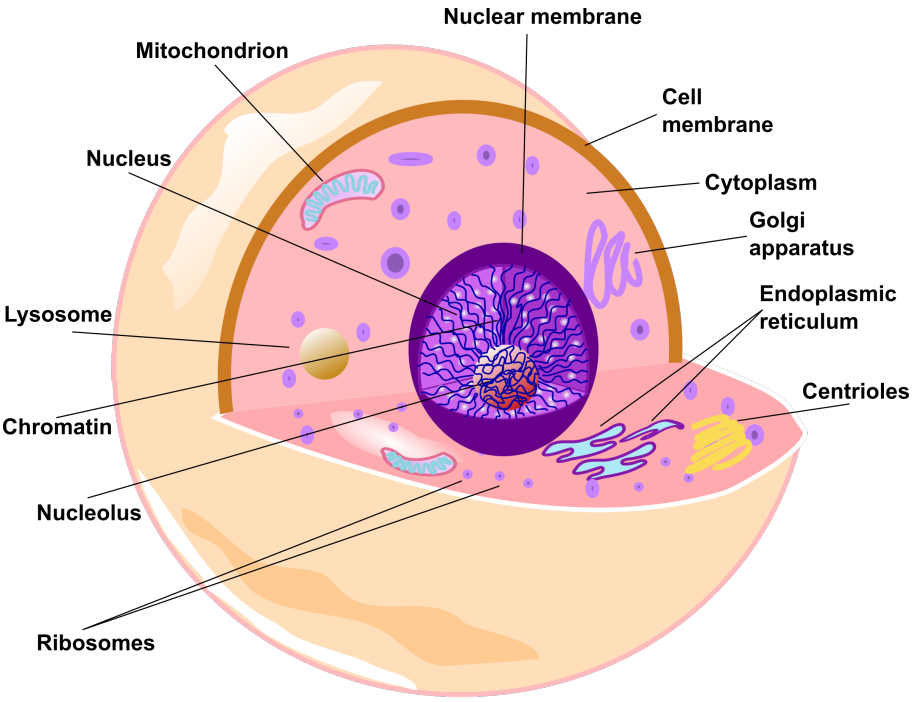


***Figure 1.*** *Structure of a cell and its cellular components.*

### 1.1.1 Cellular components and cell division

Every complex multicellular organism, including humans, started with single cell (zygote) (**Figure 1**). Cell, with the help of cellular components, starts dividing into multiple clusters of cells to form an organism. This process of cell division is called mitosis, where one cell divides into two by assembling the spindle fibers at two polls and pulls the chromosomes of the cell apart (**Figure 2a**). Major components of a single cell comprise of cytoplasm and nucleus containing DNA through which most molecular machinery functions to keep up the cellular integrity. The key cellular processes are maintenance of cell homeostasis, response to an external stimulus, cell-to-cell communication or cell signaling, and maintenance of genomic integrity. It is important for a cell to perform these vital functions without any defect to keep an organism in a healthy state. A minor hindrance to one of these processes will lead to disease or even can cause death. A simple example is, uncontrollable cell division leading to proliferative and invasive cells, that can form cancerous tissue. Further stages of these cells can activate factors responsible for the formation of new blood vessels (angiogenesis) to supply oxygen and nutrients necessary for cancerous cells or tissues[1]. These cancerous cells also have the ability to spread to other tissues or organs and it is called metastasis. These abnormal behaviors of cancerous cells are called as hallmarks of cancers[2,3]. Therefore, currently investigations are going on to address the exact molecular mechanism through which cell division occurs.

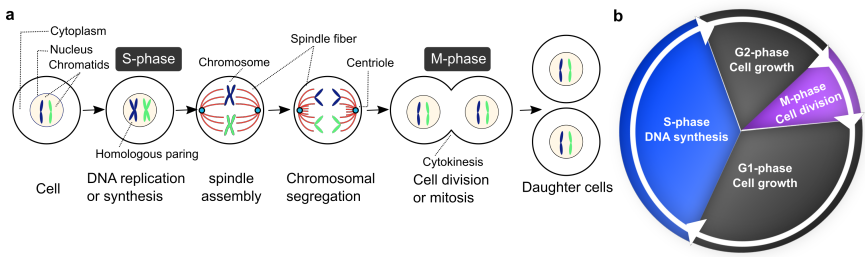


***Figure 2.*** *Different stages of cell division and cell cycle.* ***a)*** *Schematic of somatic cell division process.* ***b)*** *Representation of cell cycle phases such as G1 (cell growth phase), S-phase (synthesis or DNA replication phase), G2 (second phase of cell growth) and Mitosis phase (cell division).*

### 1.1.2 Cell cycle

Before a cell undergoes mitosis, it has to make sure if the cell is ready for cell division, and this occurs in interphase. Cell goes through different stages of interphase before entering into mitosis, such as G1, S, and G2 phase (**Figure 2b**). During this interphase, the cell grows bigger in size to make one more copy of DNA or chromosome and centrioles. Later this cell with a duplicated copy of cell components divides into two. This whole process of the cell cycle is vastly controlled by phosphorylation and dephosphorylation of cyclin-dependent kinase (CDK) protein complexes that determine a cell's entry into the next stages of the cell cycle. One more important factor, such as cell cycle checkpoints, gets activated when there is a problem during DNA replication or chromosome segregation. These checkpoints delay the progress of cells into the next stage until the damage is repaired[4,5]. However, sometimes these checkpoints are not completely foolproof and are prone to errors.

In the G1-phase, the newly formed cell undergoes growth, and the components of the cell are duplicated to prepare it for the replication stage. Next, the cell enters into S-phase to makes a copy of DNA without changing their chromosomal numbers or ploidy (decondensed sister chromatids, see **section 1.2.1** for information about chromatids). Before entering into mitosis, the cell grows again by entering into G2-phase by synthesizing necessary proteins and enzymes important during cell division. Once the cell passes the G2 checkpoint, the cell division or mitosis starts by condensing the chromatid into compact chromosomes (**section 1.2.1**). At the same time, the ribosomal machinery disappears, the nucleus disintegrates and, spindle fibers form around the chromosomes. These spindle fibers pull individual chromosomal copy apart with the help of centrioles (as seen in **Figure 2a**). Now spindle fibers disappear, forming a nuclear envelope, stretches chromosomes back into chromatids, and pair of nucleus re-appears for each duplicated copy. Each nucleus forms individual daughter cells by the process of cytokinesis. These individual daughter cells again enter into G1-phase for the maintenance of the cell cycle regulation. Defective cell cycle regulation can cause cell to divide exponentially. One of the major cancer cell hallmarks is the cell cycle, where the cells grow and divide uncontrollably. Therefore, it is important to investigate their detailed molecular mechanisms that can lead to cancerous cells[5]. Studies have been conducted to understand normal and chronic cell proliferation during cell division. Importantly, the response from cell cycle regulators such as CDKs and growth factors during cell cycle checkpoints are

shown to play a significant role in promoting cell proliferation. In some cancers, CDK inhibitors are also used as treatment strategy[6]. Moreover, studies have also shown that apart from protein-coding genes, long noncoding RNAs (transcribed from the noncoding region of the genome, **section 1.3** for details) can regulate cell cycle progression by controlling CDKs and CDK inhibitors[7-10].

**Box 1: Difference between mitotic and meiotic cell division**

In humans, mitotic cell division occurs in somatic cells where a diploid cell (2n or 23 pairs of chromosomes) gives rise to two more diploid daughter cells (2 x 2n or 2 x 23 pairs of chromosomes) with the help of DNA replication and chromosomal duplication. In contrast, meiosis occurs in germ cells where a diploid cell (2n or 23 pairs of chromosomes) results in two cells with haploid chromosomes (1n + 1n, one pair from each parent). The resulting haploid cells are called gametes, egg cells in female and sperm cells in the male. This is the reason when egg and sperm fuses (fertilization), zygote gets one set of the chromosome from the female, and another set from the male forming a diploid zygote cell.

### 1.1.3 Cell division and cancer

The cell cycle is the core process in eukaryotic organisms that helps in the growth, survival, and renewal of cells in eukaryotes. As discussed earlier, there are various stages involved in the process of the mammalian cell cycle. Among these, S-phase and M-phase stages play an important role in successful cell division. In somatic cells, when one of these processes fails, it may lead to defective cell division or, in most cases, uncontrollable cell division[5]. One such a common error includes a mutation in cell cycle regulating protein kinases (CDKs), which in turn fails to respond to DNA damage that occur during DNA replication[11]. In this situation, a cell either undergoes apoptosis (cell death) or starts inheriting the mutations in subsequent daughter cells (insertion, deletion, or nucleotide mismatch). This accumulation of mutation load leads to genomic instability causing cancer[12]. Some of these mutations can lead to a benign tumor where the cells are not cancerous and do not spread to other tissues or organs. It is very rare that benign tumors turning into cancerous form[13,14]. Although it is less probable that these benign tumors cause problems unless it surpasses nearby blood vessels, nerves, or tissues are causing distress and

develop other complications. In the case of malignant tumor cells, it causes adverse effects by spreading proliferated cells to other tissues or organs (metastasize) interfering with their core cellular processes[15]. There are various environmental factors that may cause DNA damage or replication errors, such as radiation exposure, viruses, and chemo-induced (**Figure 3**). Additional factors that can influence cancer are genetic (transgenerational inheritance), aging, hormonal imbalances, lifestyle, and diet. Though there are many factors responsible for forming malignant cells, the resulting outcome is always aberrant cell division and growth. Hence cell cycle, specifically DNA replication during S-phase, has a strong association with cancerous cells due to replication errors[16]. Thus, detailed molecular mechanisms and pathways involved in S-phase, controlling cell proliferation are needed to be explored for treatment strategies in cancer[17].

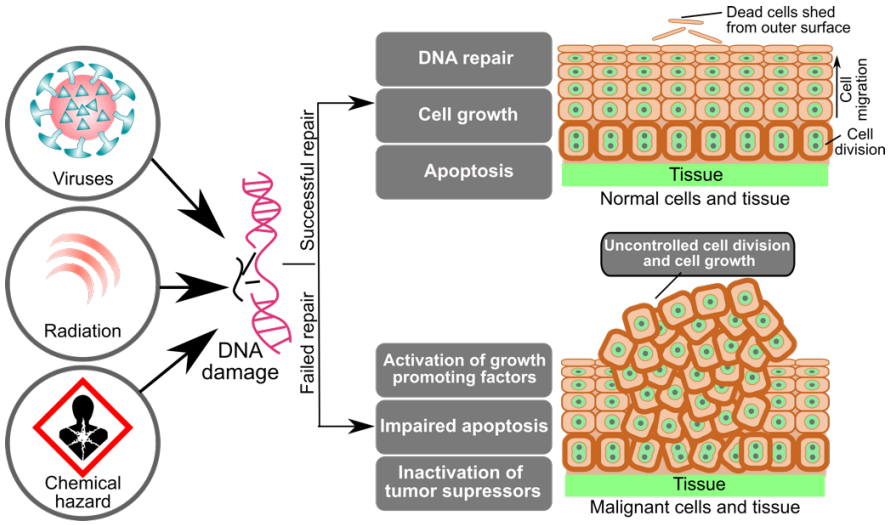


***Figure 3.*** *DNA damage causing external stimuli can be viruses, radiation exposure and exposure to hazardous chemicals. Once the DNA is damaged, it can be either successfully repaired or fail. Successful repair mechanism involves DNA repair, cell growth and apoptosis leading to healthy normal cells. Failure can make the cell growth uncontrollably by activating growth-promoting factors, inhibiting apoptosis and tumor suppressors.*

### 1.1.4 Cell division during embryo development

After meiosis, the resulting haploid sperm cells from the male will have either X or Y sex chromosomes. On the other hand, the haploid egg cells from female (XX) will only have all cells with X chromosomes. During fertilization, when the sperm and egg cell fuses, it forms a complete diploid cell called zygote where sperm determines the sex of the fetus. Zygote enters into a cell cycle as the somatic cell does, and subsequent cell divisions of this zygote result in stem cells, progenitor cells, embryo, tissues, and organs. Unlike the cell cycle in adult somatic cells, the zygotic cell cycle and cell division result in different cell types at different stages of the development process. Firstly, during preimplantation stages of development, the zygote forms two daughter cells and then four-cells, eight-cells and morula[18,19](**Figure 4**). These cells are known as totipotent cells. After morula, the cells undergo rapid cell division forming a cell mass called blastocyst containing pluripotent cells, which later gives rise to embryonic stem cells. These embryonic stem cells form three distinct embryonic germ layers, such as ectoderm, mesoderm, and endoderm (**Figure 4**). Germ layers contain multipotent cells and have the ability to form multiple cell types called progenitor cells. Lastly, these individual progenitor cells undergo lineage commitment and differentiate into various organs and organ systems[20](**Figure 4**). This shows that there are some similarities between the adult somatic cell cycle and the zygotic cell cycle, were one diploid cell divides and forms two diploid daughter cells. Though there are similarities in the outcome, the early embryonic cell cycle only switches between S-phase (DNA synthesis) and M-phase (mitosis) without entering the G1 or G2 phase[21]. Since these cells also undergo DNA replication and cell division, there is a probability for DNA damage and replication errors[21,22]. Therefore, it is essential to study the molecular mechanism involved in embryonic cells during development that mimic the aberrant cell division in cancerous cells[23].

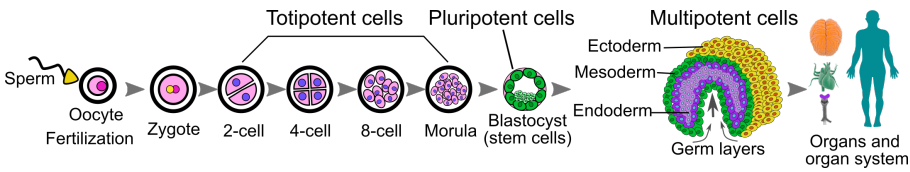


***Figure 4.*** *Schematic shows sperm-oocyte fertilization, preimplantation (zygote to blastocyst) and post-implantation stages of an embryo. These stages are followed by organogenesis to form different organs and systems.*

## 1.2 Overview of chromatin and genome

Genome contains a stretch of nucleotides called DNA. This supercoiled DNA is wrapped around proteins called histones. These continuous chains of histone and DNA form nucleosome and higher-order chromatin structure (**Figure 5**). Chromatin structures together are tightly arranged in the form of chromatids. A pair of sister chromatids are known as chromosomes (**Figure 5-6**). All these complex structures contribute to the activation or inactivation of gene transcription (**Figure 5**).

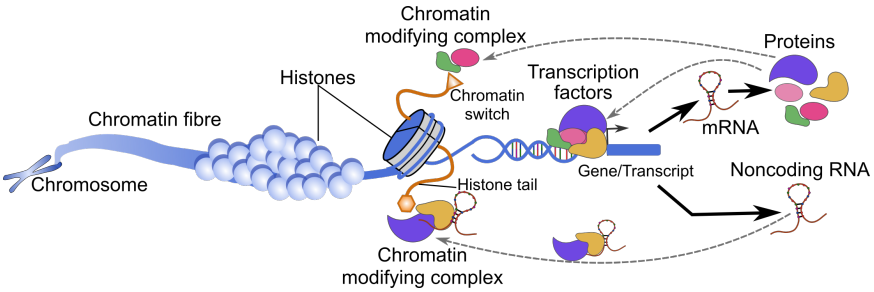


***Figure 5.*** *Schematic diagram with a complete picture of chromatin and genome. This shows the detailed structure of chromosome, chromatin, and histones wrapped by DNA. Histone tails are prone to post-translational modification by chromatin-modifying complexes. These histones modifications act as a switch to activate or repress gene transcription. Transcription of genes involves the binding of DNA unwinding proteins, various transcription factors, proteins and noncoding regulators. After transcription, a gene in the genome gives rise to either protein-coding or noncoding mRNAs. The protein-coding genes are transported from nucleus to cytoplasm for translating mRNA into proteins.*

### 1.2.1 Chromatid, chromatin, and chromosome

The cell consists of a nucleus and cytoplasm, where key cellular machinery functions to maintain cell homeostasis[24]. The nucleus contains genetic material in the form of nucleosome, chromatin, and chromosome structures (**Figure 5**). Moreover, nucleus carries out functions such as DNA replication during the cell cycle and maintenance of gene expression or transcription profiles. Chromatin in the nucleus is made of DNA wrapped around an octamer protein complex, two copies of four canonical histone molecules (H2A, H2B, H3, and H4). Individual histones are wrapped around with 1.67 turns of DNA containing 146 bases, forming a structure called a nucleosome. Multiple nucleosomes bind together like a chain and form a higher-order chromatin structure (**Figure 6**). During cell division, these chromatids get condensed to form a compact structure called chromosomes. Each chromosome makes a copy of its genetic material for its daughter cells, and these identical pairs of chromosomes are called chromatids (sister chromatids). These sister chromatids are held together by a structure called centromere.

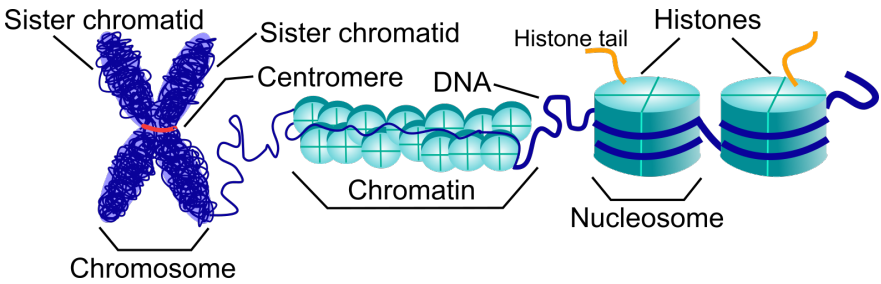


***Figure 6.*** *Structure of chromosome, chromatin and nucleosomes. Chromosome is made of condensed pair of sister chromatids. These chromatids contain densely packed chromatin (set of nucleosomes). Nucleosomes are histones wrapped by DNA.*

### 1.2.2 Gene transcription and translation

Genome contains deoxyribonucleic acid (DNA) which is a double-helix structure made of a stretch of four nucleotides arranged in a complementary manner. These nucleotides are complementarily held together with the help of strong hydrogen bonds where nucleotide adenine (A) pairs with thymine (T) and nucleotide guanine (G) pairs with cytosine (C). This base-pairing rule is known as Chargaff's rule. Some portions of this stretch of nucleotide sequences are functional and are called genes. Genes in the genome are frequently read from 5' to 3' direction, the order by which the gene takes its own copy to become a functional entity. Major components of a gene are transcription start site (TSS), exons, introns, and transcription end site. Exons code for functional RNA or protein, whereas introns are noncoding elements between two exonic regions. A gene becomes functional with the help of transcriptional and translational machinery. In the nucleus, the polymerase enzyme recognizes and binds to DNA sequence of the gene promoter to begin the transcription. Transcription is the process where the gene makes its own complimentary copy called pre-messenger RNA (pre-mRNA) (**Figure 7**). It is called ribonucleic acid because while copying, the thymine (T) gets demethylated and forms uracil (U). This happens to protect the actual DNA from breaking down by different enzymes. Also, the RNA molecules are less stable than the DNA. During transcription, the transcriptional machinery recognizes the stop signal at the transcription end site of the gene to complete the pre-mRNA synthesis. Next, the pre-mRNA undergoes polyadenylation (addition of poly(A) tail at 3' UTR), 5' capping, and intron splicing to become matured and stable mRNA. This stable mRNA is then transported out of the nucleus to the cytoplasm, where it is delivered to the ribosomal machinery. Ribosomes read this mRNA genetic code in triplets (codon) and translate into a stretch of amino acid sequences or polypeptides to form a functional protein. Translation begins by recognizing the start codon (AUG) and terminates with the stop codons (UGA/UAG/UAA) signal (**Figure 7**).

In the human genome, there are approximately 20,000 protein-coding genes covering 1.5 to 2% of the nucleotides in the whole genome. The remaining portion of the genome is considered as the noncoding genome, and some of these active regions give rise to functional noncoding mRNAs. These mRNAs from the noncoding genome do not undergo translation and does not code for any protein and are known as noncoding RNAs. These noncoding RNAs undergo transcription like protein-coding genes but do not translate. Instead,

they stay as mRNAs and perform various cellular functions in the form of mRNA secondary structures. These noncoding RNAs are classified into different categories based on their length, function, and localization (see **section 1.3**).

