

**Genetic variants within long non coding RNA: Role in  
Cancer**

**THESIS SUBMITTED FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY (SCIENCE) IN  
DEPARTMENT OF LIFE SCIENCE AND  
BIOTECHNOLOGY**

**TROYEE DAS**

**DEPARTMENT OF LIFE SCIENCE AND  
BIOTECHNOLOGY**

**JADAVPUR UNIVERSITY**

**2024**





# Bose Institute

**Dr. ZHUMUR GHOSH**  
**Associate Professor**  
Department of Biological Sciences


**Unified Academic Campus**  
EN 80, Sector V, Bidhan Nagar  
Kolkata - 700091 WB India  
Email: [zhumur@jcbose.ac.in](mailto:zhumur@jcbose.ac.in)  
Also: [ghosh.jhumur@gmail.com](mailto:ghosh.jhumur@gmail.com)

## TO WHOM IT MAY CONCERN

This is to certify that the thesis entitled “**Genetic Variants within long non coding RNA: Role in Cancer**” submitted by **Smt. Troyee Das** who got her name registered on 30.08.2019, for the award of Ph. D. (Science) degree of Jadavpur University, is absolutely based upon her own work under the supervision of **Dr. Zhumur Ghosh** and that neither this thesis nor any part of it has been submitted for either any degree / diploma or any other academic award anywhere before.

*Zhumur Ghosh 15/03/2024*

(Signature of the Supervisor with date and official seal)

 **डॉ. झुमुर घोष / Dr. Zhumur Ghosh**  
**एसोसिएट प्रोफेसर / Associate Professor**  
**जैविक विज्ञान विभाग / Department of Biological Sciences**  
**बसु विज्ञान मंदिर / BOSE INSTITUTE**  
**ईएन 80, सेक्टर V, बिधाननगर/EN 80, Sector V, Bidhan Nagar**  
**कोलकाता / Kolkata-700 091 (भारत/India)**





## ACKNOWLEDGEMENT

---

---

I am deeply grateful to all those whose support, guidance, and encouragement have played a pivotal role in the completion of my doctoral thesis. My academic journey has been enriched by the invaluable contributions of numerous individuals from within academia and beyond. Their collective support, guidance, and encouragement have been invaluable.

First and foremost, I would like to express my heartfelt gratitude to my advisor, Dr. Zhumur Ghosh. Your wisdom, patience, and encouragement have been a constant source of inspiration. Your mentorship has not only enriched my research but also my personal growth. I am deeply thankful for your guidance throughout this journey.

I would like to extend my appreciation to the member of my thesis committee, Dr. Angshuman Bagchi. Your insightful feedback and constructive criticism have pushed me to refine my work and think critically. I am honored to have had the opportunity to learn from you.

I thank CSIR and Bose Institute for providing research fellowship for my work.

I am also indebted to the staff and faculty at Division of Bioinformatics, Bose Institute, whose dedication to fostering an intellectually stimulating environment has been instrumental in my development as a researcher. The resources and opportunities provided by the Institute have been indispensable to my success.

I want to express my deepest gratitude to Ma and Baba, whose unwavering support and sacrifices have made it possible for me to reach this significant stage in my education. Your belief in me has been my greatest motivator, and I am forever indebted to you for your love and encouragement. I also want to thank my Pisin and Jethun who always wished for nothing but the best for my career. Jethun, your blessings have been a source of strength throughout this journey, and I carry your memory with me in my heart. To my Jethun and my mentor, Sunil Sir, who are no longer with us, I wish you could have witnessed this moment, for this accomplishment will always feel incomplete without you beside me. Your support and love have been the foundation upon which I've built my success, and I will forever cherish the memories that we've shared.

I'd like to extend a special thanks to Anita. We sometimes discover friendship in the most unexpected places. A friend's warmth provides solace and comfort during the most turbulent times. Unknowingly, Anita became that person. Our friendship is something I hold dear in my heart. Thank you for always listening to me, cheering me on and believing in me.

My most heartfelt thanks to Lee Jin Ki and SHINee, as their music have become inseparable from my life. They are my constant source of emotional strength. Kim Jong Hyun, you have now become the brightest star in the sky. Whenever I find myself in doubt, I hold onto your words of wisdom, and it gives me the strength to move forward.

## *Acknowledgement*

I would also thank my three wonderful colleagues and dear friends, Byapti Ghosh, Abhirupa Ghosh and Debadrita Basu aka the "Batch of 2017 Girls." From the early days of orientation to the final stages of thesis completion, we forged bonds that went beyond academia. Through the ups and downs, we celebrated each other's victories and offered solace during moments of frustration. Together, we faced the challenges, shared knowledge, and created a network of support that extended far beyond our research topics.

To my lab seniors and colleagues who have been my companions through the highs and lows of this journey, thank you for your camaraderie and support. Your presence has made the long hours of research and writing more enjoyable. Sincere thanks to seniors Dr. Arijita Sarkar, Dr. Aritra Deb, Dr. Arpana Verma, Gourab Das, Pritha Sengupta, Satakshi Bagchi, Sudip Mondal, Namrata Bhattacharya, Adrija Das.

Lastly, I want to express my gratitude to all the researchers, scholars, and institutions whose work has paved the way for my own. Your contributions to the field have inspired and informed my research.

In closing, this thesis is the result of the collective efforts and support of numerous individuals and institutions. While I may not be able to name everyone, please know that your contributions, no matter how small, have left an indelible mark on my journey.

This journey has been more than just academic achievements; it's been about personal growth. I've discovered the significance of determination, perseverance, and the delight of learning. It's been a transformative experience.

# Contents

PREFACE .....	1
CHAPTER 1  A BRIEF REVIEW .....	3
1. INTRODUCTION .....	3
1.1. FEMALE CANCERS: CAUSES AND CONCERN .....	3
1.2. LONG NON-CODING RNA (LNCRNA).....	4
1.3. GENETIC VARIANTS WITHIN THE LNCRNA LOCI .....	5
1.4. ASSOCIATION BETWEEN SNPS AND CANCER RISK .....	6
2. REVIEW OF STATUS OF RESEARCH AND DEVELOPMENT IN THE SUBJECT .....	7
2.1. INTERNATIONAL STATUS:.....	7
2.2. NATIONAL STATUS:.....	10
3. CONSEQUENCE OF THE PRESENCE OF SNP WITHIN LNCRNA .....	11
4. IMPORTANCE OF THE WORK IN THE CONTEXT OF CURRENT STATUS	12
5. REFERENCES.....	13
CHAPTER 2  LNCRBASE V.2: AN UPDATED RESOURCE FOR MULTISPECIES LNCRNAS.....	17
ABSTRACT:.....	17
1. INTRODUCTION .....	17
2. MATERIALS AND METHODS.....	19
IMPROVED CONTENT AND NEW FEATURES .....	19

2.1.	DATA PROCUREMENT.....	19
2.2.	DATA PROCESSING AND REFINEMENT .....	20
2.2.1.	REDUNDANCY CHECK AND ASSIGNING ALIAS ID.....	20
2.2.2.	ASSOCIATION WITH CPG ISLAND, REPEAT ELEMENTS AND SMALL NON-CODING RNAS.....	20
2.2.3.	DETERMINING PUTATIVE LNC-PRI-MIRNAS .....	20
2.2.4.	CODING CAPACITY OF LNCRNAS:.....	21
2.2.5.	SUB-CELLULAR LOCALIZATION OF LNCRNAS .....	21
2.2.6.	TFBS IN LNCRNA PROMOTER REGION.....	21
2.2.7.	LNCRNA PROFILE ACROSS MULTIPLE TISSUES.....	22
2.3.	DATABASE IMPLEMENTATION: .....	22
2.3.1.	THE "TRANSCRIPTS" MENU:.....	22
I.	SEARCH BY LNCRNA ACCESSION ID:.....	22
II.	SEARCH BY LNCRNA GENE SYMBOL:.....	22
III.	BROWSE BY LNCRNA SUBTYPE, CODING POTENTIAL AND SORF OVERLAP:.....	23
IV.	SEARCH FOR ASSOCIATED GENOMIC ELEMENTS:.....	23
V.	SEARCH FOR ASSOCIATION WITH SMALL NCRNA: .....	23
2.3.2.	SPECIES-SPECIFIC LNCRNA EXPRESSION PROFILE: .....	23
3.	RESULTS AND DISCUSSION:.....	25
3.1.	DISTRIBUTION OF LNCRNA SUBTYPES BASED ON THEIR GENOMIC LOCATION .....	25
3.2.	DISTRIBUTION OF REPEAT ELEMENTS WITHIN LNCRNA LOCI.....	28

3.3. ABUNDANCE OF PIWIL INTERACTING RNAS (PIRNAS) WITHIN LNCRNA LOCI .....	29
3.4. CGI ASSOCIATION WITH LNCRNA PROMOTER REGION.....	30
3.5. TISSUE SPECIFIC DISTRIBUTION OF LNCRNAS: .....	31
3.6. BIFUNCTIONAL LNCRNAS: .....	32
3.7. LNCRNAS HARBOURING EMBEDDED MIRNAS: .....	33
3.8. CELLULAR LOCALIZATION OF LNCRNAS: .....	35
3.9. TRANSCRIPTION FACTOR BINDING SITE IN LNCRNA PROMOTER REGIONS: .....	36
3.10. LNCRNA TARGET PARTNERS AND DISEASE ASSOCIATION: .....	36
4. CONCLUSION: .....	37
5. AVAILABILITY .....	38
6. REFERENCES: .....	38
CHAPTER 3  CLINICLSNP: A DATABASE HOSTING GENETIC VARIANTS IN LNCRNAS FOR CANCER PATIENTS .....	41
ABSTRACT: .....	41
1. INTRODUCTION .....	41
2. MATERIALS AND METHODS .....	42
2.1. RAW DATA CORRESPONDING TO THE THREE CANCER SYSTEMS: ...	42
2.2. RAW DATA ANALYSIS AND FEATURE MAPPING: .....	42
2.2.1. SNP DETECTION: .....	42
2.2.2. MAPPING SNP ASSOCIATION WITH REPEAT AND CGI .....	43
2.2.3. MAPPING TFBS OVERLAP WITH SNP .....	43

2.2.4.	MAPPING TAG SNPS .....	43
2.2.5.	DETERMINING STRUCTURE PERTURBATION SCORE DUE TO PRESENCE OF SNP .....	43
2.3.	DATABASE IMPLEMENTATION: .....	43
2.3.1.	SEARCH OPTIONS UNDER SECTION “CLINICLSNP”:.....	45
I.	SEARCH BY SNP ID/POSITION:.....	45
II.	SEARCH BY CELL LINE: .....	45
III.	SEARCH BY DISEASE:.....	45
3.	RESULTS AND DISCUSSION:.....	46
3.1.	MAPPING OF DETECTED SNPS WITHIN LNCRNA: .....	47
3.2.	CLINICAL RELEVANCE OF THE LNCRNA-SNPS IN THE THREE CANCER SYSTEMS:.....	48
3.2.1.	MAPPING WITH CLINVARIDS: .....	49
3.2.2.	MAPPING TAG SNPS: .....	49
3.2.3.	IDENTIFICATION OF POTENTIALLY PATHOGENIC VARIANTS:.....	49
3.3.	REGULATORY SNPS (RSNPS) WITHIN LNCRNA LOCI:.....	50
3.4.	CPG-SNPS WITHIN LNCRNA-LOCI: .....	51
3.5.	SNP ASSOCIATED WITH REPEAT ELEMENTS:.....	51
3.6.	EFFECT OF SNP ON LNCRNA SECONDARY STRUCTURE:.....	52
3.7.	CASE STUDY WITH CLINICLSNP RESULTS: .....	53
4.	CONCLUSION: .....	54
5.	AVAILABILITY .....	54

6. REFERENCES:.....	54
CHAPTER 4  SHARED LNCRNA VARIANTS IN FEMALE CANCERS .....	57
ABSTRACT:.....	57
1. BACKGROUND AND OBJECTIVE OF THE STUDY .....	57
2. MATERIALS AND METHODS.....	59
2.1. RAW DATA CORRESPONDING TO THE THREE CANCER SYSTEMS: ...	59
2.2. DETECTION OF DIFFERENTIALLY EXPRESSED LNCRNAs IN THE THREE CANCER SYSTEMS .....	60
2.3. SNP DETECTION IN PATIENT SAMPLES: .....	60
2.4. DETECTION OF COMMON DE LNCRNA-SNPS.....	60
2.5. CULTURE OF CELL LINES.....	61
2.6. GENOTYPING BY THE TAQMAN PCR ASSAY .....	61
2.7. REAL-TIME PCR ANALYSIS:.....	62
3. RESULTS AND DISCUSSION:.....	63
3.1. ANALYSIS OF BREAST, CERVICAL AND OVARIAN CANCER DATASETS REVEALED ONLY ONE COMMON DE LNCRNA-SNP ACROSS THE THREE SYSTEMS: .....	63
3.2. MIR4435-2HG HAS REPORTS OF REGULATION IN BREAST, CERVICAL AND OVARIAN CANCER BUT NOT THE SNP:.....	65
3.3. TAQMAN GENOTYPING ASSAY VALIDATED THE PRESENCE OF HETEROZYGOUS AND HOMOZYGOUS RECESSIVE ALLELE OF THE VARIANT IN CANCER AND CONTROL SYSTEMS .....	65
3.4. REAL-TIME PCR ANALYSIS REVEALED THE UPREGULATED EXPRESSION OF LNCRNA MIR4435-2HG IN THE CANCER CELL LINE CARRYING THE HETEROZYGOUS ALLELE OF RS1045267 .....	66
4. CONCLUSION: .....	68