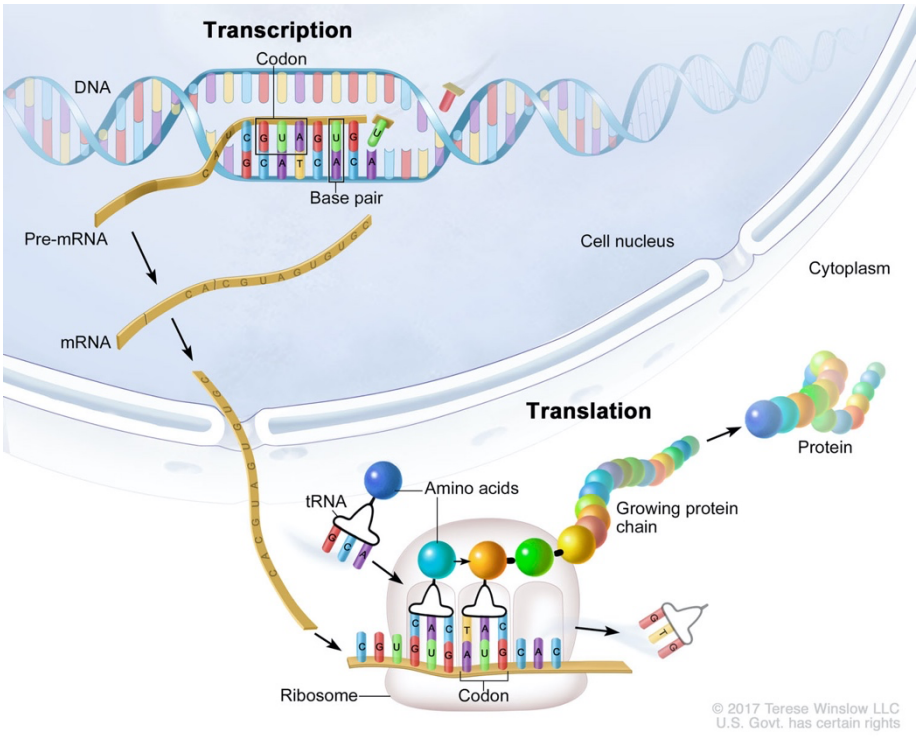


***Figure 7.*** *Mechanism of gene transcription and translation. National Cancer Institute © 2017 Terese Winslow LLC, U.S. Govt. has certain rights.*

### 1.2.3 Epigenetics and transcription regulation

***Histone modifications***: Out of four (H2A, H2B, H3, and H4) histone proteins in the complex, two of the proteins (H3 and H4) have tails and are more prone to post-translational modifications. These H3 and H4 modifications are very stable and can be maintained in the chromatin for generations and it is known as epigenetic inheritance. The common types of histone modifications are

methylation, acetylation, phosphorylation, ubiquitylation, and SUMOylation. Histone H3 is predominantly acetylated at positions lysine 9, 14, 18, 23, and 56, methylated at arginine 2 and lysine 4, 9, 27, 36, and 79, and phosphorylated at serine 10, serine 28, threonine 3, and threonine 11. Similarly, histone H4 is acetylated at lysine 5, 8, 12, and 16, methylated at arginine 3 and lysine 20, and phosphorylated at serine 1. In addition to core histones, there is a linker histone H1, which is not part of nucleosome. Histone H1 stays on top of the nucleosome (core histones and DNA) to keep the wrapped DNA in place. The post-translational modifications of histones play a major role in defining the chromatin structure and, in turn, affect the gene transcription (**Figure 8**). Certain modifications increase the chromatin accessibility (open or euchromatin) facilitating the transcription factors binding and promoting transcription[25-27]. Similarly, some modifications keep the chromatin very tight (close or heterochromatin) inaccessible to transcription factors[28,29].

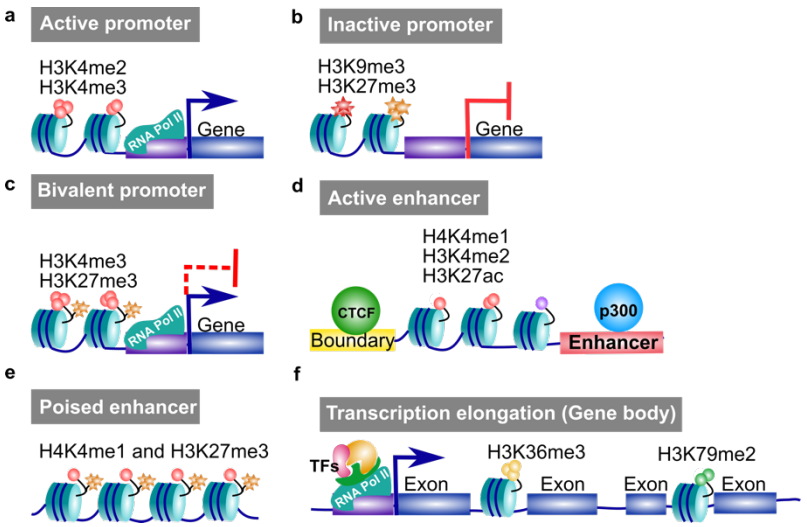


***Figure 8.*** *Different types of chromatin modifications.* ***a)*** *Active promoter with H3K4me2 and H3K4me3 modifications.* ***b)*** *Inactive promoters with H3K9me3 and H3K27me3 modifications.* ***c)*** *Bivalent promoters with H3K4me3 and H3K27me3 modifications.* ***d)*** *Active enhancer marked with p300 HAT and with H3K4me1, H3K4me2 and H3K27ac modifications.* ***e)*** *Poised enhancer with H3K4me1 and H3K27me3 modifications.* ***f)*** *Gene body with transcription elongation marks, H3K36me3 and H3K79me2. The blue arrow represents active transcription, the red arrow represents inactive transcription, and the gene with both arrows represents temporal behavior.*

**Chromatin and transcriptome-based integrative approaches to profile functional lncRNAs**

The active gene promoters are usually marked with H3K4me2/3 modifications where the chromatin is accessible (open) to transcription factors and other DNA binding proteins (**Figure 8a**). Whereas, inactive promoters are enriched with repressive chromatin marks such as H3K9me3 or H3K27me3 (**Figure 8b**). Additionally, studies have shown that the bivalents promoters where the histone has both H3K4me3 and H3K27me3 modification plays an important role in modulating development associated genes in an stage-specific manner[30-32](**Figure 8c**). Similarly, active enhancers are marked with H3K4me1/2 and H3K27ac and the enhancer boundary element is enriched with CTCF binding (**Figure 8d**). Enhancer regions can be predicted by p300 (histone acetyltransferase, HAT) binding locations[33]. Poised enhancers are marked with H3K4me1 and H3K27me3 (**Figure 8e**). Moreover, H3K36me3 and H3K79me2 chromatin modification are shown to be involved in transcription elongation process during mRNA synthesis (**Figure 8f**). The H3K36me3 modification has high affinity towards 5' end whereas H3K79me2 modification towards 3' end of the gene body[28,34,35].

***Chromatin modifiers***: Chromatin or histone modifications are established with the help of chromatin-modifying enzymes. Usually, these modifiers or histone-modifying complexes are part of co-regulatory proteins (**Figure 9**). These chromatin modifiers cannot recognize DNA targets, and it need mediators such as DNA methylation, transcription factors, or other noncoding RNAs to recruit them into the chromatin or DNA (**Figure 9**). It is known that the EZH2 protein subunit from PRC2 (Polycomb Repressive Complex 2) chromatin-modifying complex has methyltransferase property and tri-methylates Histone 3 (H3) lysine 27 (K27) to establish chromatin silencing mark (H3K27me3)[36]. Following this, the chromo-box homolog protein subunit (CBX) from the PRC1 complex binds to DNA by recognizing H3K27me3 histone modification and stabilizes the silencing[37]. These mechanisms can lead to long term epigenetic silencing of chromatin (repressive or closed chromatin) important for stem cells and embryo development. MLL forms a complex by interacting with many proteins. WDR5 is the catalytic subunit of MLL complex involved in the tri-methylation of H3K4 (H3K4me3). This H3K4me3 modification makes the chromatin accessible to many transcription factors and proteins to promote transcription of associated target genes. Studies have shown that the WDR5 has RNA binding capacity and is known to interact with many long noncoding RNAs (lncRNAs). Some of these lncRNAs recruit MLL complex via WDR5 interaction to the chromatin and modulates the expression of target genes to maintain embryonic pluripotency. This MLL-WDR5 complex maintains open or active chromatin, which in turn supports active gene transcription[38].

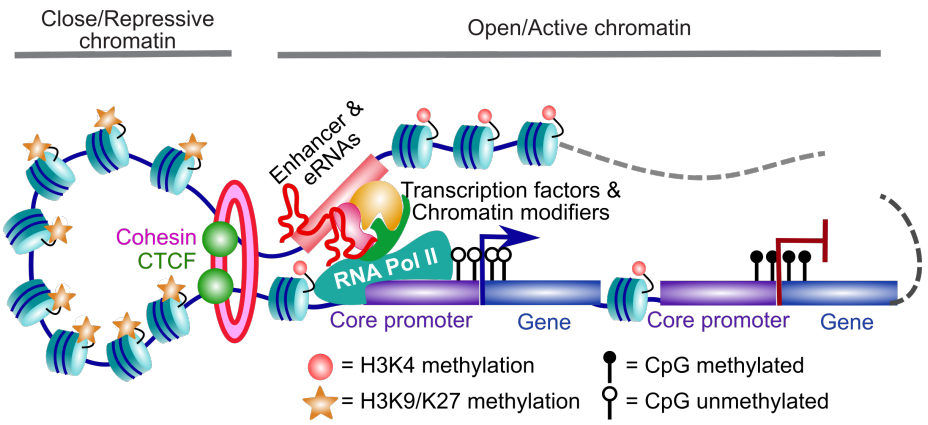


***Figure 9.*** *Schematic shows enhancer/eRNAs, transcription factor, chromatin-modifying enzymes, histone methylation and DNA methylation-based gene transcription regulation.*

***DNA methylation***: This is a reversible covalent modification by methylation that occurs in 5' position of a cytosine nucleotide on the gene promoter containing CG islands (cytosine-guanine dinucleotide, CpG promoters). Usually, this 5'-methylcytosine (5mC) modification at the CpG promoters are associated with gene repression by making the DNA unrecognizable for transcription promoting factors or proteins during development. Moreover, this epigenetic modification also leads to the transgenerational inheritance of this repression state. However, there is an exception when the CpG methylation occurs on the gene body. In many cancers, this gene body methylations are shown to be associated with active transcription of a gene[39]. Therefore, it is evident that CpG methylation in the promoter and gene body has opposite effects on gene transcriptional status. In addition to development, the DNA methylation and demethylation is also frequently seen in various cancer to control the expression of oncogenes or tumor suppressors (**Figure 9**).

***RNA polymerase***: During the process of transcription, the RNA polymerase enzyme recognizes the gene promoters in the DNA and starts making a copy of RNA (messenger RNA). RNA polymerase II (RNA pol II) occupancy is a key factor of an active gene (**Figure 9**). This whole process involves three

important stages, such as pre-initiation or promoter binding, initiation by assembling transcription factors, and elongation by mRNA synthesis[40,41].

***Transcription factors***: Promoters sequences are found near to the upstream region of transcription start sites (TSS) of a gene and are targets of transcription initiation factors (**Figure 9**). The transcription factors (TFs) are the proteins that contain DNA binding domains (DBDs) to recognize the DNA sequence of the gene promoters. Some TFs do not directly bind to gene promoters; rather, they form a multimeric complex with other proteins that recognizes RNA polymerase. Individual gene promoters can have multiple transcription factors (TFs) recognizing sequences or motifs depending on their role in multiple cellular processes or tissue-specific regulation. Once the TFs bind to the gene promoter, which either promotes the transcription (activator) or blocks the RNA polymerase recruitment to repress the transcription (repressor). A well-known example of transcription factors are clusters of Hox genes encoding transcription factors. During the development of an embryo, these Hox transcription factors decide which part of the embryo to be developed into the head, abdomen, limb, thorax, or other organs. These Hox TFs regulates this process of organogenesis via temporal activation and repression of target gene expression[42-44].

***Enhancers***: Unlike gene promoters, enhancers are found far from the gene that can amplify or promote the transcription rate of a particular gene. Chromatin conformation and DNA looping play an important role in bringing distant enhancer-promoter interactions. Enhancer brings transcription factors, chromatin modifiers, and mediator complexes to the RNA polymerase II bound gene promoter to enhance transcription. Cohesin and CTCF mediated chromatin looping can mediate this distant enhancer and gene promoter contacts[45]. In such cases, enhancer-promoter interaction falls in a close proximity to CTCF boundaries (**Figure 9**). Some of the enhancer regions transcribe and produce enhancer RNAs (eRNAs) which in turn interact with cohesin protein to maintain the stability of chromosomal looping[46].

### 1.2.4 Epigenetic changes during embryo development

Epigenetic-based transgenerational inheritance is a natural phenomenon where epigenetic changes (no changes in actual nucleotide sequences or bases) are transmitted to offspring over generations through germline[47,48]. These epigenetic changes majorly involve histone modifications (H3K4 and H3K27

methylation) and DNA methylation (5-methylcytosine)[49]. The chromatin and methylation domains are established to activate or repress the gene expression in a stage-specific manner. As we know that the genes with promoter methylation are kept silent and the unmethylated genes are active. Similarly, H3K4me2/3 is an active promoter mark that supports active transcription. On the other hand, H3K9/H3K27me3 is a repressive mark deposited over inactive gene promoters[28,35]. These are the three (DNA methylation, H3K4me3/2 and H3K9/H3K27me3) essential factors, that play a key role in determining the transcription of stage-specifically modulated genes during development (**Figure 10**). Germline (gametes) carries certain regions with epigenetic changes and needed to be maintained for producing successful offspring. This process of carrying epigenetic memory over generations is termed as genomic imprinting. Both the maternal (egg) and paternal (sperm) gametes are shown to involve in genomic imprinting[49-55]. DNA methylation domains established in gametes are known to regulate transcription of genes in offspring. These gametic-derived DNA methylation domains are rapidly lost upon fertilization and regained in later stages of development. Additionally, studies have shown that maternal inherited broad active chromatin domains (H3K4me3 marked) are important for maintaining the gene expression during zygotic genome activation (ZGA) after fertilization[56]. Moreover, gene expression during embryo development is also regulated by parentally imprinted H3K27me3 domains in an DNA methylation independent manner. As we discussed earlier in the **section 1.1.3**, after fertilization, zygotic cell regains the property of totipotency to further develop into an embryo. During early mammalian development, the chromatin undergoes temporal changes to maintain cell lineage identity. The chromatin and methylation factors determine the open and repressed chromatin states which in turn, defines the transition of pluri-/multi-potent cells into differentiated/adult cells[57]. In the zygote, the methylation domains are first erased to activate transcripts essential for regaining the totipotency of the cells. It has been shown that this totipotency is also achieved due to paternally (male germline) inherited bivalent chromatin domains consisting of both active (H3K4) and repressive (H3K27) chromatin marks[30-32]. Once the zygote reaches the morula stage, the DNA methylation is re-established to silent the totipotent genes and this is followed by pluripotent genes being unmethylated. These pluripotent genes are activated for a proper transition of morula into a cell mass called a blastocyst. Later these pluripotent genes are permanently repressed by the establishment of DNA methylation after differentiating into embryonic stem cells (ESCs). Once the ESCs are formed, the developmental genes are expressed by establishing an active H3K4 methylation mark. This temporal establishment of chromatin and methylation with correlated stage-specific expression patterns are followed throughout the development process[58,59].

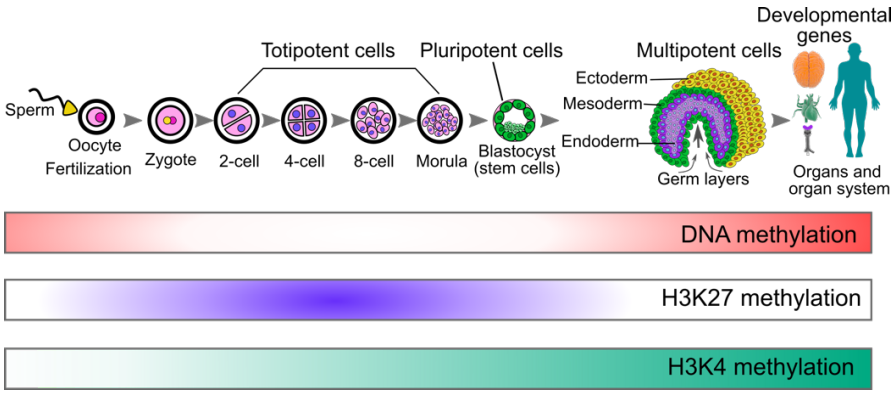


***Figure 10.*** *Dynamics of epigenetic modifications (DNA, H3K27 and H3K4 methylation) during mammalian development.*

### 1.2.5 Epigenetic changes in cancer

DNA methylation in cancer is one of the key factors modulating the expression of oncogenes and tumor suppressors[60,61]. This involves methylation of 5-methylcytosine at CpG dinucleotide promoters by DNA methyltransferases[62]. Methylation of CpG promoters hinders the binding of transcription factors thereby silencing the gene expression[63]. Additionally, these methylated regions can be recognized by certain methyl-binding proteins ("readers") to activate or repress the gene transcription[64,65]. In clear cell renal cell carcinoma, *GREM1* promoter methylation (hypermethylation) is shown to have a poor prognosis in patients due to enhanced proliferation and angiogenesis[66]. Genome-wide profiling of methylation patterns in CLL (Chronic lymphocytic leukemia) patients revealed two of the lncRNAs with promoter methylation stratifies the survival outcome for patients. The low expressed lncRNA *CRNDE* (tumor suppressor) was having hypermethylated promoter and this higher methylation levels predicts poor prognosis in CCL patients. In contrast, the highly expressed lncRNA *AC012065.7* (oncogenic) promoter is marked with lower methylation and this hypomethylation predicts poor survival in CLL patients[67]. In contrast to this promoter methylation patterns, recent cancer studies have shown that methylation in gene body can promotes gene transcription[39,67,68]. In addition to promoter methylation, it is important to explore the detailed

mechanisms to improve the methylation-based treatment strategies in cancer[69,70]. The demethylation occurs during embryo development leads to expression of developmental genes, but later disappears in adult somatic cells (**Figure 10**). However, these genes are re-expressed by demethylation in many cancers[23]. These similarity in patterns of methylation and demethylation between early embryonic and cancer cells can help us improve future cancer therapy.

**Chromatin and transcriptome-based integrative approaches to profile functional lncRNAs**

## 1.3 Long noncoding RNAs – An introduction

The transcribed RNAs contain two major classes, such as protein-coding and noncoding RNAs (ncRNAs). These noncoding RNAs (ncRNAs) are categorized into housekeeping and regulatory noncoding RNAs (**Figure 11**). The housekeeping noncoding RNAs are functionally divided as transfer (tRNA), ribosomal RNA (rRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA). However, the regulatory noncoding RNAs are classified based on their length such as short and long noncoding RNAs. The noncoding RNAs shorter than 200 nucleotides are considered to be short ncRNAs. These short ncRNAs are micro RNAs (miRNAs), small interfering RNAs (siRNAs) and piwi-interacting RNAs (piRNAs). Major portion of regulatory ncRNAs contains long noncoding RNAs (lncRNAs) having length greater than 200 nucleotides. LncRNAs can be put in five broad categories, such as sense, antisense, intronic, intergenic, and bidirectional (transcription occurs from protein-coding gene as well as the overlapping promoter of lncRNA from the opposite strand)[71,72]. Since there is no defined classification for long noncoding RNAs (lncRNAs), they are classified differently in a different context. Some lncRNAs regulate their neighboring genes (*cis acting*) and some act as a distant regulator (*Trans* acting). In cancer, they are classified as oncogenes which promotes tumor growth, tumor suppressors, and lncRNAs having dual roles (acting differently in different cancer)[73]. Several studies with the help of experimental and computational approaches classified lncRNAs based on their cellular localization as nuclear and cytoplasmic[74-78]. Additionally, recent studies have classified lncRNAs based on their genomic arrangement or orientation with respect to nearby protein-coding genes such as XH, divergent head-to-head transcripts; XT, tail-to-tail; XI, lncRNA-Inside (lncRNA within a protein-coding gene); XO, lncRNA-Out (protein-coding gene within a lncRNA; sense (same direction as protein-coding genes) separated by a certain distance[79]. LncRNAs show highly dynamic, cell-type and tissue-specific expression[80-82]. Some of these lncRNAs plays an important role during development, cancer and diverse cellular contexts[83-86].

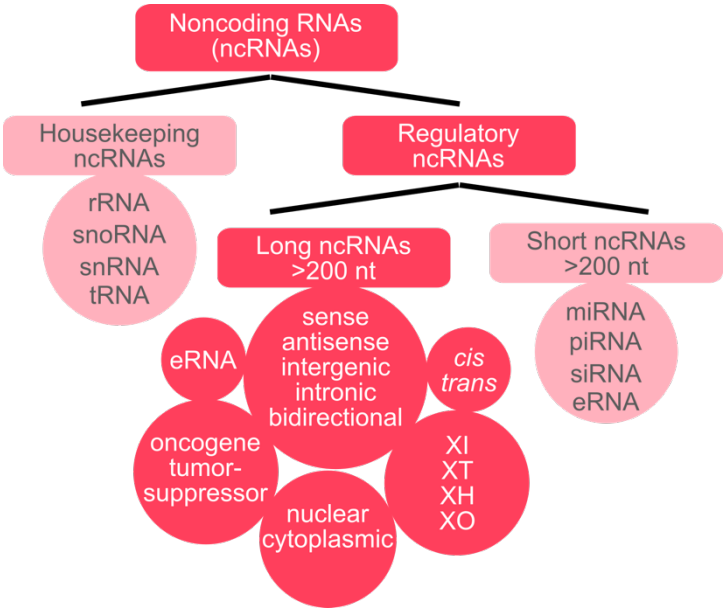


***Figure 11.*** *Classification of noncoding RNAs (ncRNAs) based on their functions, regulatory roles and length.*

### 1.3.1 Properties of lncRNAs

LncRNAs undergo transcription to form pre-mRNA followed by polyadenylation, 5' capping, and intron splicing to become matured mRNA. Unlike mRNAs produced from protein-coding genes, noncoding mRNAs do not undergo translation. These RNAs lack ability to get translated because of absence of open reading frame (ORF) (**Box 2**). However, there are some exceptions where some variants of lncRNAs are shown to code for small peptides or micropeptides[87-89]. Majority of lncRNAs performs their functions in the form of RNA secondary structure. These secondary structures are mostly single-stranded with short multiple stem-loop structures. LncRNAs can fold into complex secondary structures, and these structures can interact with DNA, other RNAs, protein, or protein complexes modulating the function of their targets in *cis* and/or *trans*[82,90-93].

**Box 2: Open reading frame (ORF)**

During translation, the mRNAs are read in triplet codes called codons. These codons begin with start codon (AUG) and ends with stop codons (UGA/UAG/UAA) and this continuous stretch is known as open reading frame (ORF). Individual triplet codon codes for an amino acid and this sequence of amino acids forms peptides, polypeptides and proteins.

### 1.3.2 Evolutionary conservation of lncRNAs

Long noncoding RNAs are evolutionally less conserved compared to strongly conserved protein-coding genes. This shows that the lncRNAs are more species-specific and may have a major role in the adaptive selection of species[94]. However, predicting the conservation of lncRNAs are more complex than the simple nucleotide sequence-based conservation as seen in the protein-coding genes. In lncRNAs, the evolutionary conservation can be sub-divided into four categories such as, primary sequence, secondary structure, synteny or chromosomal localization and functional conservation. Compared to intronic and intergenic regions of the genome, the nucleotide sequence of lncRNAs show moderate conservation but not to the extent of protein-coding genes. Also, comparative studies showed that only a fraction of sequences or short contigs are conserved among species[95]. Moreover, lncRNAs with less or partial sequence conservation are shown to have a functional conservation between species. These functional conservations were proved by genomic and RNA targeting approaches. For example, lncRNA Cyrano from zebrafish and mammalian orthologs were having functional conservation during embryogenesis[95]. Comparative analysis between zebrafish and mammalian lncRNAs showed syntenic conservation by having the same protein-coding neighbors. Though, they differed in their length and number of splice sites (exons), similar functions were perturbed. Additionally, the advanced predication algorithms also found evolutionarily conserved secondary structures of lncRNAs among many species[96]. Recent study using computational approach also predicted conserved *k*-mers (consensus motif sequence) profiles in a evolutionarily unrelated lncRNAs having similar function[97]. Despite their low sequence similarities, some lncRNAs are shown to be structurally and functionally conserved to perform its functions in different species[98].

### 1.3.3 Functional significance of lncRNAs

LncRNAs are characterized based on their nature of tissue or cell type specificity[7,99]. Unlike, protein-coding genes, lncRNAs function differently in different cell types and tissues. Despite, the lacking evidence that the lncRNAs have a consensus mechanisms, they have been implicated in various cellular and molecular functions[100].

***Cell cycle regulation***: There are different stages of the cell cycle (**section 1.1.2**) and it is maintained by specific regulators such as Cyclin kinases (CDKs), CDK inhibitors, apoptotic factor (p53) and cell division check point regulator (phosphorylated RB). Upon DNA damage during the cell cycle progression, certain lncRNAs are activated to initiate cell cycle arrest. Also, some lncRNAs induce p53 mediated DNA damage response leading to apoptosis[8]. Additionally, *MALAT1* lncRNA in fibroblast cells is known to regulate the expression of cell cycle-associated genes necessary for G1/S transition and mitosis[101]. Similarly, *HOTAIR* lncRNA acts as an oncogene (promotes tumor progression) having elevated expression in lung cancer. *HOTAIR* promotes cell cycle progression by disrupting the activated restriction barrier during G1/S transition[102]. Thus, it is important to study lncRNAs that are associated with the cell cycle in both normal and disease progression.

***Chromatin regulation***: Most of the lncRNAs are known to function by interacting with chromatin and chromatin modifiers to maintain the chromatin architecture[100,103-105]. During mammalian development one of the maternal X-chromosome is transcriptionally kept silent. The higher expression of *Xist* lncRNA is seen on the inactivated X-chromosome rather than the active one. *Xist* keeps whole chromosome silent by suppressing expression of genes in *cis*. Another lncRNA called *Tsix* originated from *Xis*t antisense locus specifically blocks accumulation of *Xist* in *cis*. Thus, it is important to keep *Tsix* repressed for successful inactivation of X-chromosome by allowing *Xist* RNA accumulation[106]. Some lncRNAs are shown to interact with chromatin and recruit chromatin-modifying enzymes to makes the chromatin accessible or keep it repressed[107]. Paternally transcribed lncRNA *KCNQ1OT1* also interacts with H3K9me3 and H3K27me3, recruits PRC2 to the chromatin to transcriptionally suppress target protein-coding genes *in cis* within *KCNQ1* domain[92,108,109]. This helps in maintenance of lineage-specific (paternally imprinting) gene regulation. LncRNA *MEG3* is also known to recruit PRC2 (Polycomb repressive complex 2) chromatin modifier to the chromatin to

establish H3K27 tri-methylation and to modulate the expression of *MEG3* target genes. In the same study, several lncRNAs are shown to be associated with inactive chromatin by interacting with both PRC2 and H3K27me3 repressive chromatin mark[91]. Additionally, in Facioscapulohumeral muscular dystrophy (FSHD), *DBE-T* lncRNA recruits chromatin-modifying Trithorax Ash1L protein to chromatin which di-methylates H3K36 to activate transcription of FSHD genes[110]. Also, *HOTAIR* is known to function by interacting with chromatin in development and disease[92,111,112]. Detailed chromatin-based lncRNA mechanisms are discussed in **section 1.5**.

***Developmental regulation***: The transcriptome-wide studies have proved that lncRNAs play an important role during development[113]. There are clusters of lncRNAs know to have a role during embryonic germ layer lineage commitments (ectoderm, endoderm and mesoderm) after implantation[114-118]. Additionally, lncRNA *Fendrr* is expressed specifically in mesodermal cell types in murine cells. It helps in the development of heart and body wall in a lineage-specific manner[119]. Co-expression analysis revealed that the lncRNAs with similar expression patterns to neighboring protein-coding genes were having a key role during the cleavage stages of embryonic development. Importantly, these lncRNAs are involved in several functions such as cell cycle, transcription and translation during embryo development[120]. It was previously noted that the developmentally-associated RNAs are derived from the oocyte (maternal expressed)[121]. Sperms were considered to be transcriptionally silent due to their compact nuclei and chromatin configurations. It was considered that the sperm merely deliver the paternal genome to oocyte during fertilization. Recent studies have shown many natural antisense lncRNA transcripts (NATs) are expressed during spermatogenesis[55]. These studies have proved that despite their compact nucleus, sperm carries RNAs that play a transcriptionally important role during embryo development[122-128]. Thus, lncRNAs are needed to be extensively studied to unveil their exact molecular mechanism during early embryo and organismal development.

**Chromatin and transcriptome-based integrative approaches to profile functional lncRNAs**

## 1.4 LncRNAs in cancer

The emergence of high-throughput sequencing and completion of the human genome project paved a way to characterize functional lncRNAs. LncRNAs are shown to be critical regulators in various cancers. These can either promote (oncogene) or suppress (tumor suppressor) cancer progression through different mechanisms depending on the cancer type (**Figure 12**). Significant efforts from TCGA (The Cancer Genome Project) has identified hundreds of lncRNAs in various cancers[129-132].

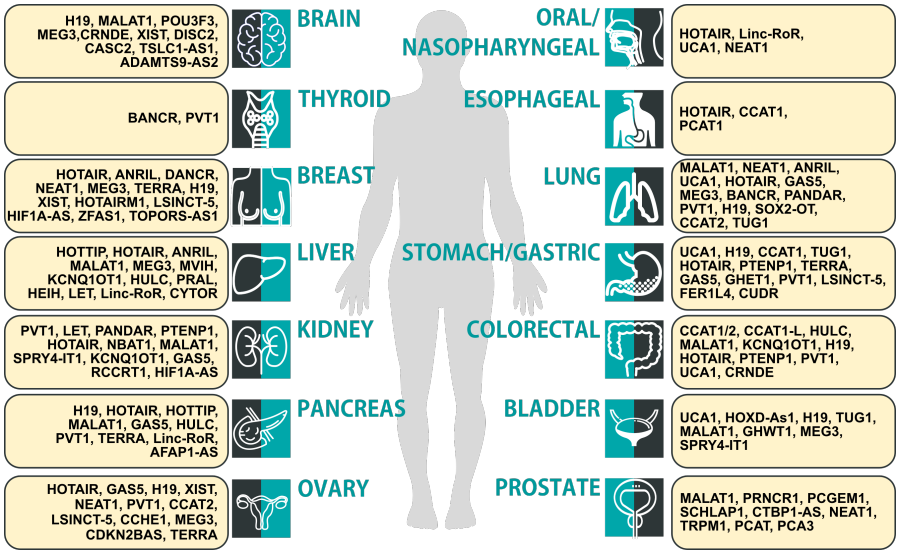


***Figure 12.*** *List of oncogenic and tumor suppressor lncRNAs in different cancers.*

### 1.4.1 Oncogenic properties of lncRNAs

***HOTAIR***: It is an intergenic lncRNA, known to promote tumor growth, invasion, migration, and metastasis via various signaling pathways (**Figure 12**). *HOTAIR* serves as a predictive oncogenic biomarker in multiple cancers such as breast, liver, kidney, pancreatic, ovarian, oral, esophageal, lung, gastric and colorectal cancers. *HOTAIR* interacts with chromatin-modifying enzymes (PRC2 or LSD1) and recruits them to the target gene loci to alter methylation patterns (H3K27 tri-methylation and H3K4 demethylation) at the promoters. In many cancers, this H3K27 tri-methylation and H3K4 demethylation causes inactivation or silencing of tumor suppressor genes leading to metastasis[111,133-135].

***MALAT1***: *MALAT1* is one of the first lncRNA shown to be associated with cancer[136]. It is a highly predictive marker for cancer cell metastasis in lung adenocarcinoma[136,137]. Recent evidences also suggest oncogenic properties of *MALAT1* in many cancers such as brain, colorectal, liver and pancreatic cancers (**Figure 12**).

***CRNDE***: *CRNDE* has both protein-coding and noncoding properties. One of the *CRNDE* transcript variant or isoform codes for short peptide[88]. LncRNA *CRNDE* (Colorectal neoplasia differentially expressed) is found to be highly expressed having greater specificity and sensitivity in colorectal cancers (CRC)[138](**Figure 12**). Also, it is shown that demethylation (hypomethylation) of *CRNDE* lncRNA promoter leads to its elevated expression in patients with chronic lymphocytic leukemia (CLL). In the healthy individuals, the *CRNDE* promoter is hypermethylated, suppressing the expression of the lncRNA[67]. The higher and lower methylation levels of *CRNDE* promoters in CLL patients predicts the survival outcome of the patients[67].

***PCAT1 and NEAT1***: *PCAT1* is shown to be highly expressed in prostate cancer and contributes to cancer cells proliferation (**Figure 12**). It is a transcriptional repressor lncRNA that modulates its cell division associated target genes by a *trans*-regulation mechanism[139]. Recent studies have also shown that *PCAT-1* lncRNA activates AKT and NF-kB signaling in castration-resistant prostate cancer patients[140]. *NEAT1* has an oncogenic property known to control cancer-specific cellular processes such as, cell proliferation, invasion, migration, apoptosis and more importantly DNA damage in prostate cancer cells[141-143].

### 1.4.2 LncRNAs as tumor suppressors

***NBAT1 and CASC15***: Neuroblastoma is a type of extracranial solid tumor; that causes improper neuronal differentiation at the sympathetic nervous system. The high-risk patient groups show lower levels of *NBAT1* and *CASC15* expression. Overexpression of these lncRNAs promote proper neuronal differentiation by modulating pathways related to cell proliferation and migration. Higher expression of these lncRNAs in NB shows a good prognosis in patients by serving as a tumor suppressor lncRNAs[86,144].

***GAS5***: Patients with lower levels of *GAS5* have poor survival compared to patients having higher *GAS5* in gastric and colorectal cancers[145,146](**Figure 12**). Lower levels of *GAS5* promotes cell proliferation in gastric cancer[145]. Moreover, the lower levels of *GAS5* lncRNA is also seen in pancreatic cancer[147]. *GAS5* overexpression in pancreatic cancer cells causes decreases the proliferation, by inhibiting CDK6, which in turn decreases G1/G0 and increases S-phase stages of cell cycle[148].

***MEG3***: *MEG3* lncRNA is downregulated in breast cancer, whereas its over expression inhibits tumor progression and growth (**Figure 12**). It inhibits miR-21-mediated activation of PI3K/Akt signaling pathway by suppressing the activity of miR-21[149]. Moreover, chromatin-associated *MEG3* also targets TGF-beta pathway genes in breast cancer cells by forming RNA-DNA triplex structures[91].

**Chromatin and transcriptome-based integrative approaches to profile functional lncRNAs**

## 1.5 Mechanisms of lncRNAs

As described earlier, lncRNAs regulate their target genes via multiple mechanisms. Their known mechanisms in nuclear compartment includes regulation of chromatin loops, regulation of gene splicing by recruiting or inhibiting splicing factors and recruiting transcription factors to the gene promoters and chromatin modifiers to the histones. Additionally, lncRNAs also perform some functions in cytoplasmic compartment such as controlling rate of protein translation and helping in mRNA decay (**Figure 13**).

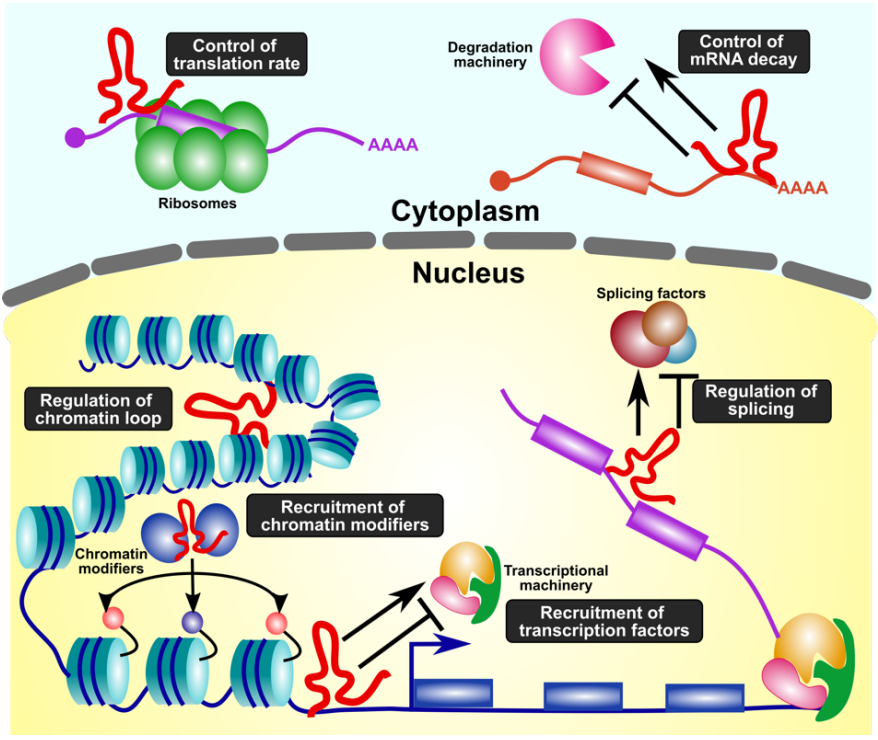


***Figure 13.*** *Different mechanisms by which lncRNAs regulate their target genes, genome and chromatin. Red colored secondary structure in the figure represents lncRNAs.*

### 1.5.1 Transcriptional and translational regulation

Transcription factors and enhancers are the key players in modulating gene expression. There are different means that the transcription factors bind to the target gene promoters (**Figure 13**). However, lncRNAs are shown to guide these transcription factors to the promoters to activate or suppress the transcription of the genes. The lncRNA transcribed from *DHFR* alternate upstream promoter (*lncRNA-DHFR*) is known to reduce the occupancy of transcription factor TFIIB to the major promoters *in cis*, which in turn disrupts the assembly of the transcription preinitiation complex. In this case, the lncRNA regulates the transcription by forming a stable RNA-DNA triplex complex with the major promoters and also by binding to transcription factor TFIIB protein[150]. During the process of transcription, a gene transcribes into pre-mRNA and this pre-mRNA undergoes splicing to become matured mRNA. There are many transcription factors involved in the splicing mechanism. It is important that splicing factors binds to pre-mRNA to complete the transcriptional regulation (**Figure 13**). *MALAT1* lncRNA recruits splicing factor SF2/ASF to the actively transcribing neuronal genes. Upon knockdown of *MALAT1*, the hippocampal neurons are significantly reduced due to defects in SF2/ASF recruitment[151]. However, *MALAT1* lncRNA is also shown to regulate TGF-beta binding protein (LTBP3) by recruiting Sp1 transcription factor to *LTBP3* promoter. Moreover, mouse *Evf2* lncRNA binds and forms a complex with DLX2 transcription factor protein to regulate their target protein-coding genes *DLX5* and *DLX6*[152]. Hence lncRNAs play an important role in promoting or suppressing transcription of genes by direct or indirect interaction with the transcription factors[153,154](**Figure 13**). The transcribed protein-coding genes are exported to the cytoplasm to become a stable functional protein with the help of translation machinery. In HeLa cells (cervical carcinoma cells), the RNA-binding protein HuR binds to *lincRNA-p21* and recruits *let7* to decrease the stability of *lincRNA-p21*. This leads HuR to target cancer signaling JunB and β-catenin mRNAs and increase their translation. When the HuR levels are lowered, *lincRNA-p21* is accumulated in the cells and this decreased the translational efficiency of JunB and β-catenin mRNAs[154,155]. Overall, this shows the importance of lncRNAs in transcription, splicing and translational regulation (**Figure 13**).