5. REFERENCES:	68
CHAPTER 5  DEVELOPING LNCRNA-SNP BASED BREAST AND OVARIAN CANCER RISK PREDICTION MODELS AND ITS EFFECT ON GENE REGULATION	71
ABSTRACT:	71
1. BACKGROUND AND OBJECTIVE OF THE STUDY	71
2. MATERIALS AND METHODS:	73
2.1. RAW DATA CORRESPONDING TO BREAST AND OVARIAN CANCER SYSTEMS AND THEIR NORMAL COUNTERPARTS:	73
2.2. PRE-PROCESSING AND SNP DETECTION THROUGHOUT TRANSCRIPTOMIC DATA	74
2.3. SELECTION OF EXCLUSIVE BREAST AND OVARIAN LNCRNA-SNPS.	74
2.4. DETECTION OF LNCRNA-SNP EXPRESSION AS WELL AS DIFFERENTIALLY EXPRESSED TRANSCRIPTS IN THE TWO CANCER SYSTEMS AGAINST THEIR NORMAL COUNTERPARTS	75
2.5. SCREENING INTERACTING GENE PARTNERS OF LNCRNA-SNPS THROUGH CORRELATION ANALYSIS, RNA-RNA INTERACTION INFORMATION AND CE(COMPETITIVE ENDOGENOUS)-RNA NETWORK ANALYSIS	75
2.6. MACHINE LEARNING BASED ANALYSIS AND DEVELOPING CANCER RISK PREDICTION MODEL	76
2.7. CULTURE OF BREAST AND OVARIAN CANCER CELL LINES	84
2.8. GENOTYPING BY THE TAQMAN PCR ASSAY	84
2.9. REAL-TIME PCR ANALYSIS:	85
3. RESULTS AND DISCUSSION:	86
3.1. SELECTION OF LOW FREQUENCY VARIANTS EXCLUSIVE IN BREAST AND OVARIAN CANCER PATIENTS:	86



3.2. ADDRESSING IMBALANCED CASE-CONTROL DATA AND INSIGNIFICANT ODDS RATIOS THROUGH MACHINE LEARNING-BASED APPROACH .....	86
3.3. LNCRNA-SNP MEDIATED EFFECTS ON GENE REGULATION FOR IMPROVED MODEL EFFICIENCY .....	87
3.4. PATHWAY AND GENE SET ENRICHMENT OF THE GENES CONSTITUTING THE FEATURE SET .....	89
3.5. SCREENING THE MOST PROMISING LNCRNA-SNPS FOR WET LAB VALIDATION.....	91
3.6. VALIDATING THE PRESENCE OF THE OVARIAN CANCER SPECIFIC SNP RS9510420 HARBOURED BY THE LNCRNA LINC00621 AND QRT-PCR VALIDATION OF THE CORRESPONDING LNCRNA .....	92
3.7. VALIDATING THE PRESENCE OF THE BREAST CANCER SPECIFIC SNP RS2366152 HARBOURED BY THE LNCRNA HOTAIR AND QRT-PCR VALIDATION OF THE CORRESPONDING LNCRNA .....	93
4. CONCLUSION: .....	93
5. REFERENCES:.....	94
CHAPTER 6  GENERAL DISCUSSION AND FUTURE PERSPECTIVES.....	97
APPENDIX.....	99



## **PREFACE**

Long non-coding RNAs (lncRNAs) form a subgroup of noncoding RNAs (ncRNAs) with a wide range of functions within the cellular system. These molecules, which are typically longer than 200 nucleotides, share some similarities with messenger RNAs (mRNAs), such as possessing a 5' cap, a polyadenylated tail, and multiple exons. However, they lack the coding capacity found in mRNAs. The recent revelation of lncRNAs as regulatory entities in the cellular system has reshaped our understanding of the functional capacity of the genome. This shift in perspective has challenged the notion that a substantial segment of the genome is non-functional "junk." LncRNAs have been experimentally shown to play diverse roles, including transcriptional regulation, X chromosome inactivation, genomic reprogramming, miRNA interaction, nuclear-cytoplasmic transport, involvement in RNA splicing, and participation in apoptosis. However, these known functions represent only the tip of the iceberg.

Single Nucleotide Polymorphisms (SNPs), often pronounced as "snips," are common genetic variations found in the DNA of individuals. SNPs are the most prevalent form of genetic diversity in the human genome and contribute significantly to the variations observed among individuals. They can lead to phenotypic changes and influence susceptibility to diseases, including cancer. Most SNPs are situated in non-coding regions of the genome, which include lncRNAs. SNPs within functional regions of the genome have the potential to be linked with phenotypic alterations and susceptibility to diseases, including various types of cancer. There is substantial evidence in the literature connecting cancer risk to lncRNAs containing SNPs. This association holds true for some of the most commonly occurring female cancers globally, such as breast, cervical, and ovarian cancer.

Several genome-wide association studies (GWAS) have pinpointed genetic variants (SNPs) linked to diseases, including ovarian, breast, and cervical cancer, often in or near lncRNAs. However, such works are limited and the precise mechanistic implications of SNPs within these lncRNAs, such as how they modify lncRNA interactions with their targets, ultimately influencing diverse biological processes and cellular functions, have not been extensively investigated.

This thesis is dedicated to further elucidating the role of GWAS-associated SNPs within lncRNAs in the context of three predominant female cancers: breast, cervical and ovarian cancer. Our goal is to identify shared and distinct functional SNP-containing lncRNAs across different female cancer systems. Our aim is to explore the functional implications of these SNPs and validate their presence within lncRNA loci. By conducting in-depth in-silico analyses and experimental validations, we aspire to elucidate the intricate interplay between genetic variations, lncRNAs, and cancer biology, ultimately advancing our knowledge of these critical female malignancies.

## *Preface*

**Chapter 1**, titled '**A Brief Review**' provides insights into the impact of SNPs located within lncRNA loci on the development of female cancers. It also offers an overview of the international and national research landscape in this field.

**Chapter 2**, titled '**LncRBase V.2: An updated resource for multispecies lncRNAs**' introduces the updated version of the LncRBase database. This version includes a comprehensive collection of lncRNA data from six additional species beyond humans and mice, accompanied by enhanced features of lncRNAs. The chapter details the in-silico methods employed for curating and analyzing lncRNA data.

**Chapter 3**, titled '**ClinicLSNP: a database hosting genetic variants in lncRNAs for cancer patients**,' presents a database integrated into LncRBase V.2. It provides exhaustive information regarding the SNPs present within lncRNA loci by extensive in-silico analysis of patient RNA-seq data; it also provides an assessment of their clinical importance. This resource is an invaluable tool for both researchers and clinicians, offering insights into lncRNA-SNPs specific to breast, cervical, and ovarian cancers, along with their relevant information.

**Chapter 4**, titled '**Shared lncRNA Variants in Female Cancers**,' delves into common SNP-associated lncRNAs shaping the tumor landscapes of three female cancers. The significant lncRNA-SNP uncovered through patient transcriptome data was examined for its role through wet lab validation, revealing potential influence on the upregulated expression of the lncRNA in these cancers.

**Chapter 5**, titled '**Developing lncRNA-SNP based Breast and Ovarian cancer risk prediction models and its effect on gene regulation**,' focuses on gene expression-based Breast and Ovarian Cancer risk probability models for predicting the risk factor, particularly in patients with conditions that heighten cancer susceptibility. By integrating biological knowledge and statistical analysis, exclusive lncRNA-SNP associations and their gene partners are identified for both cancers. Wet lab validation underscores the significance of these previously overlooked lncRNA-SNP associations in disease progression.

**Chapter 6**, titled '**General Discussion and Future Perspectives**,' provides a comprehensive summary of the main finding outlined in the preceding chapters. It also addresses current constraints in the research and outlines potential avenues for future enhancements and developments in the field.

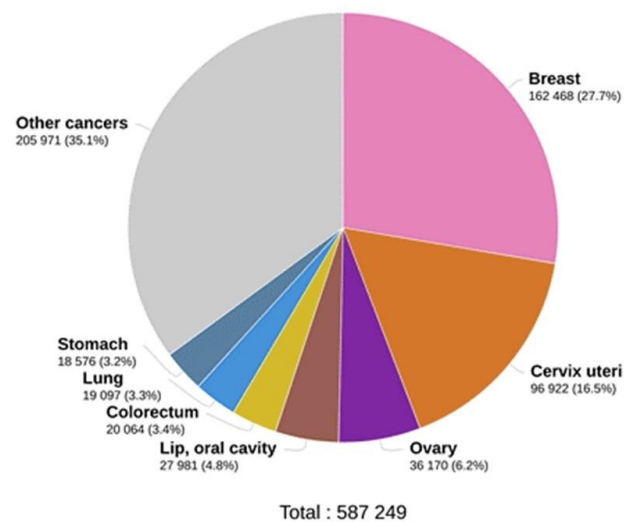
The **Appendix** includes a comprehensive compilation of my publications, featuring reprints of key publication, as well as my curriculum vita

## CHAPTER 1| A BRIEF REVIEW

### 1. Introduction

#### 1.1. Female Cancers: Causes and Concern

In recent years, cancer has risen to become one of the top ten causes of death in India, with more than 800,000 new cases diagnosed annually. According to data from 2009, the country has witnessed nearly 2.5 million cancer cases, resulting in approximately 400,000 cancer-related deaths. The age-adjusted cancer incidence rates vary from 44 to 122 per 100,000 population in males and from 52 to 128 per 100,000 population in females. (Figure1)[1].



**Figure 1:** Top three cancers among women in India are (i) Breast Cancer (ii) Cervical Cancer and (iii) Ovarian Cancer affecting women mostly between age 35-64 and incidence rates vary from 52-128 per 100,000 females. (modified adaptation from <https://www.breastcancerindia.net/>, 2018 survey)

Breast cancer and gynecological cancers are prominent among the leading diseases affecting Indian women, as reported by the National Cancer Registry Program. Specifically, breast, ovarian, and cervical cancers are the most prevalent, primarily affecting women in the age group of 35 to 64. Over the past two decades, the incidence of breast cancer has been increasing in all urban registries across India[2, 3]. It's worth noting that approximately 5% to 10% of breast cancers are attributed to genetic anomalies, with the two breast cancer-associated genes, BRCA1 (Breast Cancer gene 1) and BRCA2 (Breast Cancer gene 2), accounting for a significant portion, possibly up to 10% of all breast cancer cases (source: [www.breastcancer.org](http://www.breastcancer.org)).

Cervical cancer presents a significant health challenge in India, with 122,844 women being diagnosed with the disease each year, resulting in 67,477 deaths[4]. It ranks as the second most common cancer among Indian women and the third most common cancer

# Chapter 1

affecting women globally. [5]. While chronic infection with the human papillomavirus (HPV) serves as the predominant cause of cervical cancer and its precursor lesions, there is notable evidence of familial clustering of cervical carcinoma, similar to several other cancers[6]. Moreover, the observed trend of heightened familial relative risk with closer biological relatedness indicates a significant contribution of genetic factors to the familial clustering of cervical cancer.[7]. Yet, our current comprehension of the genetic foundations of cervical cancer remains restricted.

Ovarian cancer is another cancer type on the rise in India. Between 2001 and 2006, the age-standardized incidence rates for ovarian cancer ranged from 0.9 to 8.4 per 100,000 women in different registries across the country. Women affected by this devastating disease often face challenges in early cancer detection, primarily due to the absence of specific and sensitive biomarkers for ovarian tumors[8].

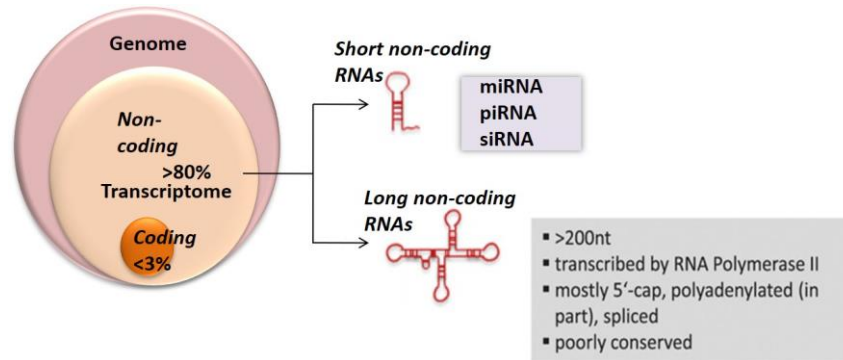
It's important to note that mutations in BRCA1 and BRCA2 genes are also alinked to an elevated risk of ovarian cancer. These genetic mutations contribute to approximately 5-10% of ovarian cancer cases in women (source: <http://www.cancer.org/cancer/ovariancancer/>).

Research has shown shared genetic characteristics between basal-like breast tumors and high-grade serous ovarian tumors[9]. This suggests a related etiology and the potential for similar therapeutic approaches for both of these cancers. However, when it comes to cervical cancer, Human papilloma Virus (HPV) infection is a necessary but not sole risk factor for its development. Studies conducted by Wang S et al, Madeleine MM et al, and Hosono S et al. have conducted research to examine the significance of genetic variations in cervical cancer using both candidate gene association studies and genome-wide association studies. These studies have established a connection between the genetic makeup of disease loci and susceptibility to malignancy development in cervical cancer[10-12]. As a result, there remains an open question regarding whether these gynecological malignancies and breast cancer share a network of common cellular events guided by genetic variations within the genome.

## 1.2. Long Non-Coding RNA (lncRNA)

Long non-coding RNAs (lncRNAs) are RNA transcripts characterized by their length, typically exceeding 200 nucleotides, and their lack of involvement in protein synthesis as templates. Many lncRNAs are transcribed by RNA polymerase II, undergo splicing, and polyadenylation[13]. These molecules typically exhibit lower expression levels in comparison to protein-coding transcripts and are primarily found within the nucleus, in contrast to mRNAs, which are more abundant and primarily located in the cytoplasm[14]. While lncRNAs were once considered to be non-functional "junk" RNA, they are now recognized for their diverse roles in various cellular processes. These roles include the regulation of gene expression, involvement in genomic reprogramming, participation in X-chromosome inactivation, contribution to genomic imprinting, influence on nuclear compartmentalization, facilitation of nuclear-cytoplasmic trafficking, involvement in

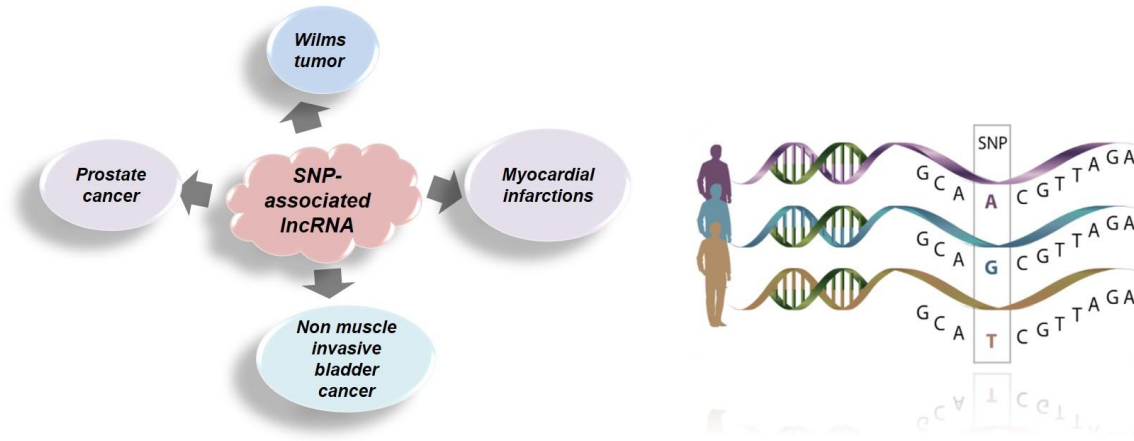
RNA splicing, and control of critical cellular processes such as the cell cycle and apoptosis, among others[15] (**Figure 2**). LncRNAs constitute a substantial portion of noncoding genes in mammals and other eukaryotes, playing diverse roles in cellular systems including cancer[16, 17].



**Figure 2:** The non coding world: A significant portion of the transcriptome is made up of long non coding RNAs (lncRNAs) characterized by greater than 200 nucleotides in length, 5' capped, polyadenylated and poorly conserved across species.

### 1.3. Genetic variants within the lncRNA loci

Single Nucleotide Polymorphisms (SNPs) represent the most prevalent genetic variations within the human genome, often located in functional regions that could influence phenotypic traits and susceptibility to diseases, including cancer.(**Figure 3**)[18, 19]. A significant proportion of SNPs are located in non-coding regions of the genome[20], suggesting their regulatory roles in disease outcomes[21]. These non-coding regions include long intergenic non-coding RNAs (lincRNAs), a subclass of long noncoding RNAs (lncRNAs)[22]. A large majority of single nucleotide polymorphisms (SNPs) identified in genome-wide association studies (GWAS) are located in intergenic and intronic regions[23]. This observation underscores the crucial role of non-coding regions of the genome in predisposition to diseases. Some GWAS SNPs have been shown to have effects on the expression of the genes transcribed from the nearby or distant loci, known as expression Quantitative Trait Loci or eQTL [20]. Moreover, lncRNAs containing SNPs can experience changes in their functions, potentially leading to phenotypic alterations. Growing evidence suggests that SNP-associated lncRNAs are linked to a wide range of disease phenotypes, including Wilms' tumors[24], myocardial infarction[25], bladder cancer[26], prostate cancer[27], and more. Since the structural features of lncRNAs are crucial for their functions, SNPs within lncRNA transcripts may impact the secondary structure of these molecules, affecting their stability and functions. This, in turn, can disrupt lncRNA interactions with target genes, ultimately influencing disease phenotypes[28].



**Figure 3:** SNP associated LncRNAs related to different diseases including cancer

## 1.4. Association between SNPs and cancer risk

SNPs are determined through the genotyping of groups of individuals, and subsequently their association with specific traits is examined. These association rates are determined through multivariate logistic regression analysis, yielding the odds ratio (OR) accompanied by corresponding confidence intervals (95% CIs). The OR, within the framework of case-control studies, quantifies the ratio of two odds: firstly (as per Equation 1), the likelihood of the allele's occurrence alongside a specific SNP contributing to tumor development.

**Equation 1:**

$$Odds\ ratio = \frac{p(\frac{cancer}{allele} - A) / (1 - p(\frac{cancer}{allele} - A))}{(\frac{cancer}{allele} - B) / (1 - p(\frac{cancer}{allele} - B))}$$

and second (Equation 2), the likelihood of developing a cancer disorder in the presence of the alternative allele.

**Equation 2:**

$$Odds\ ratio = \frac{Odds(A)}{Odds(B)}$$

The odds (as per Equation 3) are defined as the ratio of the probability of developing cancer to the probability of not developing cancer.

**Equation 3:**



$$Odds = \frac{p}{1 - p}$$

An OR value greater than 1 indicates that allele A is a risk factor for cancer development, while an OR value less than 1 suggests that allele A provides protection against tumors[29].

## 2. Review of status of Research and Development in the subject

### 2.1. International Status:

Numerous GWAS have identified disease-associated genetic variants, specifically SNPs, situated within or in close proximity to lncRNAs. For instance, Kumar et al. investigated the association between SNPs and long intergenic RNA (lincRNA) expression levels in various human tissues and found strong genotype-lincRNA expression correlations, often linked to disease- or trait-associated SNPs[21].

The chromosomal region 9p21, once referred to as a "gene desert," contains the lncRNA ANRIL, and SNPs within or near ANRIL have been linked to atherosclerotic vascular disease susceptibility, coronary disease, intracranial aneurysms, and type 2 diabetes[30, 31]. The ANRIL locus contains three critical coding genes (CDKN2A, CDKN2B, and ARF) that play a crucial role in the retinoblastoma (RB) and p53 tumor suppressor networks[32]. It is widely acknowledged that SNPs in the upstream region of CDKN2A and CDKN2B predominantly influence the function of lncRNA ANRIL[31]. These disease-associated SNPs have been shown to impact ANRIL lncRNA expression[33] and affect a repressive STAT1 binding site near the ANRIL locus[34]. Consequently, derepressed ANRIL negatively regulates the neighboring CDKNB gene[34, 35] through binding to CBX7, a constituent of the PRC, they induce the suppression of CDKN2B expression in a manner dependent on PRC1 and histone H3K27me [35]. In the context of coronary artery disease and its inflammatory component, ANRIL SNPs associated with this condition have been found to elevate ANRIL expression in response to pro-inflammatory stimuli like interferon-gamma (IFN $\gamma$ ) or TNF $\alpha$ [34, 36]. In such circumstances, ANRIL engages with the YY1 transcription factor to enhance the expression of crucial mediators involved in the inflammatory response[36]. **(Figure 4a).**

The CCAT2 gene region contains a SNP associated with colorectal cancer, specifically rs6983267. Mechanistically, CCAT2 binds to TCFL2, leading to the upregulation of WNT signaling pathway genes[37]. CCAT2 itself is a target of the WNT signaling pathway, creating a positive feedback loop that amplifies pathway activity, ultimately resulting in the increased expression of the MYC proto-oncogene[37]. Additionally, the CCAT2 SNP region, including rs6983267, has been reported to interact with the MYC promoter, suggesting that CCAT2 may directly influence MYC expression by acting as an enhancer-like element[38]. The G-allele of the CCAT2 rs6983267 SNP has

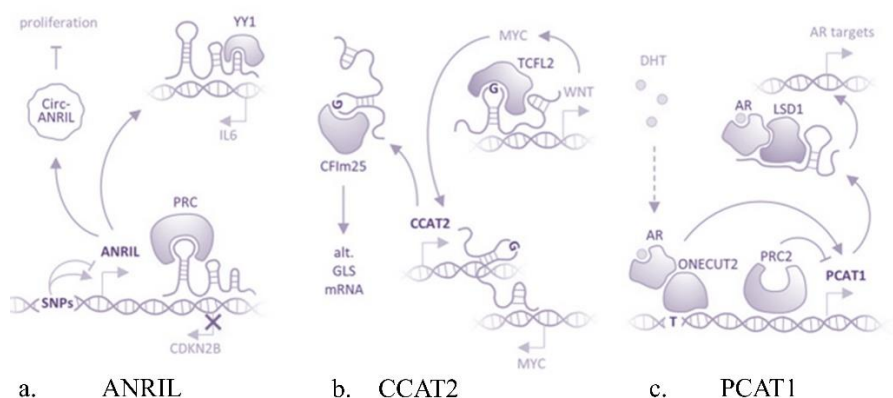
## Chapter 1

additionally demonstrated a preference for binding to the CFIm25 subunit of the cleavage factor I (CFIm) complex, thereby promoting the alternative splicing of Glutaminase (GLS) mRNA. This alternative splicing event contributes to cancer cell metabolism and proliferation[39]( **Figure 4b**).

SNP rs7463708 located in a PCAT1 enhancer region, has been found to increase the binding of the androgen receptor (AR) in conjunction with the ONECUT2 transcription factor. This enhanced binding leads to elevated PCAT1 expression[40]. PCAT1 lncRNA has also been shown to interact with an AR/LSD1 complex, promoting the expression of androgen-stimulated genes involved in prostate cancer progression[40](**Figure 4c**). The risk variant rs72725854 (A > T) is located within a prostate-specific enhancer and has been associated with increased risk. This variant promotes the expression of PCAT1 through the recruitment of the SPDEF transcription factor[41].

The rs7763881 SNP within the HULC locus has been linked to a decreased risk of developing hepatocellular carcinoma in patients with persistent hepatitis B virus (HBV) infection[42]. Similarly, in studies related to colorectal and esophageal cancer, this SNP, rs7763881, has shown a protective effect[43, 44]. Although the association of the lncRNA HULC with cancer is well-established, the exact mechanistic role of the rs7763881 SNP in HULC's function is yet to be fully understood.

Additionally, Jin et al. identified an lncRNA-related SNP (rs3787016) associated with prostate cancer risk[45]. Similarly, in papillary thyroid carcinoma (PTC), a SNP (rs944289) located near the lincRNA PTCSC3 may influence its expression and thus contribute to PTC susceptibility[46].



**Figure 4:** The mechanistic effect of the presence of SNPs in cancer progression in (a). ANRIL (b.) CCAT2 and (c.) PCAT1. (Modified adaptation from Aznaourova et.al)

In the context of female reproductive system health, the **HOXA** region, which regulates embryogenesis and **ovarian carcinogenesis**, contains lncRNAs such as HOXA10-AS, HOXA11-AS, and HOTTIP. In serous epithelial ovarian cancer (EOC), an SNP within HOXA11-AS (rs17427875) has been linked to reduce EOC risk, suggesting a potential tumor suppressor role for this lncRNA[47]. **HOTAIR** has been implicated in elevating

the risk of **breast cancer**[48]. In an association study conducted on a female breast cancer patient, three specific SNPs were identified, having associations with cancer risk. Specifically, rs920778 and rs12826786 were found to elevate the breast cancer risk, while a negative correlation was observed for the rs1899663 SNP[49].

Moreover, certain SNPs within lncRNAs can act as expression Quantitative Trait Loci (eQTLs) for protein-coding genes. For example, SNPs within an lncRNA AC008392.1. can serve as eQTLs for **CARD8**, a gene associated with immune responses and apoptosis, potentially influencing virus-induced **cervical cancer** risk[50]. Similarly, **ZNRD1-AS1**, an lncRNA antisense to ZNRD1 involved in immune responses against HPV infection and **cervical cancer**, harbors several SNPs. Some of these SNPs (rs3757328, rs6940552, and rs9261204) have been associated with a decreased risk of cervical cancer[51]. SNP rs6983267 in the **CCAT2** locus has been linked to heightened MYC expression and is thought to contribute to the promotion of proliferation and rapid growth of cervical **squamous cell carcinoma** (SCC) compared to lower-grade carcinomas.[52]. This genetic variation can contribute to the more aggressive behaviour observed in some cervical SCC cases[52]. A comprehensive list of SNPs mapped within lncRNA loci in Breast, Ovarian and Cervical Cancer as mentioned in literature has been catalogued in **Table1.1**

While databases cataloguing SNP associations with lncRNA loci exist, they often lack detailed mechanistic insights into how these SNPs modify lncRNA interactions with their targets, thereby influencing various biological processes and cellular functions. Furthermore, these databases may not fully address the impact of lncRNA-SNP on female malignancies.