### 1.5.2 Chromatin regulation

Several lncRNAs are known to interact with chromatin modifiers to define the chromatin states of the genome. The chromatin-associated lncRNAs recruit chromatin modifiers to the chromatin to maintain the chromatin states in a cell-specific manner[156](**Figure 13**). *HOTAIR* is the first developmentally associated lncRNA shown to epigenetically repress its target genes in *trans* by H3K27me3 deposition through recruitment of PRC2 (Polycomb repressive complex 2) at the chromatin[157]. Paternally transcribed lncRNA *KCNQ1OT1* also interacts with H3K9me3 and H3K27me3, recruits PRC2 to chromatin and maintains lineage-specific (paternally imprinted) transcriptional silencing of its target protein-coding genes *in cis* within *KCNQ1* domain[92,108,109]. Similarly, *MEG3* chromatin-interacting lncRNA modulates expression of TGF-beta target genes by recruiting PRC2 to chromatin and establishing H3K27me3 at the distal promoters *in trans*[91]. PRC2 is a histone methyltransferase complex involved in epigenetic silencing of the genome during development and disease. In addition to recruitment of chromatin modifiers, lncRNAs also take part in chromatin conformation (3D) or chromatin looping. The genome contains enhancers regions that amplify transcription of other genes by assembling chromatin modifiers and transcription factors. These enhancer regions transcribe and produce a set of long noncoding RNAs called enhancer RNAs (eRNAs). The activity of enhancers at the genome is shown to be well correlated with levels of eRNAs[158-160]. Chromatin loops or enhancer-promoter loops are formed when distant enhancers and promoters are come into contact. Studies have found that upon eRNA knockdown the chromatin loop structures are impaired or having defective loops[160-162]. These eRNAs interact with the cohesin protein complex to maintain the stability of the chromatin loop. Cohesin together with DNA-binding protein CTCF forms a complex and are known to establish chromatin loops[163](**Figure 9**). In summary, lncRNAs can both recruit chromatin modifiers and also helps in forming stable chromatin loops to maintain the chromatin architecture[46](**Figure 13**).

### 1.5.3 Role in mRNA decay

The levels of mRNA dictate the efficiency of the production of proteins. These mRNA levels are determined in the cells by two key factors such as the rate of transcription and mRNA decay (**Figure 13**). Staufen 1 (STAU1)-mediated messenger RNA decay (SMD) involves the formation of STAU1 binding site (SBS) by lncRNA *1/2-sbsRNA* (polyadenylated long noncoding RNA). SBS is formed by improper base pairing between Alu elements at the 3'-UTR regions of target mRNA and the Alu element at the lncRNA *1/2-sbsRNA*. This leads to binding of SBS-transactivated STAU1 to the target mRNA and in turn recruit UPF1 protein to mediate mRNA decay[164]. In mouse lncRNA *1/2-sbsRNA* forms SBS by short interspersed element (SINE) to mediate SMD degradation pathway influencing developmental process[165]. Thus, lncRNAs are involved in very complex mechanisms and sometimes these mechanisms vary between organisms (**Figure 13**).

## 1.6 Next generation sequencing techniques

Major challenges addressing problems in molecular biology are greatly tackled by recent advancements in high-throughput sequencing technologies. With the help of recent sequencing techniques and high computational power from sequencing platforms, we are now able to sequence millions of DNA or RNA fragments in parallel. Currently available high-throughput sequencing technologies are commonly known as next-generation sequencing. NGS can be broadly classified as transcriptome-based and genome-based.

### 1.6.1 Transcriptome based

***RNA-seq***: As described earlier, genes in the genome undergo transcription and gives rise to mRNA. Using this transcriptome sequencing technique (RNA-seq), the number of transcribed mRNAs (expression) per genic location can be captured and quantified. The RNA-seq technique is useful in estimating the global expression patterns of transcribed RNAs. With RNA-seq, we can quantify the difference in gene expression between control vs. treatment (conditioned) cells; patient vs. healthy normal tissues; temporal expression patterns of genes at different time points and tissue-specific expression of genes.

*Library preparation*: The protocol starts with isolating the RNA from the cells (two or more conditions) of interest. Next, the isolated RNA is fragmented into shorter fragments. This is because most of the sequencing machines can only handle fragments between 36-300 bp length. Since this follows random fragmentation, it gives the advantage of getting overlapping fragments for the same location. Later fragmented RNA is converted into double stranded DNA (dsDNA). These dsDNA fragments are more stable than RNA and can be easily amplified. In the next step, the adapters are added to both ends of dsDNA fragments (adapter ligation). The adapters are the tags that sequencing machines can easily recognize. Since different samples (conditions) can be treated with different adapters, the machine can sequence multiple samples at the same time. In the final step, the fragments with adapters are subjected to

PCR amplification to get a complementary DNA (cDNA) and the adapter-ligated PCR amplified fragments are then enriched in the prepared library. Before these fragments are sequenced, it undergoes quality checks to know the size of the library and the fragment length.

*Sequencing*: Obtained DNA fragments are attached to a grid called flow cell. This flow cell contains fluorescent probes color-coded according to the type of nucleotide (A: adenine, T: thymine, G: Guanine and C: cytosine). Next the probes are allowed to attach to the first nucleotide of the DNA fragments, captures the colors and then washes off the colors in the grid. Similarly, the colors are captured for the entire length of all fragments base-by-base. In some regions of fragments, the color of probes is not brighter to be read by the sequencing machine. Those bases are assigned as low-quality bases by the sequencing machine by giving them low-quality scores. Additionally, if there too many probes of the same color at the same regions (low diversity bases), those are also read as low-quality regions. Usually, this occurs at the beginning of the fragments when the machine starts recognizing its first few bases. These regions are later recognized as overrepresented sequence regions while performing bioinformatics analysis. Depending on the demand, the sequencing can be done using slightly different protocols such as single-end (sequence from one end), paired-end (sequence both ends of the fragments) and strand-specific library (maintaining the directionality of the fragment in the genome, sense or antisense). The detailed information about raw data format and the analysis of RNA-seq is described in the **section 4.3**.

### 1.6.2 Genome based

***WGS***: Whole Genome Sequencing (WGS) techniques are used to sequence the entire genome of an organism. WGS is implemented to serve several aspects, including 1) evolutionary studies to perform comparative analysis, 2) assembling genomes by *de novo* sequencing for newly discovered species for which reference genomes are not available, 3) finding somatic mutations (SNV and CNV) genome-wide, and 4) identifying large-scale structural alterations and genomic rearrangements.

*Library preparation and sequencing*: WGS involves similar library preparation and sequencing steps. One exception is, instead of extracting transcribed RNA, the whole genome is fragmented and subjected to sequencing.

***Exome-seq***: Whole Exome Sequencing (WES) is similar to WGS except, this is a targeted approach that covers only coding regions of the genome. WES techniques are used to find somatic mutations (SNV and CNV) and genomic rearrangements, and identify fusions at the DNA level, but only on the coding regions. This approach is cost and time-efficient compared to WGS since it involves sequencing only part of the genome (exome).

### 1.6.3 Epigenome based

***ChIP-seq***: Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) is a chromatin-based sequencing approach to find locations in the genome bound by proteins. As we know that the chromatin consists of DNA wrapped around histones and the states of chromatin determines the transcription of the genes. In DNA, all kinds of the protein binds either to promote or suppress the transcription of the genes, and these factors are known as transcription factors (TFs) (refer **section 1.2.3**). Therefore, it is important to know the gene or genomic targets of a particular transcription factor. In addition to that, it is also important to know the localization of different types of histone modifications at the genome to know the chromatin states (refer **section 1.2.3**). Transcription factors (TFs) and histone modification profiles differ between different biological model systems. We perform ChIP-seq to find gene targets of particular transcription factors or to characterize chromatin-wide histone modification profiles. For this purpose, specially designed antibodies are available for various known transcription factors and histone modifications.

*Library preparation and sequencing*: First formaldehyde cross-linking is performed to cross-link all the proteins to DNA. This includes all DNA binding proteins, not just one. Next, the DNA or the chromatin is fragmented into shorter fragments (~300 bp). In the next step, we isolate the proteins by immunoprecipitation using special antibodies (example, TF protein: anti-CTCF or histone protein: anti-H3K4me3). The formaldehyde cross-linking is then reversed using heating technique. As a final step we isolate the DNA by washing away the proteins and histones. The whole process usually done chromatin-wide in a pool of at least 6 million cells. Once we get the DNA, we follow the similar procedure described in RNA-seq library preparation (Adapter ligation, PCR amplification, check library size and length of the fragments) and subject to sequencing. Computational analysis of ChIP-seq is described in **section 4.4**.

***DNase-seq***: DNase-seq is used for purifying nucleosome-free DNA regions. If the chromatin in the genome is nucleosome-free, means the DNA is unwrapped from histones and is accessible to any proteins or transcription factors. These nucleosome-free regions are regulatory regions in the genome where the transcription factors bind and performs its function in a cell type specific manner. DNase-seq takes an advantage that these regulatory regions are sensitive to the DNase-I enzyme that digest nucleic acids. DNase-I enzyme cuts the nucleosome-free DNA regulatory region into small pieces of DNA fragments.

*Library preparation and sequencing*: Obtained DNase-I hypersensitive regulatory DNA fragments are subjected to library preparation and sequencing as described in **section 1.7.1**.

***WGBS-seq***: Whole Genome Bisulfite sequencing (WGBS-seq) technique allows us to detect genome-wide DNA methylated regions. This allows us to evaluate differentially methylated regions (DMRs) in the genome between two biological conditions (control vs. treatment cells, patient vs. healthy normal tissues etc.,). Also, this helps in identifying the epigenetic states (active: unmethylated or silent: methylated) of gene promoters. In contrast, recent studies have shown that gene body methylation has opposite effects on gene transcription (active: methylated or silent: unmethylated)[39,67,68].

*Library preparation and sequencing*: This protocol involves fragmentation of genomic DNA. Next the fragmented DNA is treated with sodium bisulfite. Upon bisulfite treatment, the unmethylated cytosines are converted into U (uracil) which is later converted back to thymine (T) while subjected to sequencing. On the other hand, the methylated cytosines or 5-methylcytosine (5mC) are protected from the conversion and remain as cytosine throughout the process of sequencing. This bisulfite conversions distinguishes the methylated regions from unmethylated regions. One close alternative for this technique is RRBS-seq (Reduced-Representation Bisulfite sequencing) where it uses an enzyme that specifically targets only methylated sites. This method is cost-efficient compared to WGBS-seq but also has a disadvantage that it does not capture all methylated regions or promoters[166]. There are also other variants to WGBS-seq such as MeDIP-seq and MBD-seq/MethylCap-seq. In MeDIP-seq antibody against 5-methylcytosine (5mC) is used to enrich those methylated regions. Whereas in MBD-seq, it uses methyl-binding domain protein to target those methylated regions. After obtaining the DNA with bisulfite conversion (WGBS-seq) or without conversion (MeDIP-seq or MBD-seq), the library preparation is continued (Adapter ligation, PCR amplification,

check library size and length of the fragments) and subjected to sequencing as described in **section 1.7.1**.

### 1.6.4 Interactome based

***ChRIP-seq***: As described in **section 1.7.3**, in a given ChIP-seq technique we can identify the enriched genomic regions by a particular protein of interest. In ChRIP (Chromatin RNA immunoprecipitation), the RNA is extracted from immunopurified chromatin (antibody pulldown against protein/histone of interest) and subjected to sequencing. ChRIP-seq can isolate specific RNA transcripts from chromatin and its sub-compartments. This method is useful in finding RNA transcripts interacting with different chromatin compartments (active or repressive chromatin)[91,92,167,168].

*Library preparation and sequencing*: First the cells are treated with Actinomycin D (ActD) to block ongoing transcription and to discard nascent transcripts attached to the chromatin. Next, wash-off residual ActD and treat the cells with formaldehyde. Treatment of formaldehyde is then followed by UV radiation crosslinking. In the next step, an antibody against proteins or histones are used for chromatin immunoprecipitation. For example, to extract active chromatin interacting RNA transcripts, one can use an antibody against H3K4me3 and the active chromatin modifier WDR5 (methylates H3K4). At last, proteinase K treatment is done to get rid of proteins followed by the reversal of formaldehyde cross-linking by heating. Then the chromatin-interacting RNA transcripts are isolated using TRIzol RNA extraction protocol. The DNA contamination in the prepared samples is avoided by treating the extracted RNA with DNase I and again extract the chromatin-interacting RNA using TRIzol. In the final step, the chromatin-interacting RNA transcripts are subjected to sequencing. This follows same sequencing procedure used for RNA-seq in **section 4.3**. For control or input DNA sample, the chromatin without any antibody treatment is used.

***ChOP-seq***: Chromatin Oligo affinity Purification technique is used for detecting RNA-DNA interactions genome-wide. This is a targeted approach to find binding sites or occupancy of a particular RNA in the genome[91,169,170].

*Library preparation and sequencing*: This involves designing of *in silico* antisense DNA probes (biotin labeled) covering a full-length RNA of interest. It is also important to discard off-target probes mapping to multiple locations

in the genome. As a negative control, the probes against reporter genes (example, GFP) are also constructed. First, cross-linking procedures are done using glutaraldehyde allowing proteins to bind to DNA. Next, the DNA or the chromatin is fragmented. Now, the RNA probes are allowed to hybridize followed by capturing the biotin labeled probes using streptavidin beads. Then, the RNA is washed away using RNA elution buffer. At last, the isolated DNA is subjected to sequencing.

***More interactome techniques***:

*GRID-seq*: Many studies have proposed that a huge number of RNAs (coding and noncoding) interacts with the genome in eukaryotes. GRID-seq (Global RNA interactions with DNA followed by deep sequencing) efficiently maps these interactions on a global scale. GRID-seq can detect all chromatin-interacting RNAs on a global scale and accurately maps the binding sites of these RNAs in the genome[171].

*HiC*: High-resolution capture of chromatin conformation space or 3D chromatin loops. This gives us information about short and long-range interactions within the genome in a chromatin space. In simple terms, this method is useful for finding DNA-DNA interacting regions within and between chromosomes in a single nucleotide resolution[172].

# 2 AIM OF THIS THESIS

Since the noncoding portion of the genome is poorly investigated, the present thesis is focused on a major class of noncoding RNAs, such as long noncoding RNAs (lncRNAs). There is a need for comprehensive tools to explore the functions of lncRNAs originated from this noncoding portion of the genome in different biological contexts, to understand its regulation fully. The focus of this thesis is to provide a comprehensive list of functional long noncoding RNAs by utilizing advanced high-throughput sequencing technologies as well as computational and statistical methods. Advancements in computational cloud-based clusters and statistical tools are becoming very useful in tackling complex problems of molecular biology and in cancer. This thesis utilized high-throughput data from various sequencing techniques as well as different computational approaches to profile functionally relevant lncRNAs.

**Paper I**: Initial study explored the possible role of cell cycle-associated long noncoding RNAs in different cancers by analyzing transcriptome-based dataset (nascent RNA capture)[173].

**Paper II**: In the second study, we focused on characterizing active chromatin-associated lncRNAs by analyzing sequencing data from chromatin RNA immunoprecipitation technique (ChRIP-seq)[174].

**Paper III**: The final study deals with exploring the transcription and chromatin dynamics of sperm-derived long noncoding RNAs throughout development and their role in cancer[175].

**Chromatin and transcriptome-based integrative approaches to profile functional lncRNAs**

# 3 RESULTS

## 3.1 S-phase and cancer associated lncRNA transcripts (paper I)

**Temporal expression of S-phase associated lncRNAs**

The nascent RNA capture technique followed by sequencing is optimized to find temporally expressed lncRNAs during the S-phase compartments of the cell cycle[176]. The HeLa cells were synchronized at the G1/S transition stage of the cell cycle using thymidine and hydroxyurea. Following the release of G1/S blockage, the nascent RNA transcription was allowed in the presence of EtU (5-ethynyl-uridine). Therefore, the newly synthesized S-phase nascent RNA species were specifically labeled with EtU and captured across three different S-phase time points (T1: 0-2h, T2: 1.5-3.5h and T3: 3-5h)[173](**Figure 14a**). Additionally, the efficiency of the labeling technique was validated by comparing EtU labeled with the unlabeled RNA-seq samples (**Figure 14b**). There were 1,145 S-phase temporally expressed lncRNAs in synchronized EtU labeled samples at three time points compared with the unsynchronized sample (**Figure 14c**). Also, 394 lncRNAs from EtU labeled samples were also found in unlabeled samples (**Figure 14d**).

**S-phase Cancer-Associated lncRNA Transcripts (SCATs)**

Obtained S-phase temporally expressed lncRNAs were tested for their differential expression status between normal and cancer patients from TCGA (16 cancer types). Out of 1,145, there were 570 S-phase lncRNAs that are significantly differentially expressed in at least one of the cancer types. For this purpose, the RNA-seq samples from TCGA were analyzed for differential expression analysis. In addition to that, DNA methylation status of these lncRNAs was also verified in patients and healthy individuals to check if it correlated with the expression status of that particular lncRNA. Later top candidates were selected based on the number of cancers in which particular lncRNA is differentially expressed, having a correlation with methylation

status and their significance in predicting clinical outcomes. One of the top candidates was lncRNA *RP11-465N4.4,* and it was named as S-phase Cancer Associated Transcript 7 (*SCAT7*). LncRNA *SCAT7* was upregulated in patients with more than 4-fold and having correlative promoter hypomethylation (low levels of methylation compared to healthy individuals) (**Figure 14e**). This shows how epigenetic alterations can define the transcriptional status of lncRNAs in cancers.

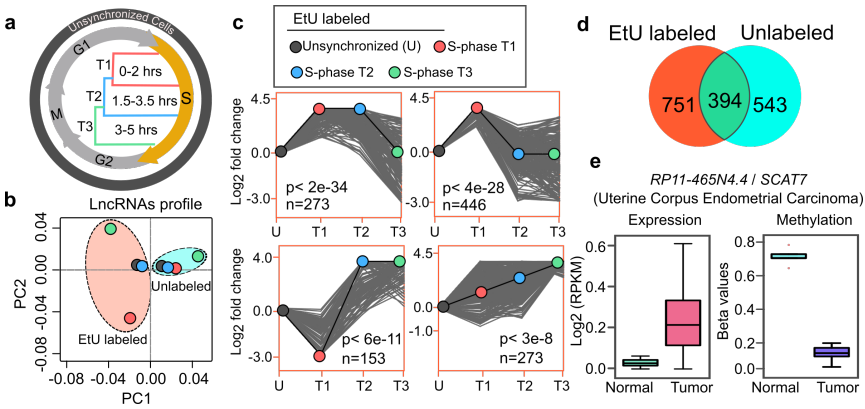


***Figure 14. a)*** *Transcripts collected at S-phase stage of cell cycle at different time points (synchronized and unsynchronized).* ***b)*** *Principal component analysis (PCA) of lncRNA profiles from three time points of EtU-labeled and unlabeled transcriptome samples.* ***c)*** *STEM analysis with significant temporal patterns of lncRNA expression in three time points compared with unsynchronized control samples.* ***d)*** *Commonly expressed lncRNAs from EtU-labeled and unlabeled.* ***e)*** *Expression and methylation status of top S-phase cancer-associated lncRNA Transcript (SCAT7) in TCGA patients.*

## *SCAT7* regulating FGF signaling via hnRNPK/YBX nucleoproteins

Knockdown followed by RNA sequencing was performed by downregulating *SCAT7* lncRNA in HeLa, A549, and CaKi-2 cells using siRNAs (Small Interfering RNAs) to validate top lncRNA candidate (**Figure 15a**). The differential expression analysis was performed between control-si and *SCAT7*-si to find altered proteins and affected pathways. The knockdown of *SCAT7* in all three cell lines revealed that *SCAT7* regulates genes involved in the cell cycle, FGF signaling, and cellular senescence pathways[177](**Figure 15b**). Importantly FGF/FGFR signaling and its downstream pathways such as

PI3K/AKT and Ras/MAPK were also affected as seen in other cancers[178-182]. To further explore the mechanism *SCAT7* in regulating FGF/FGFR signaling, additional ChOP (Chromatin Oligo-affinity Purification) experiments were performed to find *SCAT7* interacting proteins. Two of the well-known cell cycle co-functional proteins, such as YBX1 and hnRNPK[183-185], were found to be top interacting proteins with *SCAT7*. Further, the interaction of *SCAT7* RNA and the proteins YBX1 and hnRNPK was confirmed using RNA immunoprecipitation, and ChOP followed by western blot. In summary, *SCAT7* lncRNA promotes the transcription of FGF/FGFR signaling genes by recruiting YBX1-hnRNPK nucleoproteins to their promoters and, in turn, affects the FGF/FGFR downstream pathways (**Figure 15c**).