<b>lncRNA</b>	<b>SNP</b>	<b>Ref/alt allele</b>	<b>Cancer</b>
<b>Hoxa11-AS</b>	rs17427875	A>T	Ovarian Cancer
<b>HOTAIR</b>	rs1899663	G>T	
	rs4759314	A>G	Ovarian Cancer[53]
<b>HOTAIR</b>	rs920778	T>C	
	rs12826786	C>T	Breast Cancer
	rs1899663	G>T	
<b>ANRIL</b>	rs1333045	T>C	Breast Cancer[54]
<b>LINC00520</b>	rs11622641	C>G	
	rs12880540 rs2152278	T>C	Breast Cancer[55]
		C>A	
<b>AQP4-AS1</b>	rs527616	C>G	Breast Cancer[56]
<b>SOX2OT</b>	rs9839776	C>G	Breast Cancer[57]
<b>H19</b>	rs3741219	A>G	
	rs217727	G>A	Breast Cancer[58]
	rs2839698	G>A	
	rs3741216	T>A	
<b>ZNRD1-AS1</b>	rs3757328	G>A	Cervical Cancer
	rs6940552	G>A	
	rs9261204	A>G	
<b>CCAT2</b>	rs6983267	G>T	Cervical cancer
<b>CARD8</b>	rs7248320	G>A	Cervical Cancer

**Table1.1** Mapped genome-wide association studies (GWAS) reporting disease-associated genetic variants (SNPs) to, or in, the vicinity of lncRNAs in ovarian, breast as well as in cervical cancer.

# Chapter 1

## 2.2. National Status:

Siva Sankar et al. conducted a bioinformatic analysis [59] focusing on SNPs associated with breast cancer. Their study involved gathering data from various databases, including Online Mendelian Inheritance in Man (OMIM), Entrez Genome View, and the Cancer Genome Anatomy Project (NCBI). They aimed to enumerate and locate SNPs within chromosomes associated with both cancerous and non-cancerous breast tissue. The analysis also explored DNA methylation patterns and mutations in proto-oncogene hotspots implicated in breast carcinoma formation[59].

Kapur et al. explored genomic changes in breast cancer patients with a history of tobacco exposure. Their research identified copy number alterations in multiple chromosomes among these cancer patients[60].

Vinod et al. examined the role of the IL10 (-1082A/G) gene promoter polymorphism in breast cancer patients from South India. They observed a significant association between the AA genotype of the IL-10 -1082A/G polymorphism and breast cancer[61].

Tulsyan et al. explored how CD44 gene polymorphisms affect the risk and prognosis of breast cancer in the North Indian population. Their findings suggested that CD44 rs353639 could have a notable effect on breast cancer progression[62].

Mitra et al. investigated the correlation between the non-synonymous SNP rs1052133 (C8069G/Ser326Cys) situated in the exonic region of the human 8-oxoguanine DNA glycosylase (hOGG1) gene and the susceptibility to squamous cell carcinomas of the head and neck (SCCHN) in samples from the North Indian population. Their research revealed a strong association between the rs1052133 polymorphism and SCCHN susceptibility, with the mutant (G) allele potentially acting as a protective factor among North Indian subpopulations[63].

Additionally, Bhartiya et al. contributed by establishing a comprehensive database of human lncRNAs and characterizing their genomic and functional context within the cellular system. They mapped a substantial number of genomic variations (from the database dbSNP <https://www.ncbi.nlm.nih.gov/snp/>) to lncRNAs. Furthermore, they provided conservation scores for numerous sites within these lncRNAs[64].

Considering the national landscape, there appears to be a gap in detailed studies concerning the association of SNPs with lncRNAs and their functional roles in cancer. This gap underscores the importance of including such factors in genetic analyses of cancer genomes, particularly in breast, cervical, and ovarian cancers, which significantly contribute to cancer-related morbidity and mortality among women in the Indian subcontinent.

### 3. Consequence of the presence of SNP within lncRNA

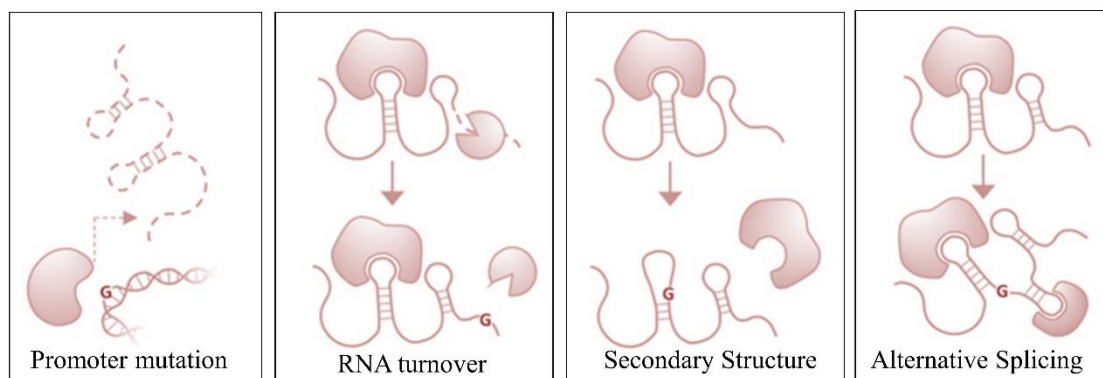
Several lncRNAs containing gene variants associated with diseases have been proposed as potential clinical markers, aiding in the stratification and prognosis of cancer patients, among other applications[65, 66]. Additionally, non-coding SNPs have been observed to exhibit a significant concentration within promoters, including DNaseI hypersensitive sites (HSSs), which are indicative of active transcription and are essential for the expression of lncRNA genes[23, 67]. Hence, recent findings substantiate the notion that SNPs can impact the expression of lncRNAs associated with diseases by altering regulatory DNA regions, such as transcription factor binding sites. [68].

Additionally, it is conceivable that nucleotide rearrangements in disease-relevant lncRNA loci, such as alterations in secondary structures or alternative splicing patterns can result from these SNPs. SNP located within lncRNA secondary structures, such as hairpin loops, have the potential to modify RNA folding, thereby influencing its interaction with other biomolecules

Furthermore, it is conceivable that rearrangements of nucleotide such as secondary structure alternation or alternative splicing patterns within lncRNA loci, can arise from these SNPs. SNPs situated within the hairpin loops, a form of secondary structure possess the potential to alter RNA folding, thereby impacting its interaction with other biomolecules.[69].

lncRNA-SNPs can also impact splice sites, leading to the generation of alternative splice variants with modified functionality and influencing their interactions with target proteins or transcripts[69]. **Figure 5** depicts potential outcomes and effects of SNPs) within lncRNA sequences.

Although there has been an abundance of recently published literature on lncRNAs, encompassing GWAS and investigations into prognostic markers, the practical significance of lncRNAs in human diseases, however, continues to be a subject of debate. This uncertainty stems from conflicting findings obtained through studies conducted in cell culture, small animal models, and clinical research.



# Chapter 1

**Figure 5:** Potential outcomes of SNPs occurring within lncRNAs. (Modified adaptation from Aznaourova et.al)

## 4. Importance of the work in the context of current status

The incidence of cancer in India is escalating due to a burgeoning population and prolonged life expectancy. Breast cancer affects 17% of the global population in India, with the country boasting the highest age-standardized incidence of cervical cancer in South Asia. Indian women face significant risks of cervical cancer, with a 2.5% cumulative lifetime risk and 1.4% cumulative death risk from this disease. Ovarian cancer, frequently identified at later stages has emerged to be prevalently malignant among women in India, leading to a grim prognosis. This is especially concerning during a woman's reproductive years.

These cancers are often influenced by the cumulative effect of numerous low-risk gene variants. SNPs, which involve single base mutations, play a crucial role in identifying gene mutations and susceptibility to cancer-causing factors. GWAS have pinpointed thousands of disease risk-associated SNPs, among which 75 are linked to breast cancer risk and 42 to cervical cancer risk.

Many of these risk-associated SNPs map to non-coding regions of the genome, indicating their regulatory roles. Approximately 7% of SNPs associated with autoimmune diseases are estimated to map to long intergenic non-coding RNAs (lincRNAs), a subclass of lncRNAs. LncRNAs are believed to play a role in regulating the expression of protein-coding genes. Consequently, lncRNA-SNPs may indirectly impact protein expression, thereby influencing disease outcomes.

Given the prominence of breast, cervical, and ovarian cancers in Indian women and their significant impact on the country's socio-economic conditions, there is an urgent need to revise existing therapeutic approaches for cancer detection, prognosis, and treatment. To address this, it is vital to conduct a comprehensive study of SNP-associated lncRNA loci in these cancers and examine their role in modulating lncRNA-target interactions, ultimately influencing disease phenotypes.

The significant knowledge gap that remains between the association of SNPs and the underlying molecular mechanisms contributing to disease risk presents both a challenge and a potential avenue for discovering lncRNAs that play pivotal roles in cancer. A genetic variation within a regulatory element has the potential to alter the abundance of a gene transcript. However, it's worth noting that a substantial proportion of SNPs reside in intergenic or intronic regions, and their connection to lncRNA function within the cellular context may be more intricate and less straightforward.

The work presented in this thesis is an effort in adding further understanding of the role of SNP associated lncRNAs in female cancers. Furthermore, this systematic study of

SNP-associated lncRNAs in breast, cervical, and ovarian cancer will shed light on the genetic architecture and potential similarities between these three disease systems. Identifying candidate lncRNA-SNPs and their target genes can help establish high-risk groups and enable early diagnosis, potentially improving the survival rates of patients facing these prevalent malignancies in the breast, cervix, and ovary.

## 5. References

1. Uma Devi, K., *Current status of gynecological cancer care in India*. J Gynecol Oncol. 2009 Jun;20(2):77-80. doi: 10.3802/jgo.2009.20.2.77. Epub 2009 Jun 29.
2. Badwe, R.A., et al., *Cancer incidence trends in India*. Jpn J Clin Oncol, 2014. **44**(5): p. 401-7.
3. Kamath, R., et al., *A study on risk factors of breast cancer among patients attending the tertiary care hospital, in udupi district*. Indian J Community Med, 2013. **38**(2): p. 95-9.
4. Sreedevi, A., R. Javed, and A. Dinesh, *Epidemiology of cervical cancer with special focus on India*. Int J Womens Health, 2015. **7**: p. 405-14.
5. Arbyn, M., et al., *Worldwide burden of cervical cancer in 2008*. Ann Oncol, 2011. **22**(12): p. 2675-2686.
6. Magnusson, P.K., P. Lichtenstein, and U.B. Gyllensten, *Heritability of cervical tumours*. Int J Cancer, 2000. **88**(5): p. 698-701.
7. Magnusson, P.K., P. Sparén, and U.B. Gyllensten, *Genetic link to cervical tumours*. Nature, 1999. **400**(6739): p. 29-30.
8. Saini, S., et al., *Epidemiology of epithelial ovarian cancer, a single institution-based study in India*. Clinical Cancer Investigation Journal, 2016. **5**(1): p. 20-24.
9. *Comprehensive molecular portraits of human breast tumours*. Nature, 2012. **490**(7418): p. 61-70.
10. Wang, S., et al., *Association of 42 SNPs with genetic risk for cervical cancer: an extensive meta-analysis*. BMC Med Genet, 2015. **16**: p. 25.
11. Madeleine, M.M., et al., *Comprehensive analysis of HLA-A, HLA-B, HLA-C, HLA-DRB1, and HLA-DQB1 loci and squamous cell cervical cancer risk*. Cancer Res, 2008. **68**(9): p. 3532-9.
12. Hosono, S., et al., *HLA-A alleles and the risk of cervical squamous cell carcinoma in Japanese women*. J Epidemiol, 2010. **20**(4): p. 295-301.
13. Mercer, T.R., M.E. Dinger, and J.S. Mattick, *Long non-coding RNAs: insights into functions*. Nat Rev Genet, 2009. **10**(3): p. 155-9.
14. Rashid, F., A. Shah, and G. Shan, *Long Non-coding RNAs in the Cytoplasm*. Genomics, Proteomics & Bioinformatics, 2016. **14**(2): p. 73-80.
15. Fang, Y. and M.J. Fullwood, *Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer*. Genomics, Proteomics & Bioinformatics, 2016. **14**(1): p. 42-54.
16. Wilusz, J.E., H. Sunwoo, and D.L. Spector, *Long noncoding RNAs: functional surprises from the RNA world*. Genes Dev, 2009. **23**(13): p. 1494-504.
17. Rinn, J.L. and H.Y. Chang, *Genome regulation by long noncoding RNAs*. Annu Rev Biochem, 2012. **81**: p. 145-66.
18. Reich, D.E., S.B. Gabriel, and D. Altshuler, *Quality and completeness of SNP databases*. Nat Genet, 2003. **33**(4): p. 457-8.
19. Erichsen, H.C. and S.J. Chanock, *SNPs in cancer research and treatment*. Br J Cancer, 2004. **90**(4): p. 747-51.

## Chapter I

20. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits*. Nat Rev Genet, 2009. **10**(4): p. 241-51.
21. Kumar, V., et al., *Human disease-associated genetic variation impacts large intergenic non-coding RNA expression*. PLoS Genet, 2013. **9**(1): p. e1003201.
22. Ricaño-Ponce, I. and C. Wijmenga, *Mapping of immune-mediated disease genes*. Annu Rev Genomics Hum Genet, 2013. **14**: p. 325-53.
23. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits*. Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.
24. Li, W., et al., *H19 gene polymorphisms and Wilms tumor risk in Chinese children: a four-center case-control study*. Mol Genet Genomic Med, 2021. **9**(2): p. e1584.
25. Ma, R., et al., *Promoter polymorphisms in the lncRNA-MIAT gene associated with acute myocardial infarction in Chinese Han population: a case-control study*. Biosci Rep, 2020. **40**(2).
26. Liu, H., et al., *LncRNA BCLET variant confers bladder cancer susceptibility through alternative splicing of MSANTD2 exon 1*. Cancer Med, 2023. **12**(13): p. 14440-14451.
27. Anil, P., S. Ghosh Dastidar, and S. Banerjee, *Unravelling the role of long non-coding RNAs in prostate carcinoma*. Advances in Cancer Biology - Metastasis, 2022. **6**: p. 100067.
28. Gong, J., et al., *lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse*. Nucleic Acids Res, 2015. **43**(Database issue): p. D181-6.
29. Minotti, L., et al., *SNPs and Somatic Mutation on Long Non-Coding RNA: New Frontier in the Cancer Studies? High Throughput*, 2018. **7**(4).
30. Burd, C.E., et al., *Expression of Linear and Novel Circular Forms of an INK4/ARF-Associated Non-Coding RNA Correlates with Atherosclerosis Risk*. PLOS Genetics, 2010. **6**(12): p. e1001233.
31. Pasmant, E., et al., *ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS*. FASEB J, 2011. **25**(2): p. 444-8.
32. Sherr, C.J., *Ink4-Arf locus in cancer and aging*. Wiley Interdiscip Rev Dev Biol, 2012. **1**(5): p. 731-41.
33. Cunningham, M.S., et al., *Chromosome 9p21 SNPs Associated with Multiple Disease Phenotypes Correlate with ANRIL Expression*. PLoS Genet, 2010. **6**(4): p. e1000899.
34. Harismendy, O., et al., *9p21 DNA variants associated with coronary artery disease impair interferon- $\gamma$  signalling response*. Nature, 2011. **470**(7333): p. 264-8.
35. Yap, K.L., et al., *Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a*. Mol Cell, 2010. **38**(5): p. 662-74.
36. Zhou, X., et al., *Long non-coding RNA ANRIL regulates inflammatory responses as a novel component of NF- $\kappa$ B pathway*. RNA Biol, 2016. **13**(1): p. 98-108.
37. Ling, H., et al., *CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer*. Genome Res, 2013. **23**(9): p. 1446-61.
38. Pomerantz, M.M., et al., *The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer*. Nat Genet, 2009. **41**(8): p. 882-4.
39. Redis, R.S., et al., *Allele-Specific Reprogramming of Cancer Metabolism by the Long Non-coding RNA CCAT2*. Mol Cell, 2016. **61**(4): p. 520-534.
40. Guo, H., et al., *Modulation of long noncoding RNAs by risk SNPs underlying genetic predispositions to prostate cancer*. Nat Genet, 2016. **48**(10): p. 1142-50.



41. Walavalkar, K., et al., *A rare variant of African ancestry activates 8q24 lncRNA hub by modulating cancer associated enhancer*. Nat Commun, 2020. **11**(1): p. 3598.
42. Liu, Y., et al., *A genetic variant in long non-coding RNA HULC contributes to risk of HBV-related hepatocellular carcinoma in a Chinese population*. PLoS One, 2012. **7**(4): p. e35145.
43. Kang, M., et al., *Long noncoding RNAs POLR2E rs3787016 C/T and HULC rs7763881 A/C polymorphisms are associated with decreased risk of esophageal cancer*. Tumour Biol, 2015. **36**(8): p. 6401-8.
44. Shaker, O.G., M.A. Senousy, and E.M. Elbaz, *Association of rs6983267 at 8q24, HULC rs7763881 polymorphisms and serum lncRNAs CCAT2 and HULC with colorectal cancer in Egyptian patients*. Sci Rep, 2017. **7**(1): p. 16246.
45. Jin, G., et al., *Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk*. Carcinogenesis, 2011. **32**(11): p. 1655-9.
46. Jendrzewski, J., et al., *The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type*. Proc Natl Acad Sci U S A, 2012. **109**(22): p. 8646-51.
47. Liu, Z., et al., *Over-expressed long noncoding RNA HOXA11-AS promotes cell cycle progression and metastasis in gastric cancer*. Mol Cancer, 2017. **16**(1): p. 82.
48. Cantile, M., et al., *Long Non-Coding RNA HOTAIR in Breast Cancer Therapy*. Cancers (Basel), 2020. **12**(5).
49. Hassanzarei, S., et al., *Genetic polymorphisms of HOTAIR gene are associated with the risk of breast cancer in a sample of southeast Iranian population*. Tumour Biol, 2017. **39**(10): p. 1010428317727539.
50. Yin, J., et al., *Expression Quantitative Trait Loci for CARD8 Contributes to Risk of Two Infection-Related Cancers--Hepatocellular Carcinoma and Cervical Cancer*. PLoS One, 2015. **10**(7): p. e0132352.
51. Guo, L., et al., *Expression quantitative trait loci in long non-coding RNA ZNRD1-AS1 influence cervical cancer development*. Am J Cancer Res, 2015. **5**(7): p. 2301-7.
52. Łażniak, S., et al., *The association of CCAT2 rs6983267 SNP with MYC expression and progression of uterine cervical cancer in the Polish population*. Arch Gynecol Obstet, 2018. **297**(5): p. 1285-1292.
53. Saeedi, N. and S. Ghorbian, *Analysis of clinical important of LncRNA-HOTAIR gene variations and ovarian cancer susceptibility*. Molecular Biology Reports, 2020. **47**(10): p. 7421-7427.
54. Khorshidi, H.R., et al., *ANRIL Genetic Variants in Iranian Breast Cancer Patients*. Cell J, 2017. **19**(Suppl 1): p. 72-78.
55. Guo, Q., et al., *Characterization of lncRNA LINC00520 and functional polymorphisms associated with breast cancer susceptibility in Chinese Han population*. Cancer Med, 2020. **9**(6): p. 2252-2268.
56. Marchi, R.D., et al., *Association between SNP rs527616 in lncRNA AQP4-AS1 and susceptibility to breast cancer in a southern Brazilian population*. Genet Mol Biol, 2021. **44**(1): p. e20200216.
57. Tang, X., et al., *Correlations between lncRNA-SOX2OT polymorphism and susceptibility to breast cancer in a Chinese population*. Biomarkers in Medicine, 2017. **11**(3): p. 277-284.
58. Hassanzarei, S., et al., *Genetic polymorphisms in long noncoding RNA H19 are associated with breast cancer susceptibility in Iranian population*. Meta Gene, 2017. **14**: p. 1-5.

## Chapter I

59. SK, S., *Genomic analysis of SNPs in breast cancer by using bioinformatics databases*. Indian Journal of Biotechnology 2007.
60. Saxena, S., et al., *Genomic alterations in breast cancer patients from Northeast India using 10K SNP arrays*. Genome Biol. 2010;11(Suppl 1):P34. doi: 10.1186/gb-2010-11-s1-p34. Epub 2010 Oct 11.
61. Vinod, C., et al., *A Common SNP of IL-10 (-1082A/G) is Associated With Increased Risk of Premenopausal Breast Cancer in South Indian Women*. Iran J Cancer Prev, 2015. **8**(4): p. e3434.
62. Tulsyan, S., et al., *CD44 Gene Polymorphisms in Breast Cancer Risk and Prognosis: A Study in North Indian Population*. PLOS ONE, 2013. **8**(8): p. e71073.
63. Mitra, A.K., et al., *Protective association exhibited by the single nucleotide polymorphism (SNP) rs1052133 in the gene human 8-oxoguanine DNA glycosylase (hOGG1) with the risk of squamous cell carcinomas of the head & neck (SCCHN) among north Indians*. Indian J Med Res, 2011. **133**(6): p. 605-12.
64. Bhartiya, D., et al., *lncRNome: a comprehensive knowledgebase of human long noncoding RNAs*. Database (Oxford), 2013. **2013**: p. bat034.
65. Zhang, X., M.H. Hamblin, and K.J. Yin, *The long noncoding RNA Malat1: Its physiological and pathophysiological functions*. RNA Biol, 2017. **14**(12): p. 1705-1714.
66. Arriaga-Canon, C., et al., *The use of long non-coding RNAs as prognostic biomarkers and therapeutic targets in prostate cancer*. Oncotarget, 2018. **9**(29): p. 20872-20890.
67. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA*. Science, 2012. **337**(6099): p. 1190-5.
68. Kulkarni, S., et al., *CCR5AS lncRNA variation differentially regulates CCR5, influencing HIV disease outcome*. Nat Immunol, 2019. **20**(7): p. 824-834.
69. Aznaourova, M., et al., *Disease-Causing Mutations and Rearrangements in Long Non-coding RNA Gene Loci*. Front Genet, 2020. **11**: p. 527484.

## CHAPTER 2| LncRBase V.2: An updated resource for multispecies lncRNAs

### Abstract:

Long non-coding RNAs (lncRNAs), exceeding 200 nucleotides, play diverse roles in cellular functions. Next-generation sequencing has identified numerous lncRNAs, cataloged in database. Despite these resources, certain aspects, such as lncRNA associations with other noncoding RNAs and genomic elements, remain underexplored. LncRBase V.2 builds on its predecessor to address these gaps. This updated version predicts transcription factor binding sites, offers insights into upstream regulatory influences, predicts subcellular localization, and maps small open reading frames within lncRNA loci. With an extended analysis to six species, disease associations, and tissue-specific expression data, LncRBase V.2 is a comprehensive resource, accessible at <http://dibresources.jcbose.ac.in/zhumur/lncrbase2/>.

### 1. Introduction

Long non-coding RNAs (lncRNAs) form a subgroup of noncoding RNAs (ncRNAs) that participate in various cellular functions [1, 2]. These molecules typically exceed 200 nucleotides in length and often share similarities with messenger RNAs (mRNAs) in terms of their biogenesis, including features such as a 5' cap, polyA tail, and multiple exons, although they lack protein-coding capacity [3]. The recent revelation of lncRNAs as regulatory entities has revolutionized our understanding of the genome's functional potential, overturning the previous notion that a significant portion of it was non-functional "junk." Experimental evidence has affirmed their diverse roles, including gene expression regulation, X-chromosome inactivation, genomic reprogramming, miRNA interactions, nuclear-cytoplasmic transport, RNA splicing, and apoptosis [4] has been experimentally proved. However, this represents only the tip of the iceberg.

The advent of next-generation sequencing technology has led to the rapid identification of an increasing number of lncRNAs. Experimental validation through traditional wet bench techniques has corroborated these in silico discoveries. Several existing databases, such as Ensembl[5], NONCODE[6], Refseq[7], Lncipedia[8] have provided information on lncRNAs from various species. Ensembl, for instance, employs a gene build pipeline to classify lncRNA transcripts into different biotypes. NONCODE offers details on the length, sequence, coding potential, and expression profiles of human and mouse lncRNAs, as well as predictions of the secondary structure of human lncRNAs. LNCipedia records lncRNA isoform names, genomic locations, coding potential scores, and locus conservation for lncRNA transcripts. Several specialized databases are available which offer comprehensive information on specific aspects of lncRNAs. For instance, CHIPBase v2.0[9] primarily focuses on categorizing regulatory molecules found within the promoter region of lncRNA transcripts listed only in the Ensembl

## Chapter2

database(<https://www.ensembl.org/>), so it may not encompass the full spectrum of lncRNA features or regulatory mechanisms. Other specialized databases catalog experimentally validated associations between lncRNAs and diseases, such as Lnc2Cancer[10], LncRNADisease2[11] and Lnc2Atlas[12].

Despite these extensive efforts in lncRNA research, certain areas have remained relatively unexplored. These include investigating lncRNA associations with other ncRNA molecules or their overlap with genomic elements that could serve as intrinsic regulatory sites. The original intent behind publishing the first version of LncRBase[13] was to delve deeper into these less-explored aspects of the lncRNA regulome. In this updated version, LncRBase V.2, we aim to expand our investigations by studying their upstream regulators through the prediction of transcription factor binding sites (TFBS) in the lncRNA promoter regions. This approach can provide valuable insights into the molecules that govern lncRNA expression. Additionally, given recent evidence highlighting the influence of subcellular localization patterns on lncRNA functionality[14], we have incorporated predictions of subcellular localization, shedding light on the activating and deactivating functions of lncRNAs. Furthermore, ribosome footprinting assays have hinted at the potential micro-peptide coding capability of some lncRNAs [14]. In response, we have mapped possible small open reading frames (sORFs) within lncRNA loci based on footprinting studies, offering a more definitive means of discerning whether a given molecule is putatively coding or noncoding.

The updated LncRBase V.2 represents a significant upgrade over its predecessor, LncRBase[13], having a range of distinctive features, as detailed in **Table 1**: (i) We have extended our analysis to encompass an additional six species, broadening our scope beyond Human and Mouse. (ii) A key advancement lies in our ability to unravel the regulatory mechanisms governing lncRNAs. By predicting TFBS within the lncRNA promoter regions, we shed light on their upstream regulatory influences. (iii) Building on mounting evidence highlighting the impact of subcellular localization patterns on lncRNA functionality[15] we have ventured into the prediction of lncRNA localization. This aspect holds sway over the activation and deactivation functions of lncRNAs.(iv) Leveraging insights from ribosome footprinting assays that suggest the coding potential of certain noncoding molecules [16]. We have meticulously mapped potential sORFs within lncRNA loci based on footprinting studies. This approach offers a more robust means of confirming the coding or noncoding status of these molecules[17]. (v) As an added resource, we have curated comprehensive information on interacting target molecules and the involvement of lncRNA genes in various diseases. (vi) Furthermore, LncRBase V.2 serves as a readily accessible repository of tissue-specific lncRNA expression data across multiple species.