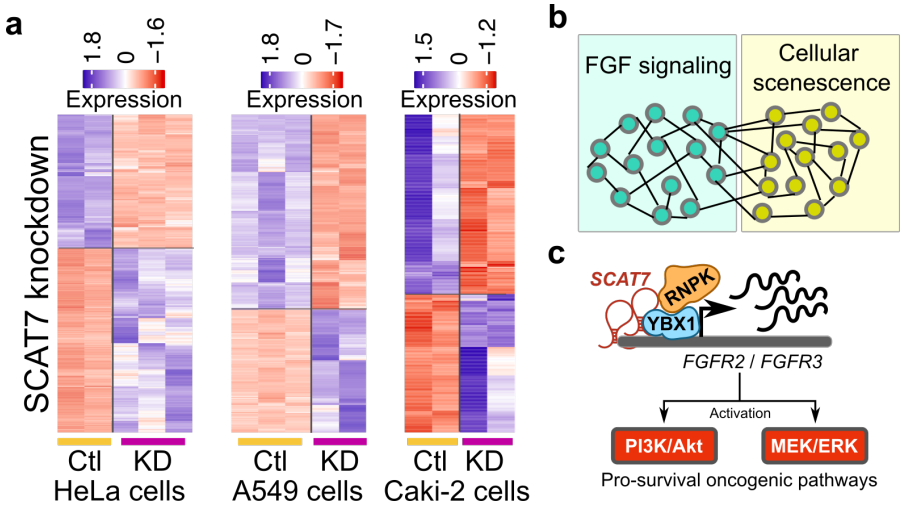


***Figure 15. a)*** *Heatmaps show differentially expressed genes from SCAT7 knockdown in Hela (cervical), A549 and Caki-2 (kidney) cell line compared to control cells. Red indicates higher expression and blue indicates lower expression.* ***b)*** *Pathways (FGF signaling and cellular senescence) affected by SCAT7 knockdown.* ***c)*** *Mechanism by which SCAT7 regulates its oncogenic signaling pathways via recruiting YBX1 and hnRNPK complex to the FGF promoters.*

## Significance and conclusion

This study is an alternative and efficient approach to find lncRNAs involved in the cell cycle (one of the major cancer cell hallmarks) using a newly optimized technique such as nascent RNA capture assay[186]. Our study only investigated and explored the mechanism of one of the top clinically relevant candidates in cancer by integrating the pan-cancer dataset. There are other lncRNA candidates that might have different mechanisms in different cancer types. It is important to explore such candidates from this study. Overall this study provides a comprehensive list of S-phase cancer-associated lncRNAs and a huge resource for future investigations.

## 3.2 Active chromatin associated lncRNAs (Paper II)

### Chromatin RNA immunoprecipitation and sequencing

Chromatin in the nucleus is either found in an active or repressive state. There are studies reporting that the lncRNAs can be associated with these chromatin states. In this study, Chromatin RNA immunoprecipitation (ChRIP) was performed, followed by high-throughput sequencing to find lncRNAs associated with active or open chromatin state[91,168,174]. To achieve this, modified ChRIP was performed in BT-549 (breast cancer cell line) cells treated with Actinomycin D (ActD) to prevent co-transcriptional crosslinking of lncRNA to chromatin. Then, formaldehyde treatment followed by nuclei isolation and sonication was done[91,168,174]. Next, the antibodies against active chromatin mark H3K4me2 and active chromatin reader WDR5 were used for pulldown. WDR5 is a part of MLL-SET1 complex, known to interact directly with di-methylated H3K4 and it is required for transition of H3K4 di-methylated state into tri-methylated state[187]. Also, H3K4me2 marks the transcription factor binding site and is also shown to have tissue-specific regulation[188,189]. Later these cells were subjected to reverse crosslinking, RNA purification, and DNase treatment. At last, the purified RNA was sequenced using a high-throughput sequencing platform.

### Active chromatin-associated lncRNAs

Obtained ChRIP-seq samples from H3K4me3 and WDR5 were compared with the input DNA sample to find enriched lncRNAs in the active chromatin compartment. There were 209 lncRNAs found to be enriched in active chromatin (H3K4me3 and WDR5) and are termed as active chromatin-associated lncRNAs (Active lncCARs) (**Figure 16a**). We also found strong evidence that these active lncCARs interact with both H3K3me3 and WDR5.

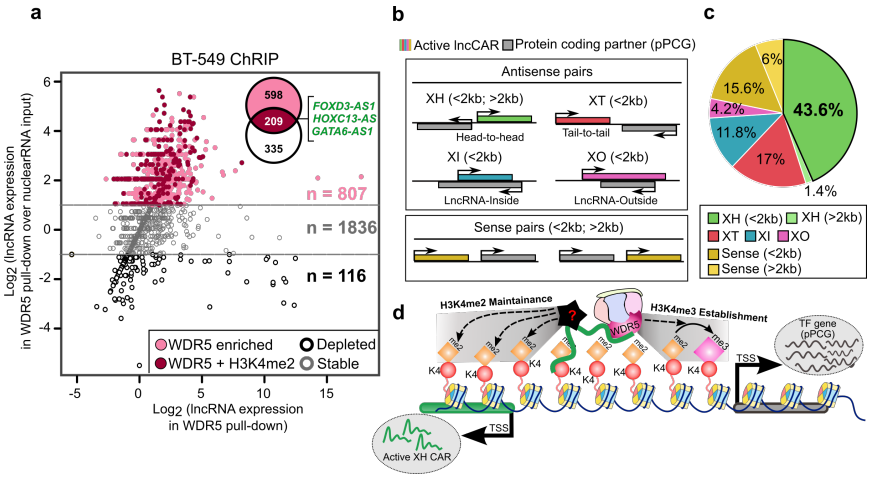


***Figure 16. a)*** *Scatter plot showing enrichment of WDR5 lncRNAs over nuclear input. The X- axis denotes log transformed expression and Y-axis denotes log-fold change of WDR5 lncRNAs over nuclear input. Venn diagram, in the Scatter plot, shows 209 lncRNAs that are commonly enriched in both H3K4me2 and WDR5 ChRIP pulldowns.* ***b)*** *Genomic organization of H3K4me2 lncRNAs (green bars) with respect to nearest protein-coding genes (grey bars). XH: where lncRNA shows Head-to-Head arrangement with nearby protein-coding gene. XT: lncRNA in Tail-to-Tail arrangement with protein-coding gene, XI: lncRNA located inside a protein-coding gene, XO: lncRNA located outside and covers the entire protein-coding gene. Sense pairs means lncRNAs in the same orientation as protein-coding gene. The notion of greater or less than 2 kb means that the partner genes (pPCGs) are located within 2 kb (<2 kb) or away from 2 kb (> 2 kb) but within 50 kb window with respect to H3K4me2 lncRNAs.* ***c)*** *Distribution of different patterns of genomic arrangements as described in* ***(b)*** *for active chromatin-associated lncRNA (ChRIP pulldown using H3K4me2 and WDR5 antibodies).* ***d)*** *Model depicting transcriptional regulation of divergent gene loci by active XH lncCARs. Active XH lncCARs transcribed from divergent transcription units are targeted to specific regions around TSS of partner protein-coding gene that is enriched with H3K4me2, through their chromatin interaction property. Promoter targeting of these active XH lncCARs helps in maintaining H3K4me2 in WDR5 independent manner. XH lncCAR-dependent recruitment of WDR5 at the active XH transcription units helps in the conversion of H3K4me2 into H3K4me3, thereby establishing transcriptionally competent chromatin at the divergent transcription units.*

**Genomic organization of Active lncCARs**

Interestingly 78% of the active lncCARs are arranged in an antisense orientation to the neighboring protein-coding genes. These active lncCARs were further classified into six categories based on their genomic arrangement with the neighboring protein-coding genes. Previous studies have shown that lncRNAs from divergent promoters are associated with active transcription of the neighboring protein-coding gene *in cis*[79]. From our observation, we found that 43.6% (n=98) of total active lncCARs were arranged in head-to-head (XH) fashion with the neighboring protein-coding genes (PCGs). These neighboring protein-coding genes are found within 2kb distance to the active XH lncCARs (**Figure 16b-c**). Gene enrichment analysis of neighboring protein-coding genes showed enrichment of transcription related terms. Arrangement of these active XH lncCARs in a head-to-head pattern with PCGs in close proximity was termed as active XH transcriptional units or divergent transcription units.

**Active XH lncCAR targeting divergent transcriptional units**

To understand the mechanism by which these active lncCARs are targeted to this divergent transcription unit, the Chromatin Oligo affinity Purification experiment followed by qPCR was performed. We found enriched active XH lncCARs over the promoters of corresponding PCG targets but not at the active XH lncCAR promoters. This suggests that active lncCARs are targeted to their partner PCGs to maintain active transcription at the divergent transcriptional units. Moreover, we also observed that upon the downregulation of active XH lncCARs, the expression of corresponding partner PCGs goes down.

**Active XH lncCAR maintains H3K4me2 and establishes H3K4me3 at XH transcription units**

The WDR5 occupancy and H3k4me2/3 enrichment at the promoters of partner PCGs are decreased upon the downregulation of active XH lncCARs. We then examined whether maintenance of H3K4me2 and H3K4me3 at the divergent transcription units were WDR5 dependent. To this extent, Chromatin Immunoprecipitation (ChIP) was performed using H3K4me2, and H3K4me3 pulldown in control cells and WDR5 depleted cells. We found that WDR5 depletion only affects H3K4me3 enrichment but not H3K4me2 over divergent transcription units. This shows that divergent transcription units maintain H3K4me2 at these promoters independent of WDR5, but the establishment of H3K4me3 through the conversion of di- to tri-methylation is WDR5 dependent (**Figure 16d**).

## Significance and conclusion

Our study explores the mechanism of active chromatin-associated lncRNAs at the divergent transcriptional units. This provides a new perspective on how genomic organization can determine the functions of lncRNAs in a global scale. It is a unique observation showing most of the active chromatin-associated lncRNAs forms divergent transcription unit genome-wide. This unique pattern was validated by comparing the genomic arrangements of previously published inactive or repressive chromatin-associated lncRNAs[91]. Our study gives more insights into how active chromatin-associated lncRNAs can regulate their target genes by establishing and maintaining the chromatin state.

## 3.3 Sperm lincRNAs in development and cancer (paper III)

**Sperm specific long intergenic noncoding RNAs (Sp-lincRNAs)**

Sperm was considered merely as a passive vehicle delivering the paternal genome to the oocyte during fertilization. An oocyte is widely known to have an active transcription profile, but not the sperm. A recent study using intracytoplasmic sperm injection (ICSI), showed the ability of sperm RNAs in rescuing defective blastocyst embryos generated from *Dicer* knockout germ cells[190]. In addition to that, caput sperm (immature sperm) derived defective post-implantation embryos were rescued by microinjection of cauda sperm-specific (mature sperm) small RNAs[127,128]. Though there are a handful of studies showing the significance of sperm derived RNAs[122-125], it is important to explore other RNA molecules in sperm and on a genome-wide scale. Our study using transcriptome-wide analysis, found some long intergenic noncoding RNAs (lincRNAs) that are specifically expressed in sperm (Sp-lincRNAs) by utilizing high-throughput RNA sequencing dataset from sperm and oocyte[175]. There were other classes of lincRNAs, such as oocyte-specific (Oc-lincRNAs) and commonly expressed lincRNAs in sperm and oocyte (SpOc-lincRNAs) (**Figure 17a**). Similarly, we found sperm-specific protein-coding genes (Sp-PCGs), oocyte-specific PCGs (Oc-PCGs), and common PCGs (SpOc-PCGs). Previous studies have highlighted the importance of oocyte expressed RNAs during development. It is important to explore the possible roles of sperm-specific lincRNAs (Sp-lincRNAs) during development.

**Sp-lincRNAs during zygotic genome activation**

Zygotic genome activation (ZGA) and maternal transcript degradation (MTD) are the two key stages that define maternal to zygote transition during the preimplantation stages of embryo[191,192]. Further, we analyzed single-cell RNA-seq samples from preimplantation embryo development. This revealed specific activation of Sp-lincRNAs from the four-cell stage until late blastocyst. This is the stage where the zygotic genome gets activated to initiate zygotic transcription, which is important for embryo development (**Figure 17b**). On the other hand, SpOc-lincRNAs start declining after the four-cell stage of

embryo coinciding with the commencing of zygotic genome activation (ZGA) (**Figure 17b**). Earlier, studies have shown that maternal (oocyte) transcripts and proteins get degraded prior to the onset of ZGA[193]. However, we found that both paternal or sperm and maternal or oocyte transcripts (SpOc) gets degraded before zygotic genome activation (ZGA). Thus, our evidence shows the importance of sperm derived lincRNAs and their coordination during preimplantation embryo development.

## Chromatin and transcriptional dynamics of Sp-lincRNAs during development

During early development, paternal chromatin patterns are known to modulate transcriptional profiles of preimplantation embryos[49,52-54,194]. Next, we sought to investigate the sperm chromatin patterns at the promoters of these sperm derived transcripts (Sp and SpOc). Interestingly, we found three distinct clusters of chromatin patterns at the Sp- and SpOc-lincRNA promoters such as high levels of H3K4me3 (high-K4, absence of H3K27me3), low levels of H3K4me3 (low-K4, absence of H3K27me3) and a third cluster without both chromatin marks ($K4^-K27^-$, absence of both H3K27me3 and H3K27me3). Sp- and SpOc-PCG promoters were having one additional cluster enriched with both H3K4me3 (high) and low H3K27me3 (bivalent promoters). Expression patterns of the sperm derived transcripts were well correlated with their chromatin patterns at their promoters. The lincRNAs from high-K4 clusters had higher expression levels in sperm compared to low-K4, bivalent, and $K4^-K27^-$ clusters. These chromatin structures at Sp-lincRNAs were lost in post-implantation stages of an embryo in three germ layers (ectoderm, mesoderm, and endoderm) and germ layer derived somatic tissues (**Figure 17c**). Also, correlating with the chromatin patterns Sp-lincRNAs from all three sperm derived chromatin clusters showed no expression in germ layers and somatic tissues (**Figure 17d**). In Sp-PCG promoters, the bivalent patterns were maintained in germ layers as well as in somatic tissues. These Sp-PCGs from bivalent clusters had higher expression in germ layers and somatic tissues compared to PCGs from high-K4 and low-K4 clusters. This corroborates with the role of bivalent domains in lineage commitment during development[30,31,195]. Overall, Sp-lincRNAs carry distinct chromatin structures in sperm correlating with the expression. Importantly, Sp-lincRNAs are active during ZGA, whereas SpOc-lincRNAs are active pre-ZGA and declines at the onset of ZGA. Moreover, Sp-lincRNAs show no expression with a loss of chromatin structures in germ layers and somatic tissues (**Figure 17d**).

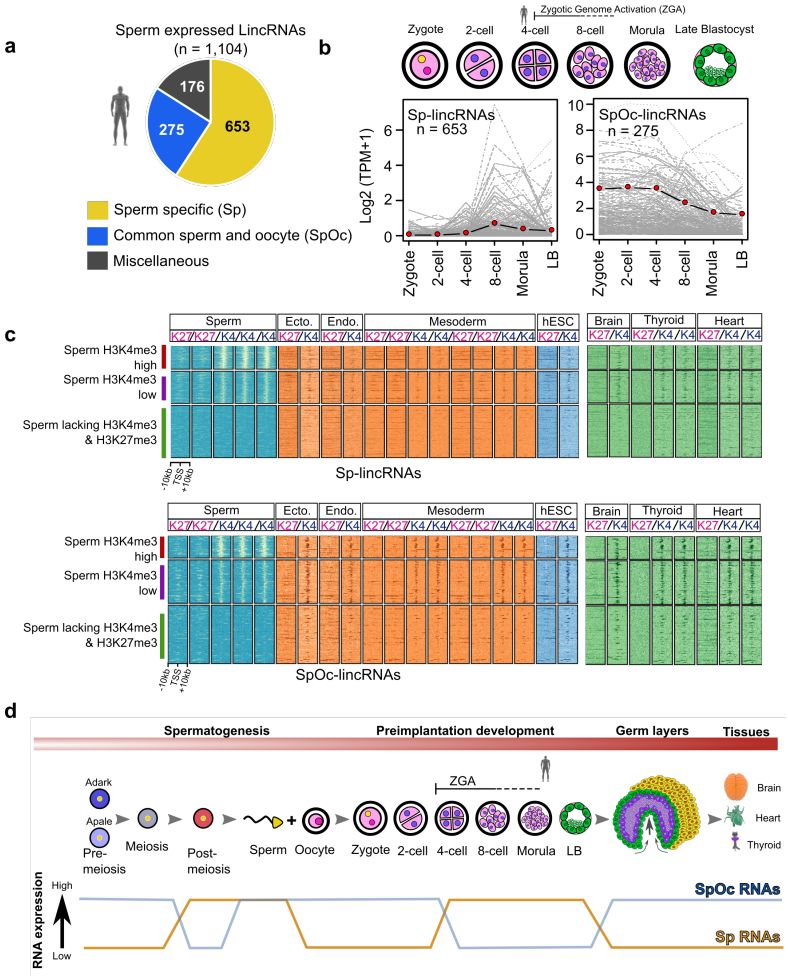


***Figure 17. a)*** *Ven diagram showing the number of sperm-specific (Sp) and commonly expressed (sperm and oocyte: SpOc) lincRNAs. Miscellaneous represents lincRNAs that are inconsistently expressed between replicate samples of gametes (not considered for further analysis).* ***b)*** *Temporal stage-specific expression patterns of Sp- and SpOc-lincRNAs.* ***c)*** *Enrichment of H3K27me3 (K27) and H3K4me3 (K4) ChIP-seq signals from the three germ layers, human embryonic stem cells (hESC), and the germ-layer-derived somatic tissues (brain, thyroid, and heart) over the promoters of Sp- and SpOc-lincRNAs (extended ±10 kb from TSS) from sperm-derived chromatin clusters.* ***d)*** *Model depicting observed expression patterns of sperm-specific and sperm-oocyte (SpOc) transcripts in spermatogenesis, gametes, preimplantation embryos, germ layers, and somatic tissues.*

## Aberrant activation of Sp-lincRNAs in tumors

Previous studies have shown the significance of RNA molecules from embryo development and spermatogenesis (cancer-testis antigens) in cancer progression[196-200]. Our study also found that sperm derived transcripts show testis-specific expression with the accumulation of transcripts in round spermatid during spermatogenesis. To explore the role of sperm derived transcripts, we utilized the RNA-seq dataset of 26 cancer types from TCGA. Strikingly, Sp-lincRNAs were showing aberrant deregulation in multiple cancer types (**Figure 18**). In addition to that, the extent of deregulation was correlating with the levels of H3K4me3 at the promoters such as lincRNAs from sperm derived high-K4 groups had higher deregulation compared to lincRNAs from low-K4 and K4$^{-}$K27$^{-}$ groups. Downregulation of top Sp-lincRNAs had strong effects on cancer cell hallmarks such as cell cycle, cell proliferation, and apoptosis.

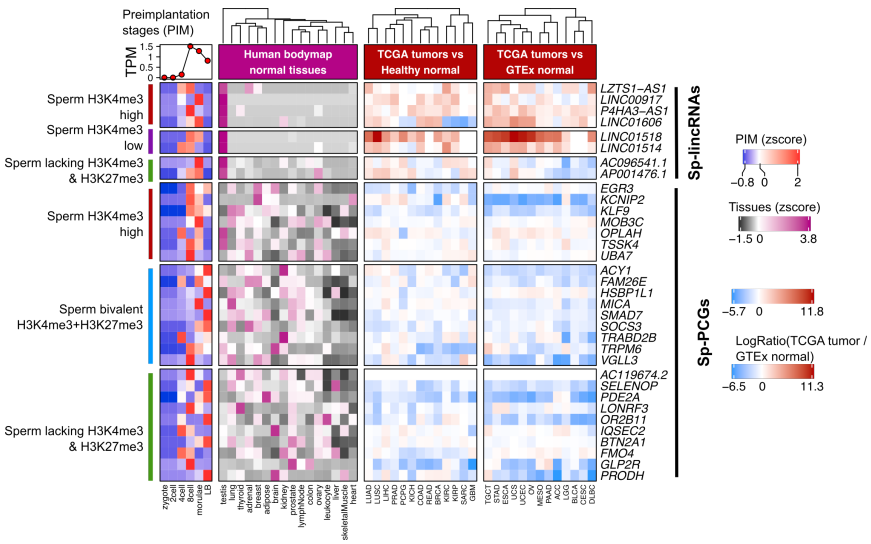


***Figure 18.*** *Status of sperm-specific transcripts (lincRNAs and PCGs) in germ cells, preimplantation stage embryos, the human body map tissues, TCGA tumors compared with the corresponding healthy samples, or TCGA tumors compared with GTEx health samples. Z score in the plots is derived from the normalized TPM expression values. The log fold change is calculated by comparing the expression of tumors with the health samples expression.*

## Significance and conclusion

This study gives a unique perspective on the importance of paternal or sperm derived transcripts during development. Additionally, we found Sp-lincRNA transcripts to be aberrantly activated in various cancers. In summary, this study explores the role of Sp-lincRNAs during development and in cancer. This list of Sp-lncRNAs has to undergo thorough pre-clinical or clinical investigations and also there is a need to perform more functional studies.

**Chromatin and transcriptome-based integrative approaches to profile functional lncRNAs**

# 4 METHODOLOGICAL APPROACHES

## 4.1 Available public resources

### 4.1.1 Sequence and knowledge-based repositories

***Ensembl***: This is an extensive and very comprehensive repository used by most of the computational biologists to get reference genome sequence (chromosome-wise) in FASTA format, gene annotations in GTF, or GFF format and genomic variants in VCF format[201]. FASTA file contains whole genome nucleotide sequences of a particular organism of interest. The GTF (Gene Transfer File) files contain names of genes (protein-coding and noncoding), gene symbols, gene isoform, exons numbers, gene orientation (sense or antisense to the direction of a genome) and genomic coordinates of all features. In our study, we used FASTA files and GTF files of human and mouse from Ensembl. This repository is more reliable compared to other available repositories since this serves as a basic resource to build their database.

***Gencode***: GENCODE servers as a specialized repository that stores and reannotates information from Ensembl[202]. This database improves the accuracy of annotation by verifying with available biological evidence. We were interested in this database since it has separate versions of annotation files (GTF or GFF) for protein-coding genes and long noncoding RNAs (lncRNAs). This database also hosts gene- or transcript-wise nucleotide sequences of lncRNAs and protein-coding genes. One more advantage of this database is that it provides statistics on different classes of genes or transcripts. This database is a useful resource for those who are exclusively working on lncRNAs.

***UCSC browser***: UCSC is a huge bioinformatic resource-based database that hosts genome sequences of many species[203]. The integrated genome browser in UCSC allows us to explore genes and genomic locations in an in-depth resolution (single nucleotide). It also allows users to load default or upload

their own tracks against the reference genome for comparative analysis. For example, uploading tracks from ChIP-seq visually gives information on the status of chromatin states or enrichment of transcription factors at that particular gene location.

***GeneCards***: This is an interesting web resource where one can find detailed information about a particular gene or transcript of interest[204]. GeneCards compiles information from Ensembl, HGNC, UniProt, gene ontology, Rfam, and more. All the information about the genes is properly referenced to their corresponding scientific publications and other reliable sources.

***Cistrome***: Genome-wide chromatin states and the transcription factor binding can be measure using Chromatin Immunoprecipitation (ChIP) technique followed by sequencing. Cistrome has a collection of analyzed ChIP-seq datasets from different species, cell lines, and somatic tissues[205]. This repository is updated frequently with the datasets from recent publications and uses a recent computational pipeline to analyze the data.

***GTEx portal***: GTEx (Genotype-Tissue Expression) provides access to completely analyzed and processed datasets such as tissue-specific expression, variants, and histology images of tissues (https://gtexportal.org). It consists of 54 different tissue types from 948 donors.

### 4.1.2 Repositories for computational tools

***Bioconductor***: There is a huge number of open-source tools available related to computational biology in the Bioconductor repository (https://bioconductor.org). Most of these tools are related to gene annotation, sequence alignment, read quantification, differential expression, motif analysis, gene enrichment analysis, basic and advance statistics. Additionally, Bioconductor also packages advanced data analysis, data processing, and visualization tools. In this thesis, we majorly utilized Bioconductor for differential expression and other statistical analysis.

***GitHub****:* One of the most useful tool development platforms for the developers is GitHub (https://github.com). This is a platform where you can interact and collaborate easily with the tool developers. For this thesis, we also used some of the visualization tools for plotting and cluster-based analysis from GitHub.

***UCSC tools***: Apart from hosting genome sequences, UCSC also provides useful standalone command-line tools to indexing, formatting, and manipulating genomic sequences (http://hgdownload.soe.ucsc.edu/admin/exe/). In this thesis, we used UCSC tools for indexing genome before alignments, converting genome into various formats.

***Ensembl tools***: In addition to UCSC, Ensembl also hosts genomic sequences and gene annotations. Additionally, Ensembl also hosts servers for sequence alignments, gene annotations, and functional predictions. Ensembl standalone tools such as BioMart, BLAST, and Variant Effect Predictor (VEP) are available for a wide range of species. This thesis mostly utilized Ensembl BioMart to extract gene or RNA specific annotations and sequences. For example, to retrieve information about particular lncRNA such as the number of gene isoforms, isoform-wise exon information.

### 4.1.3 High-throughput sequencing repositories

***ENCODE***: Encyclopedia of DNA Elements (ENCODE) is an extensive project where the member of the consortium generates, validates, and submit their sequencing data from different experimental assays[206]. The major assay types, including genome-wide sequencing (ChIP, DNase, eCLIP, WGBS, ATAC, RRBS, HiC, ChIA-PET, Repli-chip, etc.) and transcriptome (RNA-seq, Repli-seq, RIP-seq, CRISPRi, etc.) are deposited as raw sequencing file (FASTQ) as well as analyzed data. Individual samples are prepared according to ENCODE protocol, quality checked after sequencing, analyzed according to ENCODE computational pipelines, and deposited with complete information along with the corresponding control samples. Our thesis utilized raw sequencing datasets of RNA-seq and ChIP-seq samples from ENCODE.

***GEO and ENA***: These are service-based repositories from NCBI and EBI, where one can deposit and distribute the sequencing dataset from their study. GEO (https://www.ncbi.nlm.nih.gov/geo/) uses ENA (European Nucleotide Archive, https://www.ebi.ac.uk/ena/browser/) cloud-based servers to deposit sequencing dataset. Other scientists can easily access and download most of the datasets except few controlled datasets needs permission to access.

***Genomic Data Commons (GDC)***: This stores huge datasets from various consortiums such as TCGA (The Cancer Genome Atlas) and TARGET

(pediatric cancers). GDC provides sequencing datasets from WGS-seq (genome and methylome), RNA-seq (transcriptome) and WES-seq (exome) assays. TCGA covers sequencing datasets from 26 somatic tissue cancer types with their corresponding healthy somatic tissues. TARGET database contains sequencing data from pediatric and rare cancers such as Neuroblastoma, Wilms, Rhabdoin, ALL, AML, and osteosarcoma tumors. The raw sequencing data from ICGC has controlled access where one needs to apply for permission to get access.

***COSMIC***: Catalogue Of Somatic Mutations In Cancer gives information on mutations (CNV, genomic rearrangements, and fusion) in cancer patients compared with normal tissues[207]. In addition to that, the COSMIC also distributes analyzed gene expression and DNA methylation data from TCGA[208].

***Human Body Map***: In addition to GTEx, Human Body Map 2.0 project also provides a tissue-specific sequencing dataset. This consists of 16 human tissue types from various sites (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30611).

## 4.2 High-throughput sequencing analysis

Extracted RNA or DNA from the experiment is sent for sequencing. The raw sequencing samples consist of reads ranging from 36 bp to 300 bp in length. Usually, sequencing reads come as FASTQ files with four lines per fragment. This includes a line with the name of the read generated by the sequencing platforms, nucleotide sequence, followed by '+' separator, and the base quality values per nucleotide base. For any high-throughput sequencing from the Illumina platform, follows the above-described format. Unlike Illumina, ABI SOLiD varies in its raw format providing color-space FASTA files (csfasta), which needs one extra step to convert color-space reading into FASTQ format.

### 4.2.1 Transcriptome analysis

***Pre-filtering***: Analysis of transcriptome (RNA-seq or ChRIP-seq) starts by evaluating sequencing quality using the FastQC tool. This gives us information on whether the sequencing files are of good or bad quality. FastQC evaluates FASTQ files by summarizing per nucleotide base, per read quality score, GC content, distribution of read length, sequence duplication, overrepresented sequences, adapter contents, and kmer content. Based on the FastQC evaluation, if the reads contain overrepresented sequences, adapter sequences or low-quality bases at the end of the reads, the quality control steps were further extended. Cutadapt or Trimmomatic[209] was used to remove adapter sequences and overrepresented from the reads. Samples were considered for further downstream processing only if it passes the FastQC quality check. Some samples do not pass the quality check because of contamination or technical issues during library preparation.

***Alignment and quantification***: Next, the samples (FASTQ) were subjected to splice-aware alignment using HISAT2 aligner[210] (Illumina) or Lifescope 2.5 (ABI SOLiD). For this purpose, the reference genome hg19/GRCh37 or hg38/GRCh38 was downloaded from Ensembl in FASTA format and then indexed using the HISAT genome build utility or UCSC indexing tools. SAM or BAM (Sequence/Binary Alignment Map format) files are generated after

aligning to the reference genome. This BAM file contains information about genomic coordinates for individual reads mapped to the reference and their mapping quality score (MAPQ). Next, we used gene annotation (GTF) files from Ensembl (protein-coding + noncoding) or GENCODE (lncRNAs) to quantify the number of reads mapped to individual gene. It is very important to use gene annotation files that correspond to the version of the reference genome. We used featureCounts[211] from the Subread package or HTSeq-count[212] for sample-wise quantification of reads per gene. Depending on whether the sequenced samples are prepared with single/paired-end and stranded/non-stranded library preparation protocol, the parameters were adjusted on featureCounts and HTSeq-count tools. We have assigned reads to genes only if the mapping quality (MAPQ) is greater than 30.

***Normalization, time-series, and enrichment analysis***: Observed reads per gene must be normalized to library size, sequencing depth, and gene length (summed length of exons). There are different normalization approaches such as TPM (Transcripts Per Million), RPKM for single-end (Reads Per Kilobase of transcript per Million reads mapped), FPKM for paired-end (Fragments Per Kilobase of transcript per Million reads mapped) and RPM/CPM (Reads/Counts Per Million). We subjected normalized RPKM values to STEM (Short Time-series Expression Miner)[213] analysis to find temporally expressed genes or lncRNAs in different time points. For example, in **paper I** of this thesis, the RPKM values of individual lncRNA at three S-phase time points (T1, T2, and T3) were compared against unsynchronized control samples (assumed as base time point). STEM analysis resulted in significant temporal clusters across S-phase. Similarly, in **paper II**, we compared normalized values from H3K4me3 and WDR5 ChRIP-seq samples with input (control) samples to find H3K4me3-WDR5 (active chromatin) enriched transcripts. In **paper III**, the TPM values were used to derive stage-specific expression in spermatogenesis, preimplantation development, embryonic germ layers, and somatic tissues. Additionally, a comparison between cancer patients and healthy normal were also made using TPM normalized read counts.

***Differential expression analysis***: Some of the transcriptome sequencing samples were subjected to differential expression analysis. Only genes having at least 1 CPM (Counts Per Million) in at least two samples of each comparison group (control and treated) were used for DE analysis. We utilized advanced statistical tool such as EdgeR[214] from Bioconductor to find significant differentially expressed transcripts. The edgeR tool is based on the empirical Bayes statistical approach, and it can very well handle samples from experiments having a smaller number of biological replicates. This method was implemented in **paper I** to find differentially expressed genes between *SCAT7*-

si knockdown and control-siRNA samples. The significant candidates were selected using absolute log-fold change > 1 and adjusted p-value or FDR < 0.05. Since the sample size in individual groups (patient vs. healthy normal) are larger in TCGA datasets, we used a different approach for differential expression analysis. The RPKM normalized values of TCGA cancer patients and normal samples were used to find differentially expressed lncRNAs. For this purpose, we applied the Wilcoxon signed-rank nonparametric test to calculate p-values. Next, multiple testing correction was done by calculating adjusted p-values or FDR using Benjamini-Hochberg's statistical method[215].

### 4.2.2 Epigenome analysis

***Pre-filtering***: Whole genome sequencing also follows the same pre-filtering steps, as described in **section 4.3**.

***Alignment and post-filtering***: Genome sequencing deals with the whole genome rather than transcripts or transcription. Hence need for specialized splice-aware aligners like in transcriptome sequencing is not necessary. Obtained filtered or cleaned reads were subjected to genome alignment using BWA[216] or Bowtie[217]. The reads were aligned against reference genome hg19/GRCh37 or hg38/GRCh38 from Ensembl. Aligned SAM file was converted into BAM with the help of SAMtools[218]. Since genome sequencing does not involve read qualification, we must retain only high-quality reads using post-filtering steps. During post-filtering steps, we remove duplicate reads using Picard MarkDuplicates (http://broadinstitute.github.io/picard). Because we were dealing with the whole genome, it is important to remove reads from overrepresented regions of the genome called blacklisted regions using BEDTools '*intersect*'[219]. The reads with high mapping quality (MAPQ>30) were selected using '*bamutils*' from NGSUtils package[220].

***Differential enrichment/binding and clustering***: After post-processing, the mapped high-quality reads were used for further downstream analysis. Chromatin Immunoprecipitation sequencing (ChIP-seq) experiments were analyzed in both **paper II** and **paper III**. This experiment involves crosslinking proteins to DNA and pulldown using the antibody against transcription factor (TF) or histone protein of interest. Next, the TF or histone proteins are washed away to retain only DNA regions bound by the TF protein or histones of interest. ChRIP-seq (**paper II**) is different from ChIP-seq because, in ChRIP-seq, we isolate RNA bound to chromatin and remove DNA

with DNase treatment. This thesis primarily deals with H3K4me2 (**paper II**), H3K4me3 (**paper II** and **paper III**), H3K27me3 (**paper III**), and WDR5 (**paper II**) ChIP-seq analysis. The corresponding controls used in this analysis are input DNA sample (crosslinked but not immunoprecipitated). Firstly, the quality of enrichment and efficiency of antibody pulldown was evaluated using '*plotFingerprint*' utility from deepTools package[221] by comparing aligned reads from ChIP-seq pulldown and input DNA. Only samples with better enrichment profiles were considered for further downstream processing. Differential enrichment or binding was done using '*bamCompare*' from deepTools. Using '*bamCompare*', we derived the base-by-base resolution of $\log_2$ ratio between ChIP-seq pulldown and input DNA samples to find enriched regions genome-wide. The output from differential enrichment or binding is usually location of differentially enriched peaks. These localized peaks or the entire enrichment profiles are visualized/plotted over the promoters of genes or lncRNAs (**paper II** and **paper III**) using '*plotProfile*' or '*plotHeatmap*' from deepTools. Additionally, in **paper III**, we used k-means clustering (deepTools) to find combinatorial clusters between H3K4me3 and H3K27me3 over Sp and SpOc lncRNA/PCG promoters.

### 4.2.3 Sequence and proximity-based analysis

***Protein-coding potential***: To test the protein-coding capability of obtained lncRNA candidates, we used the CPC (Coding Potential Calculator) tool[222]. CPC evaluates coding potential based on the length and quality of open reading frames (ORF) in the transcript sequence. For this purpose, we extracted nucleotide sequences of transcripts or lncRNAs in FASTA format with the help of Ensembl BioMart batch analysis. By providing the FASTA sequence, CPC outputs scores for individual transcript sequences. Transcripts do not code for any peptide or protein if the CPC score is less than 0.37.

***Motif analysis***: In addition to histone modifications, transcription factors (TFs) also plays an important role in regulating gene expression. Specifically, during early embryo development, there are many transcription factors that help in regulating genes responsible for the zygote to embryo transition. Therefore, the sperm derived transcripts from **paper III** were subjected to promoter motif analysis to find TFs associated with these transcripts. For this purpose, nucleotide sequences of sperm derived transcript promoter (± 250bp from transcription start site) were extracted using '*faidx*' from SAMtools. This was done by providing gene coordinates from GTF (Ensembl) file and the reference

genome FASTA sequence (Ensembl) to '*faidx*'. HOMER[223] has a collection of consensus motif sequences database derived from available published ChIP-seq datasets of human and mouse. We used '*findMotifs.pl*' utility from the HOMER package to predict consensus transcription factor (TF) motifs enriched at the promoters of sperm derived transcripts. Predicted motifs having a p-value of less than 0.01 and found in the highest number of target gene/transcript promoters were considered to be significantly enriched.

***Genomic position and proximity-based analysis***: LncRNAs can be categorized based on their genomic positions. For example, 1) a lncRNA can be in opposite orientation in the genome or to the neighboring protein-coding gene; 2) a lncRNA can be intronic when it is present between two exons of other transcripts, and 3) a lncRNA can be intergenic when it is not present near to other transcripts or between two other transcripts. In our **paper II**, we observed that a significant number of active (H3K4me3 and WDR5 enriched) lncCARs were arranged in antisense orientation, and rest were in the sense orientation. We also found that most of the active lncRNAs were having protein-coding partners within 2kb distance. Based on these observations, we further classified these antisense active lncCARs into four sub-categories based on their neighboring protein-coding gene partners. To this extent, we found the following patterns: XH, head-to-head divergent lncRNA; XT, lncRNA arranged in tail-to-tail fashion with protein-coding gene; XI, lncRNA-Inside (lncRNA present within the protein-coding gene in the opposite direction); XO, lncRNA-Out (protein-coding gene present within lncRNA in the opposite direction). The above-described analysis involves information extracted from GENCODE (lncRNAs) and Ensembl (protein-coding genes) GTF files. The coordinates, direction of transcripts, and distance between transcripts were used to determine the patterns of arrangement. For this purpose, we used utilities from BEDTools integrated with our in-house written scripts.

### 4.2.4 Knowledge-based analysis

***Importance of knowledge-based analysis***: Differential expression or differential binding analysis provides us with a significant number of candidate protein-coding genes. There is a need to extract biologically meaningful information from the obtained candidates. There are specialized databases developed to store up-to-date information about protein-coding genes based on the evidence from the scientific publications and experimental validations. Information from these databases are retrieved and utilized by specialized tools. These tools compare our candidate genes against the databases and

extract consensus enriched functions among this users-provided list of candidates with a proper statistical inference. This whole process is done computationally, and it is termed as functional or gene enrichment analysis. With this analysis, one can predict significantly enriched biological processes or pathways in their experiments.

***Why in-house tool*?** However, there is a major problem associated with most of the available tools for gene enrichment analysis. Sometimes these tools are not up to date with their database. Because of their outdated resource, there is a possibility that the results from these tools can mislead the direction of the research[224]. Lack of updates are also found in the frequently used enrichment analysis tools such as DAVID and Gorilla. It was found that these tools were missing out on a significant number of genes from key pathways such as Cell Cycle, Chromatin organization, DNA repair, and Apoptosis. For this purpose, we developed our own tool to perform gene enrichment analysis called Gene Set Clustering based on Functional annotation (GeneSCF)[225]. To overcome problems associated with database update, we designed the tool in a way that it performs enrichment analysis by connecting to the multiple databases in real-time. We also made this tool in such a way that the user has the freedom to choose when to update the database.

***Database access and retrieval***: GeneSCF uses REST API and FTP/HTTP protocols to access and retrieve data from various databases such as Geneontology[226], KEGG[227], Reactome[228], and Network of Cancer Genes (NCG)[229]. This tool also uses the NCBI Entrez gene database to retrieve information about the user-provided candidate genes. GeneSCF does extensive statistical analysis to rank significantly enriched biological processes or pathways.

***Usage in the thesis***: GeneSCF was used in all three studies (**paper I**, **paper II**, **paper III**) in the following contexts, 1) to predict functions of neighboring protein-coding genes to the lncRNA candidates (**paper I** and **paper II**); 2) to predict functions of differentially expressed protein-coding genes between knockdown and control samples (**paper II**), and 3) to find functions associated with sperm derived protein-coding genes (Sp- and SpOc-PCGs) (**paper III**).

***URL access**:* GeneSCF tool can be accessed via, http://genescf.kandurilab.org and https://github.com/genescf/GeneSCF

### 4.2.5 Other datasets used

Some of the RNA-seq, ChIP-seq, and Methyl-seq raw and processed data used in this thesis were obtained from public and controlled repositories such as ENCODE, GEO, ENA, dbGAP, ICGC/TCGA, COSMIC, GTEx and Human Body Map 2.0 project.