Database content	LncRBase	LncRBase V.2
Number of Species	2 (Human, Mouse)	8 (Human, Mouse, Fly, Zebrafish, Rat, Chicken, Cow and <i>C.elegans</i> )
Transcript entries	216,562	549,368
Coding potential Score	CPAT	CPAT1.2.3 (Human, Mouse, Fly, Zebrafish), CPC2 (all) and PLEK (all)
sORF within lncRNA loci	No	Yes (Human, Mouse, Fly, Zebrafish, Rat , <i>C.elegans</i> )
Association with Repeat elements	Yes	Yes (all species)
Co localized miRNAs	Yes	Yes (all species)
Lnc-pri-miRNAs	No	Yes (all species)
Co localized piRNAs	Yes	Yes (Human, Mouse, Fly, Rat, <i>C.elegans</i> )
Predicted Sub cellular Localization	No	Yes (all species)
CGI in lncRNA promoter	Yes	Yes (all species)
TFBS within lncRNA promoter	No	Yes (Human)
Tissue specific expression	Yes (single dataset)	Yes (multiple datasets, all species)
lncRNA disease association (literature curated)	No	Yes (Human)
lncRNA target genes (literature curated)	No	Yes (Human)

**Table 1:** Revised Components of LncRBase V.2 as Contrasted with LncRBase (2014)

## 2. Materials and Methods

### IMPROVED CONTENT AND NEW FEATURES

#### 2.1. Data procurement

LncRNA entries have been extended to encompass six additional species, in addition to Human and Mouse. **Table 1** provides details regarding the genome builds and transcript sources for all eight species. In the case of Human, the genome build has been updated from hg19 to hg38. For *C.elegans* and Chicken, the NONCODE[6] data has been adapted to the cell and gal-gal5 genome builds, respectively, utilizing the UCSC lift over utility [18]. Data is collected in various formats, including fasta, bed, and gtf. Information pertaining to repeat elements, CGI fasta, coordinates, as well as Refseq annotated intron, 5'UTR/3'UTR/CDS exon details, is obtained from UCSC[18]. miRNA and piRNA-

## Chapter2

related information, aligned with the same genome build, is sourced from miRBase [19] and the National Centre for Biotechnology Information (NCBI)[20] respectively. For probe remapping, microarray probes are acquired from the Affymetrix website (<http://www.affymetrix.com>). Ribosome profiled data for six species is retrieved from [sorfs.org](http://sorfs.org) [21] ], while normal tissue-specific total and long RNA sequence data is obtained from the National Centre for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO)[22].

### 2.2. Data processing and refinement

#### 2.2.1. Redundancy check and assigning Alias ID

The process of eliminating redundancy and assigning identifiers follows a similar approach to that used in LncRBase [13] , as outlined: Each entry is labeled with an abbreviation for the species name, followed by "LB\_" and then the subtype designation, and finally a unique numerical identifier. In cases where multiple transcripts exist within the same chromosomal location, a numerical index is added to distinguish them. However, when a transcript belongs to multiple biotype categories, the subtype is denoted as "ambiguous" (AG), rather than assigning multiple IDs to a single transcript. For example, "hsaLB\_AG\_208424" is assigned to a transcript that overlaps both 5'UTR exon and intron regions.

#### 2.2.2. Association with CpG island, Repeat elements and small non-coding RNAs

In all species, the procedures for mapping repeat elements, identifying CGI regions, pinpointing lncRNAs associated with piRNAs (PIWI-interacting RNAs) and miRNAs (microRNAs) , and remapping microarray probes with lncRNA transcripts are consistent with the methods employed in the initial version of LncRBase.[13].

#### 2.2.3. Determining putative lnc-pri-miRNAs

The coordinates of miRNAs specific to each species have been aligned with the corresponding coordinates of lncRNA transcripts. Further, lncRNA and miRNA transcripts originating from the same genomic loci were subjected to alignment using the BLAST-like alignment tool (BLAT) [23] with specific parameters set, including a block count = 1 and a gap count = 0. Transcripts that met both criteria, namely coordinate matching and exact sequence matching, have been identified and annotated as lnc-pri-miRNAs.

#### 2.2.4. Coding capacity of lncRNAs:

In the initial version of lncRBase, the Coding-Potential Assessment Tool (CPAT)[24], was utilized as a standalone tool to evaluate the coding potential of lncRNA transcripts. However, it's worth noting that CPAT's coding potential assessment was restricted to just four species (Human, Mouse, Fruitfly, and Zebrafish). In the current version, we have employed two additional well-known tools, namely CPC2[25] and PLEK[26] alongside CPAT. CPC2 is species-agnostic, offering a broader range of applicability, while PLEK is an alignment-free tool, providing a different approach for assessing coding potential. This multi-tool approach enhances the ability to assess coding potential across a wider spectrum of species.

Scores from all three tools, along with their corresponding predictions, have been compiled and tabulated. For Human, Mouse, Fruitfly, and Zebrafish, the updated version of CPAT (CPAT1.1.3) was employed. Additionally, CPC2 and PLEK were used for all eight species. The coding potential threshold values for CPAT vary by species: Human: 0.364, Mouse: 0.44, Fruitfly: 0.39 and Zebrafish: 0.38. For CPC2 and PLEK, a common threshold value of 0.5 was applied for all species, indicating a more conservative approach in coding potential assessment compared to CPAT.

To identify putative micro peptide coding evidence, the coordinate data and related information for small open reading frame (sORF) sequences (approximately with a cutoff length of around 300 nucleotides) were obtained from sorfs.org[21] for six species (Human, Mouse, Fruitfly, Zebrafish, Rat and *C.elegans*). The following steps were taken to screen lncRNAs that potentially contain sORFs: a. Coordinate mapping of sORFs with lncRNA loci was performed. b. Sequence alignment was carried out using BLAT to compare the sORF sequences with the lncRNA sequences. This process helped identify lncRNAs that bear putative sORFs, suggesting their potential to code for micropeptides.

#### 2.2.5. Sub-cellular localization of lncRNAs

A web-based tool called LncLocator, developed by Ciao et al.[27] has been utilized to predict the subcellular localization of lncRNA transcripts for all the species in the study. LncLocator employs a Neural Network-based stacked ensemble approach by combining four base classifiers into a single model. This approach enhances prediction accuracy by leveraging the strengths of individual predictors through a stacked ensemble method, which is known for its superior performance compared to individual classifiers [28].

LncLocator provides a putative score for five subcellular locations, namely Cytoplasm, Nucleus, Ribosome, Cytosol, and Exosome, with an overall accuracy of 0.59. To determine the most probable subcellular location for a specific lncRNA transcript, the tool selects the one with the highest score among these five categories.

#### 2.2.6. TFBS in lncRNA promoter region

## Chapter2

High-confidence TF-DNA interaction data, provided as bed files, were acquired from Unibind [29]. Unibind is a database that contains binding site predictions for 232 distinct transcription factors (TFs), which were derived from the analysis of 1983 Chip-seq peak datasets. Custom scripts were employed to organize and align this data specifically within the promoter regions of lncRNAs. The mapping was constrained to a range of 1000 base pairs upstream and downstream of the transcription start site (TSS) of the lncRNAs.

### 2.2.7. LncRNA profile across multiple tissues

Raw long RNA-Seq data specific to normal tissues from various species, sourced from GEO [30], underwent a series of preprocessing steps. First, the data were quality-checked using FastQC [31]) and subjected to adapter trimming with Cutadapt [32]. Subsequently, the data were aligned to their corresponding reference genomes using Hisat2 [33]. The transcript assembly was performed using Stringtie [33], followed by the generation of transcript FPKM (Fragments Per Kilobase of transcript per Million mapped reads) files.

### 2.3. Database Implementation:

LncRBase V.2 primarily handles user queries through straightforward search options, and it retrieves information from relational databases, as illustrated in **Figure 1**. The General Output page (**Figure 1**) presents essential details about the lncRNA transcript and offers various avenues for exploring more in-depth information. The Detailed Output page provides comprehensive data about the lncRNA transcript.

#### 2.3.1. The "Transcripts" menu:

##### i. Search by lncRNA Accession ID:

Users have the option to choose the organism of interest and then input a specific lncRNA Accession ID. This Accession ID can either originate from (a) LncRBase V.2, such as "hsaLB\_LI\_10017" for Human or "dreLB\_AO\_675" for Zebrafish, or (b) any of the source databases, including Ensembl Gene90, UCSC ID, NONCODE v5.0, H-InvDB 8.0, or Lncipedia. After submitting the query, a result page will be generated, presenting detailed information tailored to the specific query feature.

##### ii. Search by lncRNA Gene Symbol:

Users have the option to choose the organism of interest and input a known lncRNA Gene Symbol to search for gene-specific transcript entries within LncRBase V.2. Upon submitting the query, the results page will showcase a list of information related to all the transcripts associated with that particular gene symbol, complete with their respective LncRBase V.2 IDs. Additionally, the page will provide literature evidence pertaining to



that specific lncRNA gene. Detailed information about individual transcripts can be accessed by clicking on the displayed LncRBase IDs.

### **iii. Browse by lncRNA subtype, coding potential and sORF overlap:**

In the first part, users can choose the organism's name, specify the chromosome number, select the lncRNA biotype, coding potential tool name, and coding potential type to access the relevant results. Additionally, users have the option to specify a genome locus if needed.

In the second part, users can select the organism's name, provide the chromosome input, and choose the lncRNA biotype to retrieve the count of small open reading frames (sORFs) associated with the retrieved lncRNAs.

### **iv. Search for associated genomic elements:**

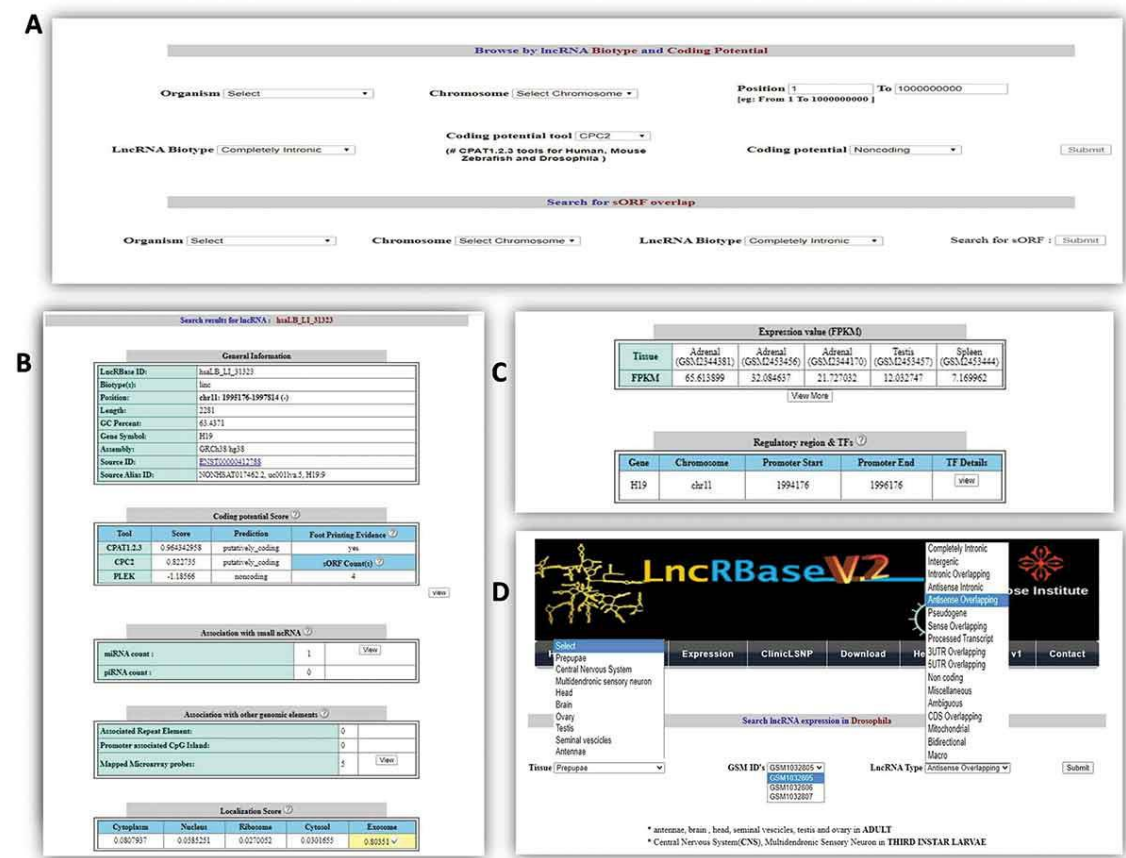
Users have the option to search for associated 'Repeat elements,' 'CGI' (CpG Islands), and remapped microarray 'probe' information by selecting their desired option, along with specifying the organism's name and chromosome number. This allows users to obtain relevant data based on their specific criteria and preferences.

### **v. Search for association with small ncRNA:**

Users can perform organism and chromosome-specific searches to identify lncRNAs associated with small non-coding RNAs, such as 'miRNAs' and 'piRNAs'. This feature enables users to pinpoint lncRNAs that are linked to these particular classes of small non-coding RNAs within a specific organism and on a specific chromosome.

### **2.3.2. Species-specific lncRNA expression profile:**

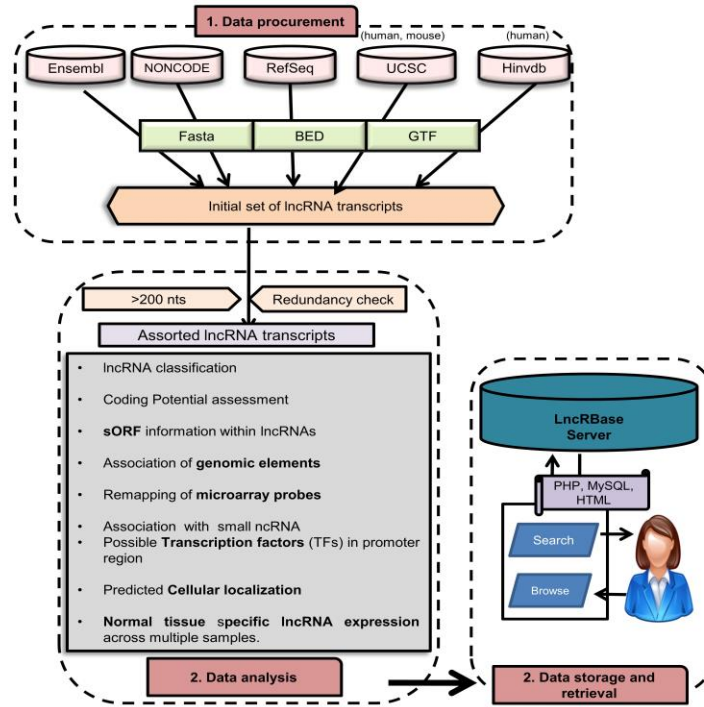
When a user clicks on an organism name under the "Expression" menu, it will lead to the organism-specific lncRNA expression query page. On this page, selecting a specific "tissue" will display all the corresponding data "accession IDs" related to that tissue. Users also have the option to refine the results by submitting a query specific to a particular "lncRNA biotype," which can help limit the number of outputs and provide more focused information.



**Figure 1:** Various Web Interfaces for Convenient Access to LncRBase V.2

A. General search page options. B. In-depth details of individual LncRBase IDs, encompassing general information, coding potential scores, connections to small ncRNAs, associations with genomic elements, and cellular localization scores. C. Comprehensive information on individual LncRBase IDs, providing tissue-specific expression values and Transcription Factor overlaps within the regulatory region. D. Tissue-specific search options for individual species, accessible across multiple Data Accession IDs.

LncRBaseV.2 has been constructed using MySQL, an open-source relational database management system. The web server operates within a Linux environment and relies on the Apache HTTP Server, which is also open-source and cross-platform. The interface layer has been crafted using a combination of HTML, CSS, and JavaScript to provide a user-friendly experience. To connect the database with the web interface and enable dynamic functionality, the PHP module has been employed. PHP, as a server-side scripting language, is responsible for generating dynamic web pages and accessing data from MySQL to produce the desired output (Figure2).



**Figure 2:** Workflow of LncRBase V.2

### 3. Results and Discussion:

LncRBase V.2, the updated version of LncRBase, is a comprehensive repository hosting information on 549,368 lncRNAs across eight different species: Human, Mouse, Rat, Fruitfly, Cow, Chicken, Zebrafish, and C.elegans. This new version brings several additional features and improvements, enhancing the significance and utility of this database. These added features contribute to the database's increased value and functionality for researchers and users.

#### 3.1. Distribution of lncRNA subtypes based on their genomic location

LncRBase V.2 offers a species-specific classification of lncRNA transcripts into a total of sixteen distinct biotypes. While some of these biotypes are newly classified, others may have existed previously. In the case of Human and Mouse, lncRNAs are distributed across all seventeen biotypes. However, for species like Cow and C. elegans, lncRNAs are distributed across ten biotypes. Below is a representation of these biotypes along with their corresponding nomenclatures as used in **Table 2**.

## Chapter2

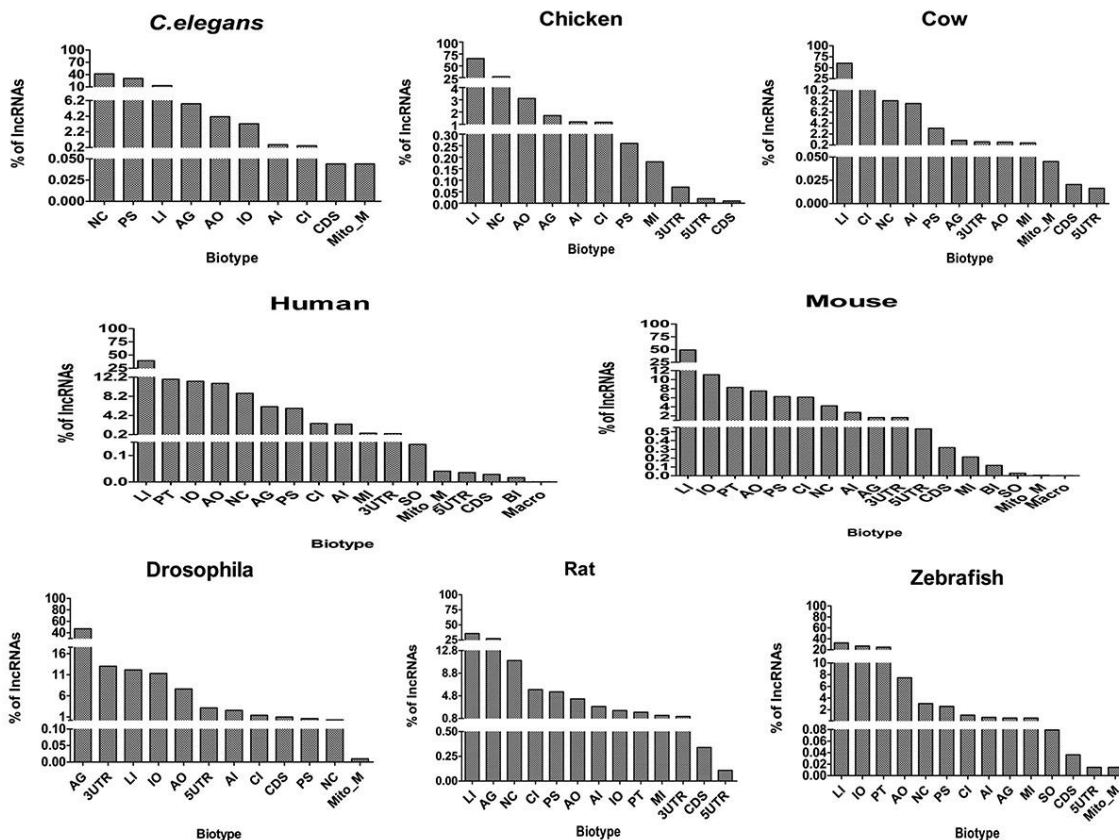
No. of biotypes	Genomic location of the LncRNA	Name of the biotype	Nomenclature for the biotype specific lncRNA
1.	<b>LncRNAs overlapping with 3'UTR exonic region in the sense strand.</b>	<b>3'UTR overlapping lncRNA</b>	<b>3UTR</b>
2.	<b>LncRNAs overlapping any 5'UTR exon in the sense strand</b>	<b>5'UTR overlapping lncRNA</b>	<b>5UTR</b>
3.	LncRNAs overlapping any CDS exon	CDS overlapping lncRNA	CDS
4.	Intergenic (linc) lncRNAs transcribed from in between two gene loci	Intergenic (linc) lncRNA	LI
5.	<b>LncRNAs intersecting any exon of a protein-coding locus on the opposite strand</b>	<b>Antisense Overlapping lncRNA</b>	<b>AO</b>
6.	LncRNAs residing within introns of a coding gene, but do not intersect any exons	Completely Intronic lncRNA	CI
7.	<b>Antisense lncRNAs completely overlapping with an intron in the opposite strand</b>	<b>Antisense Intronic lncRNA</b>	<b>AI</b>
8.	Intron Overlapping lncRNA splice variants of a gene, contain intronic sequence	Intron Overlapping lncRNA	IO
9.	Pseudogene transcripts having homology to protein coding transcripts but containing disrupted coding sequence and an active homologous gene can be found at another locus	Pseudogene	PS
10.	Sense Overlapping lncRNAs containing a coding gene in its intron on the same strand( Ensembl annotated)	Sense Overlapping lncRNA	SO
11.	Processed Transcripts not containing an ORF (Ensembl annotated)	Processed Transcripts	PT
12.	Miscellaneous RNA(miscRNA)	miscRNA	MI

**Table 2:** Naming Convention for LncRNA Biotypes' Distribution. In this version of the database, newly introduced biotypes are denoted in *italics*, while modifications to the abbreviations of existing biotypes from the previous LncRBase are indicated in **bold**.

In LncRBase V.2, several new biotypes have been introduced (indicated in *italics*), and some changes have been implemented to enhance user-friendliness, including the use of **bold text** for updated biotype abbreviations. Notably, the biotype "Ambiguous ORF"

(abbreviated as AG) from the previous version has been eliminated in this update. This change was necessitated by the fact that it is no longer supported by Ensembl 90, the version used in this work.

**Figure 3** illustrates the distribution of lncRNAs based on their biotypes across various species' genomes. It highlights the prevalence of intergenic lncRNAs in mammalian systems, as well as in Chicken and Zebrafish. In the case of *Drosophila*, the majority of lncRNAs overlap with more than one genomic location, resulting in their classification as Ambiguous (AG) type. Additionally, in *C.elegans*, most of the lncRNAs are categorized as Non-coding (NC) type because they couldn't be definitively assigned to specific genomic loci based on existing genomic annotation.

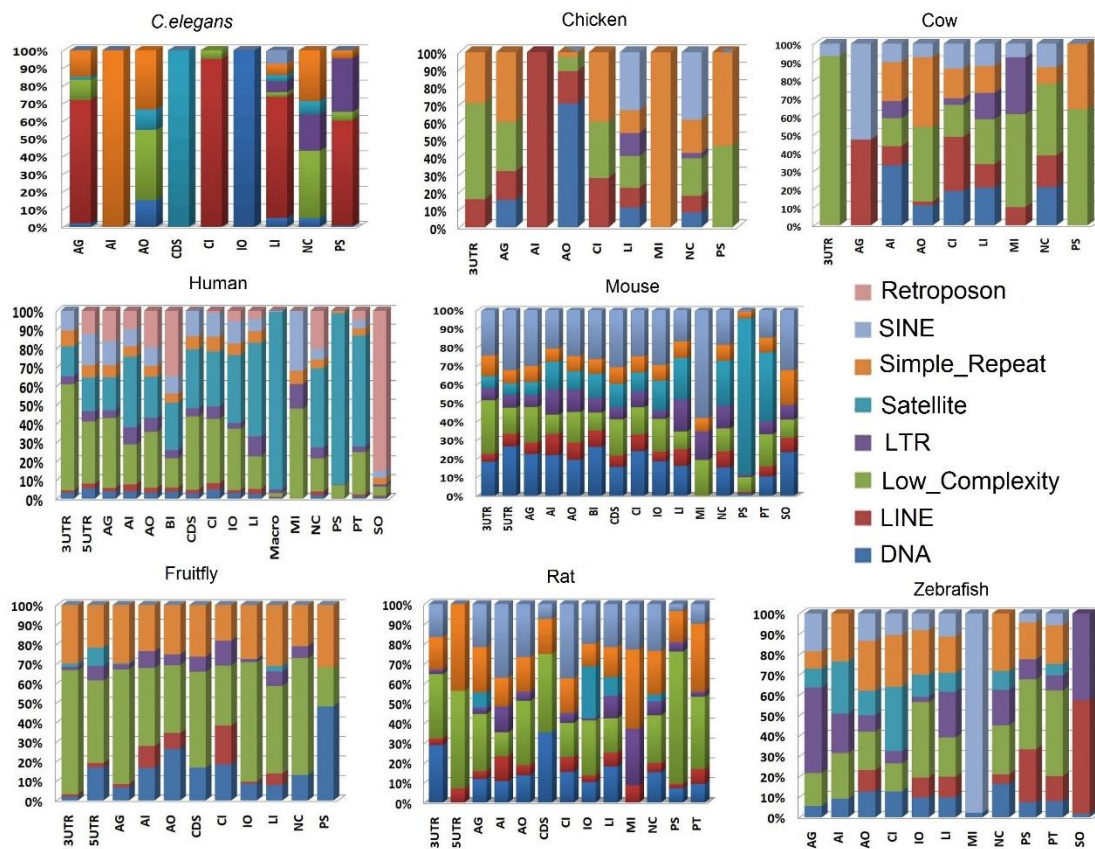


**Figure 3:** Distribution of lncRNAs Based on Biotype

Abbreviations used: NC Non-Coding, MI MiscRNA, SO Sense Overlapping, PS Pseudogenes, CI Completely Intronic, 3UTR 3'UTR overlapping, PT Processed Transcript, 5UTR 5'UTR overlapping, CDS CDS overlapping, IO Intron Overlapping, LI Long Intergenic, AO Antisense Overlapping, AI Intronic Antisense, AG Ambiguous, BI Bidirectional lncRNAs, Mito\_M Mitochondrial lncRNAs, Macro Macro lncRNAs

## 3.2. Distribution of repeat elements within lncRNA loci

Similar to the first version of LncRBase, repeat elements from various repeat classes have been mapped to lncRNA loci across all species. **Figure 4** displays the distribution of the eight predominant repeat classes, namely SINE, LINE, DNA, Simple repeat, low complexity, LTR, Satellite, and SVA Retroposon (specific to Humans), among different lncRNA biotypes. Notably, Fruitfly, Zebrafish, and Mammalian lncRNAs exhibit enrichment in low complexity repeats. The physiological significance of low complexity sequences within lncRNAs can be found in the literature. For example, within ribosomal intergenic noncoding RNA (rIGSRNA), low complexity CU/AG repeats facilitate the formation of amyloid bodies [34]. On the other hand, C.elegans lncRNAs and Chicken lncRNAs are enriched with Simple repeat and LINE repeat elements..