**Chromatin and transcriptome-based integrative approaches to profile functional lncRNAs**

# 5 SUMMARY AND FUTURE DIRECTIONS

This thesis used various computational approaches to identify lncRNAs and explore their involvement in diverse cellular functions. However, we found these lncRNAs to act differently in different biological model systems. For example, the oncogenic S-phase lncRNA *SCAT7* interacts and recruits the hnRNPK-YBX1 RNP protein complex to the promoter of FGFR members. This targeting activates the FGF/FGFR signaling pathways, which in turn activates downstream oncogenic pathways and promotes tumor progression. This thesis also identified some of the lncRNAs to be associated with the active chromatin (active lncCARs) compartment. These lncRNAs regulate the expression of neighboring transcription factor genes in breast cancer cells. Most of these active lncCARs are exhibit a divergent head-to-head pattern (XH) with adjacent protein-coding genes (active XH lncCARs). We also unraveled the mechanism by which these active XH lncCARs target and regulate the transcription of neighboring protein-coding genes. This active XH lncCARs maintains H3K4me2 at the divergent transcription units in a WDR5-independent manner. Moreover, the active XH lncCARs interact with and recruit WDR5 to these divergent transcriptional units to establish H3K4me3 by catalyzing the tri-methylation of di-methylated lysine residues. In addition to that, this thesis has taken a challenge to investigate the previously unexplored lncRNA molecules in mammalian sperms. There was a significant number of lncRNAs found to be active in the sperm compared to the oocyte. Moreover, the expression of these sperm-derived lncRNAs demonstrated a significant correlation with their chromatin patterns in sperm. Additionally, these sperm lncRNAs represent an integral component of the transcriptional machinery involved in zygotic genome activation (ZGA). Nevertheless, they also harbor oncogenic properties in multiple cancers.

All the above-mentioned observations demonstrate the uttermost necessity to explore the detailed molecular mechanisms of individual lncRNAs in various biological contexts. The significance of these lncRNAs in different biological systems proves their potential implication in therapeutic interventions for an effective cancer treatment strategy[230]. For instance, antisense oligonucleotides (ASOs) form RNA-DNA structure with target RNA molecule and activate RNase-H-mediated RNA degradation. Using an ASOs-based approach, recent studies demonstrated the efficacy of targeting lncRNAs to inhibit tumor

growth metastasis[231]. Moreover, lncRNAs interact with various chromatin remodelers, such as HATs and HDACs, to modulate chromatin configurations. Currently, several clinically approved inhibitors of HATs and HDACs are implicated in cancer treatments[232,233]. However, this approach has some barriers and limitations because of the lack of a detailed understanding of molecular mechanisms controlling the affected oncogenic pathways[232]. Using combinatorial approaches by targeting both lncRNA molecules and chromatin remodelers may fine-tune the future treatment strategies.

# ACKNOWLEDGEMENTS

First and foremost, my gratitude to **Prof. Chandrasekhar Kanduri** and the **University of Gothenburg** for believing and providing me this wonderful opportunity to pursue my Ph.D. I am very grateful to my co-supervisors, **Tanmoy** and **Erik**, for guiding me in my difficult situations. My special thanks to **Tanmoy** and **Kankadeb** for fruitful discussions from which I have gained immense knowledge about the field.

Special gratitude to my dad (**Subhash**) and mom (**Mangalam**) for being on my side during difficult times and always encouraging me to reach to this position. Thanks to my family members **Kanthilal** (brother), **Megha** (sister-in-law), **Sarvesh** (nephew), and **Shivani** (niece), **Gnanadev** (uncle), **Lakshmi** (aunt), **Ajith** (brother), **Manisha** (sister), **Tanaji** (father-in-law), **Jayashree** (mother-in-law) and **Dushyant (Babu)** (brother-in-law) for their love and support.

Particularly I want to thank my loving wife **Dakshita**, for being with me in crucial times. You gave me immense strength and encouragement while preparing for dissertation which is beyond words.

Thanks to former and present Kanduri's lab members **Gaurav**, **Matthieu**, **Roshan**, **Sanhita**, **Abiarchana**, **Alva**, **Luisa**, **Subazini**, **Prasanna**, **Vijay**, **Silke**, **Sara**, **Tanushree**, **Sagar**, **Daniel**, **Mirco** for all those wonderful memories. I highly appreciate **Silke** for being very kind and offering vegan candies specially from Germany. Also, thanks to **Kankadeb**, **Tanushree** and **Sagar** for always coming to bachelor's movie nights. **Tanushree**, though you always slept after 15 minutes of movie, you tried your best. I feel sorry and also, I must thank you guys for handling all my nonsense.

**Caroline**, **Lily**, **Daniel**, and **Silke** you all are wonderful guys I have ever seen. You are literally the sweetest persons for making those superb delicious cakes.

I appreciate **Meena's lab** and **Erik's lab** for all your support and discussions. It helped in growing conceptually.

I will never forget the Europe trip with you guys, **Prakash**, **Ram Mohan**, **Jothi**, **Dinesh (Germany)**, **Juju (Anbu)**, **Vignesh** and **Prashanth**. Those tales are everlasting and remain legendary. I would like to thank weekend buddies, **Prakash**, **Krithika** and **Madhavan (Maddy)** for all the dinner, movie and game nights. Also, I thank you guys for your valuable opinion and suggestions

while facing uncertain situations. My special thanks to **Prakash** for organizing trips and cooking awesome food (you are the best chef). I am also grateful to **Ranjith** for being such a great pal. We had wholesome time together for movies and making delicious food.

**Nandha**, the first person whom I met in Sweden, took me in when I was struggling to find accommodation during my initial days. I can never forget those difficult times we faced together. We both somehow managed to overcome all those problems and had loving memories together. You were always kind to me and being helpful.

I am extremely thankful to my school friend **Kandhan**, who trusted and followed me till Sweden. You were always there with me while facing critical problems during later stages and you have always tried your best to solve them. I wish you fulfill your dreams and I want you to achieve greater things in life.

Thanks to **Dhasarathi**, **Ram Mohan** and **Sarathy** for being very accommodative and always arranging dinner, birthday, and surprise parties. Also, it wasn't fun without our whole gang, **Kandhan**, **Lokesh**, **Karthi**, **Senthil**, **Santhosh**, **Gaddy**, **Arjun**, **Prem**, **Ranjith**, **Gopal**, **Raghu**, **Danny**, **Venki boss**, **Balaji**, **Abhilash**, **Badri** and **Dinesh**. I thank **Ranga**, **Sampath**, **Naveen** and **Vijay** for all our intellectual and philosophical discussions.

I am very lucky to have you as my course buddies, **Sukanya**, **Saber**, **Sampath** (AP), **Fahim**, **Peidi**, **Sumaiya** and **Tabassum** and I thank you for all our "so-called" group studies. Special thanks to **Saber** for being a lovely host for all our movie nights, game nights, dinner and birthday parties.

Special thanks to **Gendy**, **Lily**, **Tanmoy**, **Amita, Meeta**, **Madhavan (Maddy)** and **Dakshita** for helping me with your suggestions while writing this thesis.

Thanks to **Connie** and **Carina** from the administration for always helping me with the administrative work on right time.

I thank Uppsala Multidisciplinary Center for Advanced Computational Science (**UPPMAX**) high-performance computing (HPC) which is part of Swedish National Infrastructure for Computing (**SNIC**) for providing extensive clusters to analyze larger datasets. My thesis would not have been possible without your valuable resource.

At last, I would like to thank **Sheldon** for entertaining and inspiring me whenever I was feeling low. I literally watched your episodes every day.

# REFERENCES


1. Nagy, J.A., Chang, S.H., Dvorak, A.M. & Dvorak, H.F. Why are tumour blood vessels abnormal and why is it important to know? *Br J Cancer* **100**, 865-869 (2009).
2. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674 (2011).
3. Fouad, Y.A. & Aanei, C. Revisiting the hallmarks of cancer. *Am J Cancer Res* **7**, 1016-1036 (2017).
4. Barnum, K.J. & O'Connell, M.J. Cell cycle regulation by checkpoints. *Methods Mol Biol* **1170**, 29-40 (2014).
5. Collins, K., Jacks, T. & Pavletich, N.P. The cell cycle and cancer. *Proc Natl Acad Sci U S A* **94**, 2776-2778 (1997).
6. Cicenas, J. & Valius, M. The CDK inhibitors in cancer research and therapy. *J Cancer Res Clin Oncol* **137**, 1409-1418 (2011).
7. Liu, S.J.*, et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**(2017).
8. Kitagawa, M., Kitagawa, K., Kotake, Y., Niida, H. & Ohhata, T. Cell cycle regulation by long non-coding RNAs. *Cell Mol Life Sci* **70**, 4785-4794 (2013).
9. Zhu, S.*, et al.* Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. *Nat Biotechnol* **34**, 1279-1286 (2016).
10. Hung, T.*, et al.* Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat Genet* **43**, 621-629 (2011).
11. Ohyashiki, J.H.*, et al.* Replication errors in hematological neoplasias: genomic instability in progression of disease is different among different types of leukemia. *Clin Cancer Res* **2**, 1583-1589 (1996).
12. Sonugur, F.G. & Akbulut, H. The Role of Tumor Microenvironment in Genomic Instability of Malignant Tumors. *Front Genet* **10**, 1063 (2019).
13. Scott, R.J. & Meldrum, C.J. Missense mutations in cancer predisposing genes: can we make sense of them? *Hered Cancer Clin Pract* **3**, 123-127 (2005).
14. Clark, W.H. Tumour progression and the nature of cancer. *Br J Cancer* **64**, 631-644 (1991).
15. Khan, I. & Steeg, P.S. Metastasis suppressors: functional pathways. *Lab Invest* **98**, 198-210 (2018).
16. van der Meijden, C.M.*, et al.* Gene profiling of cell cycle progression through S-phase reveals sequential expression of genes required for

DNA replication and nucleosome assembly. *Cancer Res* **62**, 3233-3243 (2002).
17. Whitfield, M.L.*, et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* **13**, 1977-2000 (2002).
18. Schulz, K.N. & Harrison, M.M. Mechanisms regulating zygotic genome activation. *Nat Rev Genet* **20**, 221-234 (2019).
19. Niakan, K.K., Han, J., Pedersen, R.A., Simon, C. & Pera, R.A. Human pre-implantation embryo development. *Development* **139**, 829-841 (2012).
20. De Paepe, C., Krivega, M., Cauffman, G., Geens, M. & Van de Velde, H. Totipotency and lineage segregation in the human embryo. *Mol Hum Reprod* **20**, 599-618 (2014).
21. Kermi, C., Aze, A. & Maiorano, D. Preserving Genome Integrity During the Early Embryonic DNA Replication Cycles. *Genes (Basel)* **10**(2019).
22. Bui, A.D., Sharma, R., Henkel, R. & Agarwal, A. Reactive oxygen species impact on sperm DNA and its role in male infertility. *Andrologia* **50**, e13012 (2018).
23. Monk, M. & Holding, C. Human embryonic genes re-expressed in cancer cells. *Oncogene* **20**, 8085-8091 (2001).
24. van Deursen, J.M. The role of senescent cells in ageing. *Nature* **509**, 439-446 (2014).
25. Fischle, W., Wang, Y. & Allis, C.D. Histone and chromatin cross-talk. *Curr Opin Cell Biol* **15**, 172-183 (2003).
26. Nightingale, K.P.*, et al.* Cross-talk between histone modifications in response to histone deacetylase inhibitors: MLL4 links histone H3 acetylation and histone H3K4 methylation. *J Biol Chem* **282**, 4408-4416 (2007).
27. Margueron, R., Trojer, P. & Reinberg, D. The key to development: interpreting the histone code? *Curr Opin Genet Dev* **15**, 163-176 (2005).
28. Zhou, V.W., Goren, A. & Bernstein, B.E. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* **12**, 7-18 (2011).
29. Rada-Iglesias, A.*, et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283 (2011).
30. Maezawa, S.*, et al.* Polycomb protein SCML2 facilitates H3K27me3 to establish bivalent domains in the male germline. *Proc Natl Acad Sci U S A* **115**, 4957-4962 (2018).
31. Voigt, P., Tee, W.W. & Reinberg, D. A double take on bivalent promoters. *Genes Dev* **27**, 1318-1338 (2013).
32. Marcho, C., Cui, W. & Mager, J. Epigenetic dynamics during preimplantation development. *Reproduction* **150**, R109-120 (2015).

33. Raisner, R.*, et al.* Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation. *Cell Rep* **24**, 1722-1729 (2018).
34. Jambhekar, A., Dhall, A. & Shi, Y. Roles and regulation of histone methylation in animal development. *Nat Rev Mol Cell Biol* **20**, 625-641 (2019).
35. Gardner, K.E., Allis, C.D. & Strahl, B.D. Operating on chromatin, a colorful language where context matters. *J Mol Biol* **409**, 36-46 (2011).
36. Margueron, R.*, et al.* Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms. *Mol Cell* **32**, 503-518 (2008).
37. Shao, Z.*, et al.* Stabilization of chromatin structure by PRC1, a Polycomb complex. *Cell* **98**, 37-46 (1999).
38. Yang, Y.W.*, et al.* Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. *Elife* **3**, e02046 (2014).
39. Yang, X.*, et al.* Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26**, 577-590 (2014).
40. Schier, A.C. & Taatjes, D.J. Structure and mechanism of the RNA polymerase II transcription machinery. *Genes Dev* **34**, 465-488 (2020).
41. Nikolov, D.B. & Burley, S.K. RNA polymerase II transcription initiation: a structural view. *Proc Natl Acad Sci U S A* **94**, 15-22 (1997).
42. Pearson, J.C., Lemons, D. & McGinnis, W. Modulating Hox gene functions during animal body patterning. *Nat Rev Genet* **6**, 893-904 (2005).
43. Wellik, D.M. Hox patterning of the vertebrate axial skeleton. *Dev Dyn* **236**, 2454-2463 (2007).
44. Cobb, J. & Duboule, D. Comparative analysis of genes downstream of the Hoxd cluster in developing digits and external genitalia. *Development* **132**, 3055-3067 (2005).
45. Merkenschlager, M. & Odom, D.T. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell* **152**, 1285-1297 (2013).
46. Marchese, F.P., Raimondi, I. & Huarte, M. The multidimensional mechanisms of long noncoding RNA function. *Genome Biol* **18**, 206 (2017).
47. Gapp, K.*, et al.* Alterations in sperm long RNA contribute to the epigenetic inheritance of the effects of postnatal trauma. *Mol Psychiatry* **25**, 2162-2174 (2020).
48. Zhang, Y.*, et al.* Dnmt2 mediates intergenerational transmission of paternally acquired metabolic disorders through sperm small non-coding RNAs. *Nat Cell Biol* **20**, 535-540 (2018).
49. Siklenka, K.*, et al.* Disruption of histone methylation in developing sperm impairs offspring health transgenerationally. *Science* **350**, aab2006 (2015).

50. Wu, E. & Vastenhouw, N.L. From mother to embryo: A molecular perspective on zygotic genome activation. *Curr Top Dev Biol* **140**, 209-254 (2020).
51. Huypens, P.*, et al.* Epigenetic germline inheritance of diet-induced obesity and insulin resistance. *Nat Genet* **48**, 497-499 (2016).
52. Brykczynska, U.*, et al.* Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa. *Nat Struct Mol Biol* **17**, 679-687 (2010).
53. Hammoud, S.S.*, et al.* Distinctive chromatin in human sperm packages genes for embryo development. *Nature* **460**, 473-478 (2009).
54. Paradowska, A.S.*, et al.* Genome wide identification of promoter binding sites for H4K12ac in human sperm and its relevance for early embryonic development. *Epigenetics* **7**, 1057-1070 (2012).
55. Hammoud, S.S.*, et al.* Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. *Cell Stem Cell* **15**, 239-253 (2014).
56. Sankar, A.*, et al.* KDM4A regulates the maternal-to-zygotic transition by protecting broad H3K4me3 domains from H3K9me3 invasion in oocytes. *Nat Cell Biol* **22**, 380-388 (2020).
57. Saitou, M., Kagiwada, S. & Kurimoto, K. Epigenetic reprogramming in mouse pre-implantation development and primordial germ cells. *Development* **139**, 15-31 (2012).
58. Greenberg, M.V.C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* **20**, 590-607 (2019).
59. Perino, M. & Veenstra, G.J. Chromatin Control of Developmental Dynamics and Plasticity. *Dev Cell* **38**, 610-620 (2016).
60. Wajed, S.A., Laird, P.W. & DeMeester, T.R. DNA methylation: an alternative pathway to cancer. *Ann Surg* **234**, 10-20 (2001).
61. Kulis, M. & Esteller, M. DNA methylation and cancer. *Adv Genet* **70**, 27-56 (2010).
62. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nat Rev Genet* **17**, 551-565 (2016).
63. Tate, P.H. & Bird, A.P. Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr Opin Genet Dev* **3**, 226-231 (1993).
64. Schubeler, D. Function and information content of DNA methylation. *Nature* **517**, 321-326 (2015).
65. Nan, X.*, et al.* Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386-389 (1998).
66. van Vlodrop, I.J.*, et al.* Prognostic significance of Gremlin1 (GREM1) promoter CpG island hypermethylation in clear cell renal cell carcinoma. *Am J Pathol* **176**, 575-584 (2010).
67. Subhash, S., Andersson, P.O., Kosalai, S.T., Kanduri, C. & Kanduri, M. Global DNA methylation profiling reveals new insights into

epigenetically deregulated protein coding and long noncoding RNAs in CLL. *Clin Epigenetics* **8**, 106 (2016).
68. Kulis, M.*, et al.* Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat Genet* **44**, 1236-1242 (2012).
69. Koch, A.*, et al.* Analysis of DNA methylation in cancer: location revisited. *Nat Rev Clin Oncol* **15**, 459-466 (2018).
70. Mahmood, N. & Rabbani, S.A. DNA Methylation Readers and Cancer: Mechanistic and Therapeutic Applications. *Front Oncol* **9**, 489 (2019).
71. Ma, L., Bajic, V.B. & Zhang, Z. On the classification of long non-coding RNAs. *RNA Biol* **10**, 925-933 (2013).
72. Hamazaki, N., Nakashima, K., Hayashi, K. & Imamura, T. Detection of Bidirectional Promoter-Derived lncRNAs from Small-Scale Samples Using Pre-Amplification-Free Directional RNA-seq Method. *Methods Mol Biol* **1605**, 83-103 (2017).
73. Guzel, E.*, et al.* Tumor suppressor and oncogenic role of long non-coding RNAs in cancer. *North Clin Istanb* **7**, 81-86 (2020).
74. Noh, J.H., Kim, K.M., McClusky, W.G., Abdelmohsen, K. & Gorospe, M. Cytoplasmic functions of long noncoding RNAs. *Wiley Interdiscip Rev RNA* **9**, e1471 (2018).
75. Rashid, F., Shah, A. & Shan, G. Long Non-coding RNAs in the Cytoplasm. *Genomics Proteomics Bioinformatics* **14**, 73-80 (2016).
76. Guh, C.Y., Hsieh, Y.H. & Chu, H.P. Functions and properties of nuclear lncRNAs-from systematically mapping the interactomes of lncRNAs. *J Biomed Sci* **27**, 44 (2020).
77. Sun, Q., Hao, Q. & Prasanth, K.V. Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression. *Trends Genet* **34**, 142-157 (2018).
78. Gudenas, B.L. & Wang, L. Prediction of LncRNA Subcellular Localization with Deep Learning from Sequence Features. *Sci Rep* **8**, 16385 (2018).
79. Luo, S.*, et al.* Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. *Cell Stem Cell* **18**, 637-652 (2016).
80. Geisler, S. & Coller, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat Rev Mol Cell Biol* **14**, 699-712 (2013).
81. Hon, C.C.*, et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**, 199-204 (2017).
82. Zhao, J.*, et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40**, 939-953 (2010).
83. Loewer, S.*, et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**, 1113-1117 (2010).

84. Lopez-Pajares, V.*, et al.* A LncRNA-MAF:MAFB transcription factor network regulates epidermal differentiation. *Dev Cell* **32**, 693-706 (2015).
85. Mohammad, F.*, et al.* Long noncoding RNA-mediated maintenance of DNA methylation and transcriptional gene silencing. *Development* **139**, 2792-2803 (2012).
86. Mondal, T.*, et al.* Sense-Antisense lncRNA Pair Encoded by Locus 6p22.3 Determines Neuroblastoma Susceptibility via the USP36-CHD7-SOX9 Regulatory Axis. *Cancer Cell* **33**, 417-434 e417 (2018).
87. Choi, S.W., Kim, H.W. & Nam, J.W. The small peptide world in long noncoding RNAs. *Brief Bioinform* **20**, 1853-1864 (2019).
88. Szafron, L.M.*, et al.* The Novel Gene CRNDE Encodes a Nuclear Peptide (CRNDEP) Which Is Overexpressed in Highly Proliferating Tissues. *PLoS One* **10**, e0127475 (2015).
89. Ruiz-Orera, J., Messeguer, X., Subirana, J.A. & Alba, M.M. Long non-coding RNAs as a source of new peptides. *Elife* **3**, e03523 (2014).
90. Guil, S. & Esteller, M. Cis-acting noncoding RNAs: friends and foes. *Nat Struct Mol Biol* **19**, 1068-1075 (2012).
91. Mondal, T.*, et al.* MEG3 long noncoding RNA regulates the TGF-beta pathway genes through formation of RNA-DNA triplex structures. *Nat Commun* **6**, 7743 (2015).
92. Pandey, R.R.*, et al.* Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell* **32**, 232-246 (2008).
93. Gomez, J.A.*, et al.* The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus. *Cell* **152**, 743-754 (2013).
94. Kapusta, A. & Feschotte, C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet* **30**, 439-452 (2014).
95. Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H. & Bartel, D.P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537-1550 (2011).
96. Smith, M.A., Gesell, T., Stadler, P.F. & Mattick, J.S. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res* **41**, 8220-8236 (2013).
97. Kirk, J.M.*, et al.* Functional classification of long non-coding RNAs by k-mer content. *Nat Genet* **50**, 1474-1482 (2018).
98. Diederichs, S. The four dimensions of noncoding RNA conservation. *Trends Genet* **30**, 121-123 (2014).
99. Flynn, R.A. & Chang, H.Y. Long noncoding RNAs in cell-fate programming and reprogramming. *Cell Stem Cell* **14**, 752-761 (2014).

100. Akhade, V.S., Pal, D. & Kanduri, C. Long Noncoding RNA: Genome Organization and Mechanism of Action. *Adv Exp Med Biol* **1008**, 47-74 (2017).
101. Tripathi, V.*, et al.* Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. *PLoS Genet* **9**, e1003368 (2013).
102. Liu, M.*, et al.* HOTAIR, a long noncoding RNA, is a marker of abnormal cell cycle regulation in lung cancer. *Cancer Sci* **109**, 2717-2733 (2018).
103. Rinn, J.L. & Chang, H.Y. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* **81**, 145-166 (2012).
104. Mondal, T., Rasmussen, M., Pandey, G.K., Isaksson, A. & Kanduri, C. Characterization of the RNA content of chromatin. *Genome Res* **20**, 899-907 (2010).
105. Mishra, K. & Kanduri, C. Understanding Long Noncoding RNA and Chromatin Interactions: What We Know So Far. *Noncoding RNA* **5**(2019).
106. Stavropoulos, N., Lu, N. & Lee, J.T. A functional role for Tsix transcription in blocking Xist RNA accumulation but not in X-chromosome choice. *Proc Natl Acad Sci U S A* **98**, 10232-10237 (2001).
107. Hacisuleyman, E.*, et al.* Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat Struct Mol Biol* **21**, 198-206 (2014).
108. Fitzpatrick, G.V., Soloway, P.D. & Higgins, M.J. Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1. *Nat Genet* **32**, 426-431 (2002).
109. Mancini-Dinardo, D., Steele, S.J., Levorse, J.M., Ingram, R.S. & Tilghman, S.M. Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes Dev* **20**, 1268-1282 (2006).
110. Cabianca, D.S.*, et al.* A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. *Cell* **149**, 819-831 (2012).
111. Gupta, R.A.*, et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071-1076 (2010).
112. Prensner, J.R.*, et al.* The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat Genet* **45**, 1392-1398 (2013).
113. Perry, R.B. & Ulitsky, I. The functions of long noncoding RNAs in development and stem cells. *Development* **143**, 3882-3894 (2016).
114. Cabili, M.N.*, et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915-1927 (2011).

115. Amort, T.*, et al.* Long non-coding RNAs as targets for cytosine methylation. *RNA Biol* **10**, 1003-1008 (2013).
116. Ohhata, T., Senner, C.E., Hemberger, M. & Wutz, A. Lineage-specific function of the noncoding Tsix RNA for Xist repression and Xi reactivation in mice. *Genes Dev* **25**, 1702-1715 (2011).
117. Redrup, L.*, et al.* The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. *Development* **136**, 525-530 (2009).
118. Kanduri, C. Long noncoding RNAs: Lessons from genomic imprinting. *Biochim Biophys Acta* **1859**, 102-111 (2016).
119. Grote, P.*, et al.* The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev Cell* **24**, 206-214 (2013).
120. Zhang, K., Huang, K., Luo, Y. & Li, S. Identification and functional analysis of long non-coding RNAs in mouse cleavage stage embryonic development based on single cell transcriptome data. *BMC Genomics* **15**, 845 (2014).
121. Li, L., Zheng, P. & Dean, J. Maternal control of early mouse development. *Development* **137**, 859-870 (2010).
122. Ostermeier, G.C., Dix, D.J., Miller, D., Khatri, P. & Krawetz, S.A. Spermatozoal RNA profiles of normal fertile men. *Lancet* **360**, 772-777 (2002).
123. Ren, X., Chen, X., Wang, Z. & Wang, D. Is transcription in sperm stationary or dynamic? *J Reprod Dev* **63**, 439-443 (2017).
124. Alves, M.B.R.*, et al.* Sperm-borne miR-216b modulates cell proliferation during early embryo development via K-RAS. *Sci Rep* **9**, 10358 (2019).
125. Bohacek, J. & Rassoulzadegan, M. Sperm RNA: Quo vadis? *Semin Cell Dev Biol* **97**, 123-130 (2020).
126. Tsoi, M.S.*, et al.* Deposition of IgM and complement at the dermoepidermal junction in acute and chronic cutaneous graft-vs-host disease in man. *J Immunol* **120**, 1485-1492 (1978).
127. Conine, C.C., Sun, F., Song, L., Rivera-Perez, J.A. & Rando, O.J. Small RNAs Gained during Epididymal Transit of Sperm Are Essential for Embryonic Development in Mice. *Dev Cell* **46**, 470-480 e473 (2018).
128. Skerget, S., Rosenow, M.A., Petritis, K. & Karr, T.L. Sperm Proteome Maturation in the Mouse Epididymis. *PLoS One* **10**, e0140650 (2015).
129. Yan, X.*, et al.* Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. *Cancer Cell* **28**, 529-540 (2015).
130. Wang, Z.*, et al.* lncRNA Epigenetic Landscape Analysis Identifies EPIC1 as an Oncogenic lncRNA that Interacts with MYC and Promotes Cell-Cycle Progression in Cancer. *Cancer Cell* **33**, 706-720 e709 (2018).

131. Chiu, H.S., *et al.* Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context. *Cell Rep* **23**, 297-312 e212 (2018).
132. Iyer, M.K., *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47**, 199-208 (2015).
133. Loewen, G., Zhuo, Y., Zhuang, Y., Jayawickramarajah, J. & Shan, B. lincRNA HOTAIR as a novel promoter of cancer progression. *J Can Res Updates* **3**, 134-140 (2014).
134. Hajjari, M. & Salavaty, A. HOTAIR: an oncogenic long non-coding RNA in different cancers. *Cancer Biol Med* **12**, 1-9 (2015).
135. Kogo, R., *et al.* Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res* **71**, 6320-6326 (2011).
136. Ji, P., *et al.* MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**, 8031-8041 (2003).
137. Gutschner, T., *et al.* The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res* **73**, 1180-1189 (2013).
138. Graham, L.D., *et al.* Colorectal Neoplasia Differentially Expressed (CRNDE), a Novel Gene with Elevated Expression in Colorectal Adenomas and Adenocarcinomas. *Genes Cancer* **2**, 829-840 (2011).
139. Prensner, J.R., *et al.* Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol* **29**, 742-749 (2011).
140. Shang, Z., *et al.* LncRNA PCAT1 activates AKT and NF-kappaB signaling in castration-resistant prostate cancer by regulating the PHLPP/FKBP51/IKKalpha complex. *Nucleic Acids Res* **47**, 4211-4225 (2019).
141. Chakravarty, D., *et al.* The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. *Nat Commun* **5**, 5383 (2014).
142. Adriaens, C., *et al.* p53 induces formation of NEAT1 lncRNA-containing paraspeckles that modulate replication stress response and chemosensitivity. *Nat Med* **22**, 861-868 (2016).
143. Li, X., *et al.* Oncogenic Properties of NEAT1 in Prostate Cancer Cells Depend on the CDC5L-AGRN Transcriptional Regulation Circuit. *Cancer Res* **78**, 4138-4149 (2018).
144. Pandey, G.K., *et al.* The risk-associated long noncoding RNA NBAT-1 controls neuroblastoma progression by regulating cell proliferation and neuronal differentiation. *Cancer Cell* **26**, 722-737 (2014).
145. Sun, M., *et al.* Decreased expression of long noncoding RNA GAS5 indicates a poor prognosis and promotes cell proliferation in gastric cancer. *BMC Cancer* **14**, 319 (2014).

146. Yin, D.*, et al.* Long noncoding RNA GAS5 affects cell proliferation and predicts a poor prognosis in patients with colorectal cancer. *Med Oncol* **31**, 253 (2014).
147. Ji, J., Dai, X., Yeung, S.J. & He, X. The role of long non-coding RNA GAS5 in cancers. *Cancer Manag Res* **11**, 2729-2737 (2019).
148. Lu, X.*, et al.* Downregulation of gas5 increases pancreatic cancer cell proliferation by regulating CDK6. *Cell Tissue Res* **354**, 891-896 (2013).
149. Zhu, M.*, et al.* MEG3 overexpression inhibits the tumorigenesis of breast cancer by downregulating miR-21 through the PI3K/Akt pathway. *Arch Biochem Biophys* **661**, 22-30 (2019).
150. Martianov, I., Ramadass, A., Serra Barros, A., Chow, N. & Akoulitchev, A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445**, 666-670 (2007).
151. Bernard, D.*, et al.* A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *EMBO J* **29**, 3082-3093 (2010).
152. Feng, J.*, et al.* The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev* **20**, 1470-1484 (2006).
153. Faust, T., Frankel, A. & D'Orso, I. Transcription control by long non-coding RNAs. *Transcription* **3**, 78-86 (2012).
154. Zhang, X.*, et al.* Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels. *Int J Mol Sci* **20**(2019).
155. Yoon, J.H.*, et al.* LincRNA-p21 suppresses target mRNA translation. *Mol Cell* **47**, 648-655 (2012).
156. Davidovich, C. & Cech, T.R. The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. *RNA* **21**, 2007-2022 (2015).
157. Rinn, J.L.*, et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-1323 (2007).
158. Lai, F.*, et al.* Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* **494**, 497-501 (2013).
159. Melo, C.A.*, et al.* eRNAs are required for p53-dependent enhancer activity and gene transcription. *Mol Cell* **49**, 524-535 (2013).
160. Li, W.*, et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516-520 (2013).
161. Hsieh, C.L.*, et al.* Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proc Natl Acad Sci U S A* **111**, 7319-7324 (2014).
162. Pnueli, L., Rudnizky, S., Yosefzon, Y. & Melamed, P. RNA transcribed from a distal enhancer is required for activating the

chromatin at the promoter of the gonadotropin alpha-subunit gene. *Proc Natl Acad Sci U S A* **112**, 4369-4374 (2015).
163. Pugacheva, E.M.*, et al.* CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. *Proc Natl Acad Sci U S A* **117**, 2020-2031 (2020).
164. Gong, C. & Maquat, L.E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* **470**, 284-288 (2011).
165. Wang, J., Gong, C. & Maquat, L.E. Control of myogenesis by rodent SINE-containing lncRNAs. *Genes Dev* **27**, 793-804 (2013).
166. Meissner, A.*, et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33**, 5868-5877 (2005).
167. Chiesa, N.*, et al.* The KCNQ1OT1 imprinting control region and non-coding RNA: new properties derived from the study of Beckwith-Wiedemann syndrome and Silver-Russell syndrome cases. *Hum Mol Genet* **21**, 10-25 (2012).
168. Mondal, T., Subhash, S. & Kanduri, C. Chromatin RNA Immunoprecipitation (ChRIP). *Methods Mol Biol* **1689**, 65-76 (2018).
169. Mariner, P.D.*, et al.* Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol Cell* **29**, 499-509 (2008).
170. Akhade, V.S., Arun, G., Donakonda, S. & Rao, M.R. Genome wide chromatin occupancy of mrhl RNA and its role in gene regulation in mouse spermatogonial cells. *RNA Biol* **11**, 1262-1279 (2014).
171. Li, X.*, et al.* GRID-seq reveals the global RNA-chromatin interactome. *Nat Biotechnol* **35**, 940-950 (2017).
172. Lieberman-Aiden, E.*, et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).
173. Ali, M.M.*, et al.* PAN-cancer analysis of S-phase enriched lncRNAs identifies oncogenic drivers and biomarkers. *Nat Commun* **9**, 883 (2018).
174. Subhash, S.*, et al.* H3K4me2 and WDR5 enriched chromatin interacting long non-coding RNAs maintain transcriptionally competent chromatin at divergent transcriptional units. *Nucleic Acids Res* **46**, 9384-9400 (2018).
175. Subhash, S., Kanduri, M. & Kanduri, C. Sperm Originated Chromatin Imprints and LincRNAs in Organismal Development and Cancer. *iScience* **23**, 101165 (2020).
176. Meryet-Figuiere, M.*, et al.* Temporal separation of replication and transcription during S-phase progression. *Cell Cycle* **13**, 3241-3248 (2014).
177. Montes, M. & Lund, A.H. Emerging roles of lncRNAs in senescence. *FEBS J* **283**, 2414-2426 (2016).

178. Akhade, V.S., Dighe, S.N., Kataruka, S. & Rao, M.R. Mechanism of Wnt signaling induced down regulation of mrhl long non-coding RNA in mouse spermatogonial cells. *Nucleic Acids Res* **44**, 387-401 (2016).
179. Wang, J.*, et al.* Crosstalk between transforming growth factor-beta signaling pathway and long non-coding RNAs in cancer. *Cancer Lett* **370**, 296-301 (2016).
180. Jung, C., Mittler, G., Oswald, F. & Borggrefe, T. RNA helicase Ddx5 and the noncoding RNA SRA act as coactivators in the Notch signaling pathway. *Biochim Biophys Acta* **1833**, 1180-1189 (2013).
181. Zhu, Y.*, et al.* HULC long noncoding RNA silencing suppresses angiogenesis by regulating ESM-1 via the PI3K/Akt/mTOR signaling pathway in human gliomas. *Oncotarget* **7**, 14429-14440 (2016).
182. Han, Y.*, et al.* Tumor-suppressive function of long noncoding RNA MALAT1 in glioma cells by downregulation of MMP2 and inactivation of ERK/MAPK signaling. *Cell Death Dis* **7**, e2123 (2016).
183. Basaki, Y.*, et al.* Y-box binding protein-1 (YB-1) promotes cell cycle progression through CDC6-dependent pathway in human cancer cells. *Eur J Cancer* **46**, 954-965 (2010).
184. Gallardo, M.*, et al.* Aberrant hnRNP K expression: All roads lead to cancer. *Cell Cycle* **15**, 1552-1557 (2016).
185. Huarte, M.*, et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409-419 (2010).
186. Subhash, S., Ali, M.M. & Kanduri, C. S-phase cancer associated lncRNAs. *Cell Cycle* **17**, 2517-2519 (2018).
187. Trievel, R.C. & Shilatifard, A. WDR5, a complexed protein. *Nat Struct Mol Biol* **16**, 678-680 (2009).
188. Wang, Y., Li, X. & Hu, H. H3K4me2 reliably defines transcription factor binding regions in different cells. *Genomics* **103**, 222-228 (2014).
189. Pekowska, A., Benoukraf, T., Ferrier, P. & Spicuglia, S. A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Res* **20**, 1493-1502 (2010).
190. Yuan, S.*, et al.* Sperm-borne miRNAs and endo-siRNAs are important for fertilization and preimplantation embryonic development. *Development* **143**, 635-647 (2016).
191. Xia, W.*, et al.* Resetting histone modifications during human parental-to-zygotic transition. *Science* **365**, 353-360 (2019).
192. Hamm, D.C. & Harrison, M.M. Regulatory principles governing the maternal-to-zygotic transition: insights from Drosophila melanogaster. *Open Biol* **8**, 180183 (2018).
193. DeRenzo, C. & Seydoux, G. A clean start: degradation of maternal proteins at the oocyte-to-embryo transition. *Trends Cell Biol* **14**, 420-426 (2004).

194. Bui, H.T.*, et al.* Essential role of paternal chromatin in the regulation of transcriptional activity during mouse preimplantation development. *Reproduction* **141**, 67-77 (2011).
195. Tomizawa, S.I.*, et al.* Kmt2b conveys monovalent and bivalent H3K4me3 in mouse spermatogonial stem cells at germline and embryonic promoters. *Development* **145**(2018).
196. Simpson, A.J., Caballero, O.L., Jungbluth, A., Chen, Y.T. & Old, L.J. Cancer/testis antigens, gametogenesis and cancer. *Nat Rev Cancer* **5**, 615-625 (2005).
197. Bouckenheimer, J.*, et al.* Long non-coding RNAs in human early embryonic development and their potential in ART. *Hum Reprod Update* **23**, 19-40 (2016).
198. Shah, K., Patel, S., Mirza, S. & Rawal, R.M. Unravelling the link between embryogenesis and cancer metastasis. *Gene* **642**, 447-452 (2018).
199. Aiello, N.M. & Stanger, B.Z. Echoes of the embryo: using the developmental biology toolkit to study cancer. *Dis Model Mech* **9**, 105-114 (2016).
200. Hosono, Y.*, et al.* Oncogenic Role of THOR, a Conserved Cancer/Testis Long Non-coding RNA. *Cell* **171**, 1559-1572 e1520 (2017).
201. Yates, A.D.*, et al.* Ensembl 2020. *Nucleic Acids Res* **48**, D682-D688 (2020).
202. Frankish, A.*, et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773 (2019).
203. Haeussler, M.*, et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* **47**, D853-D858 (2019).
204. Stelzer, G.*, et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics* **54**, 1 30 31-31 30 33 (2016).
205. Liu, T.*, et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol* **12**, R83 (2011).
206. Davis, C.A.*, et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**, D794-D801 (2018).
207. Tate, J.G.*, et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947 (2019).
208. Sondka, Z.*, et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696-705 (2018).
209. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
210. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915 (2019).

211. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).
212. Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).
213. Ernst, J. & Bar-Joseph, Z. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* **7**, 191 (2006).
214. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
215. Bi, R. & Liu, P. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinformatics* **17**, 146 (2016).
216. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
217. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
218. Li, H.*, et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
219. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
220. Breese, M.R. & Liu, Y. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics* **29**, 494-496 (2013).
221. Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**, W187-191 (2014).
222. Kong, L.*, et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35**, W345-349 (2007).
223. Heinz, S.*, et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).
224. Wadi, L., Meyer, M., Weiser, J., Stein, L.D. & Reimand, J. Impact of outdated gene annotations on pathway enrichment analysis. *Nat Methods* **13**, 705-706 (2016).
225. Subhash, S. & Kanduri, C. GeneSCF: a real-time based functional enrichment tool with support for multiple organisms. *BMC Bioinformatics* **17**, 365 (2016).
226. The Gene Ontology, C. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330-D338 (2019).

227. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353-D361 (2017).
228. Fabregat, A.*, et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res* **46**, D649-D655 (2018).
229. Repana, D.*, et al.* The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol* **20**, 1 (2019).
230. Jiang, M.C., Ni, J.J., Cui, W.Y., Wang, B.Y. & Zhuo, W. Emerging roles of lncRNA in cancer and therapeutic opportunities. *Am J Cancer Res* **9**, 1354-1366 (2019).
231. Arun, G.*, et al.* Differentiation of mammary tumors and reduction in metastasis upon Malat1 lncRNA loss. *Genes Dev* **30**, 34-51 (2016).
232. Haery, L., Thompson, R.C. & Gilmore, T.D. Histone acetyltransferases and histone deacetylases in B- and T-cell development, physiology and malignancy. *Genes Cancer* **6**, 184-213 (2015).
233. Berner, A.K. & Kleinman, M.E. Therapeutic Approaches to Histone Reprogramming in Retinal Degeneration. *Adv Exp Med Biol* **854**, 39-44 (2016).