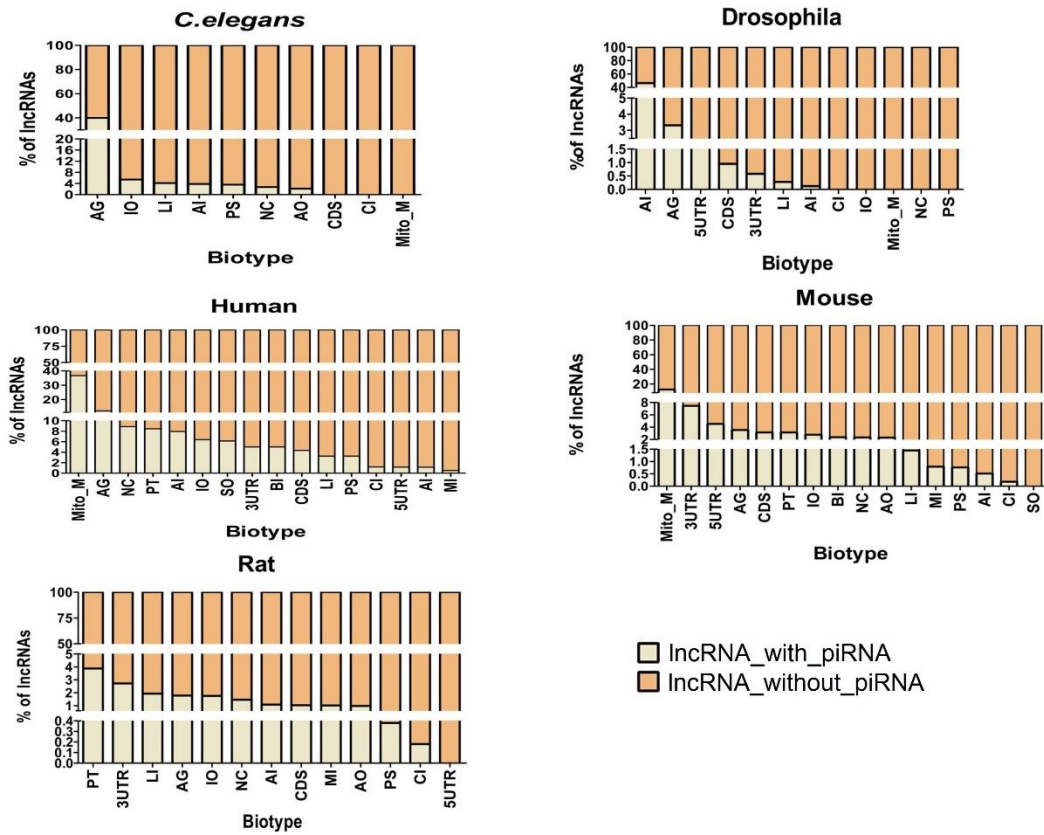


**Figure 4:** Distribution of Repeat Classes Among Various lncRNA Biotypes

### 3.3. Abundance of piwil interacting RNAs (piRNAs) within lncRNA loci

The genomic locations of piRNAs in Human, Mouse, Rat, Drosophila, and *C.elegans*, which are available in NCBI, have been mapped to their respective lncRNA loci. The analysis revealed that 55.7% of Human piRNAs, 51.5% of Mouse piRNAs, 4.5% of Rat piRNAs, 11.5% of Drosophila piRNAs, and 31.4% of *C.elegans* piRNAs overlap with their corresponding lncRNA loci. **Figure 5** presents the distribution of piRNAs within the loci of lncRNAs belonging to different biotypes for these five species.

Notably, it's interesting to observe that a significant percentage of Human and Mouse piRNAs overlap with the newly identified lncRNA biotype Mito\_M, despite the fact that only a small percentage of Human and Mouse lncRNAs fall into this specific biotype, as depicted in **Figure 3**.

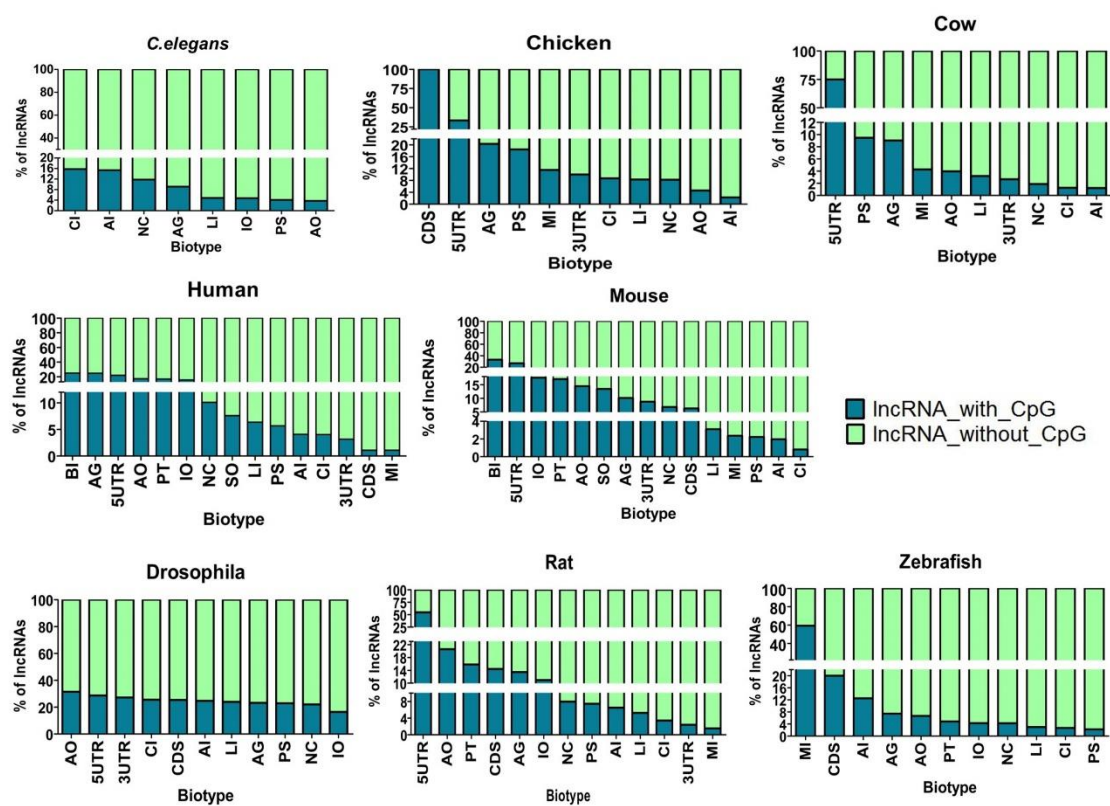


**Figure 5:** Distribution of piRNAs Overlapping with lncRNA Loci Across Different Biotypes for Eight Species.



### 3.4. CGI association with lncRNA promoter region

**Figure 6** illustrates the distribution of CpG islands (CGI) within the promoter regions of lncRNAs, spanning 1000 bases upstream and downstream of the Transcription Start Site (TSS). The overall pattern reveals that only a minority of lncRNA promoters are associated with CpG islands, with percentages as low as 2% for Cow and as high as approximately 23% for Fruitfly. However, when considering various biotypes, it is observed that the promoters of 5UTR lncRNA transcripts are predominantly associated with CpG islands and exhibit a high CpG content.



**Figure 6:** Distribution of CpG Islands in the Promoter Regions of lncRNAs Based on Biotype

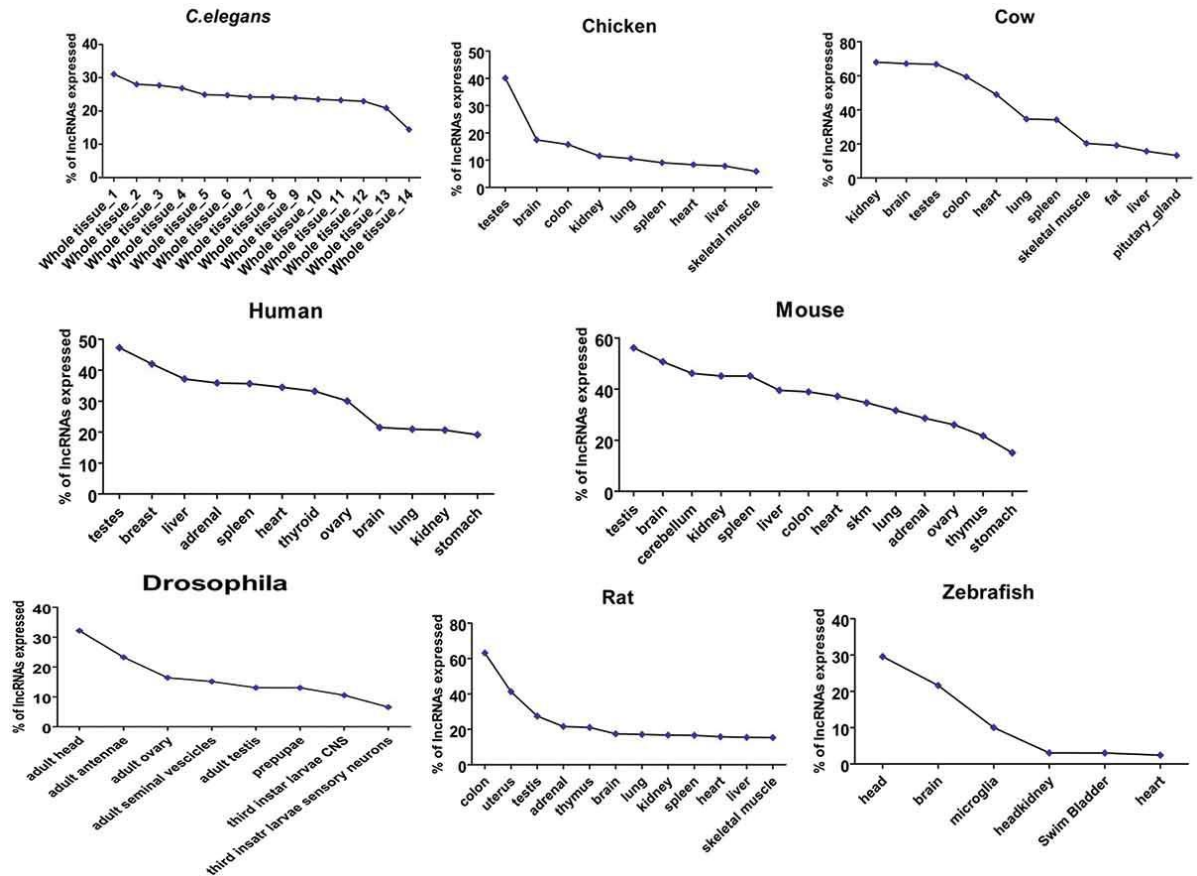
The association of CpG islands along with tissue specificity of lncRNAs is known to have a significant impact on various diseases. lncRNAs like MALAT1 and NEAT1, which have established roles in multiple disorders and carcinomas affecting different tissues, are found to have CGI-associated promoters, as indicated in our analysis. In lncRBase V.2, we have delved into the relationship between CGI association and lncRNA tissue specificity, which is discussed further in the next section.



### 3.5. Tissue specific distribution of lncRNAs:

lncRNAs are widely recognized for their tissue-specific expression patterns [35, 36]. Furthermore, it's worth noting that a significant proportion of tissue-specific gene promoters are not associated with CpG islands [37]. This aligns with the results of our CpG island analysis, where we observed that the majority of lncRNA gene promoters do not reside within CGI regions, as discussed in Section 1.

Hence, obtaining a clear understanding of lncRNA expression across various tissue types is crucial for deciphering the tissue-specific functions of these molecules. Our analysis encompassed fourteen whole-tissue RNA-seq datasets for *C.elegans*. For Fruitfly, we examined tissue data for all three developmental stages: Prepupae, third instar larvae, and adults. In the case of Zebrafish, we focused on RNA-seq data from head, microglia, swim bladder, head kidney, and heart tissues. Additionally, we considered various tissues such as brain, colon, kidney, spleen, lungs, skeletal muscle, testes, and others for the four mammalian species and chicken. Our analysis unveiled that lncRNA expression is predominantly observed in the testis of Human, Mouse, and Chicken, whereas in *Drosophila* and Zebrafish, its expression predominates in the head. In Cow and Rat, predominance is noted in the kidney and colon, respectively. The tissue-specific lncRNA expression profiles for these eight species are depicted in **Figure 7**.



**Figure 7:** Distribution of lncRNAs Across Various Tissues (Healthy) for the Eight Species

## Chapter2

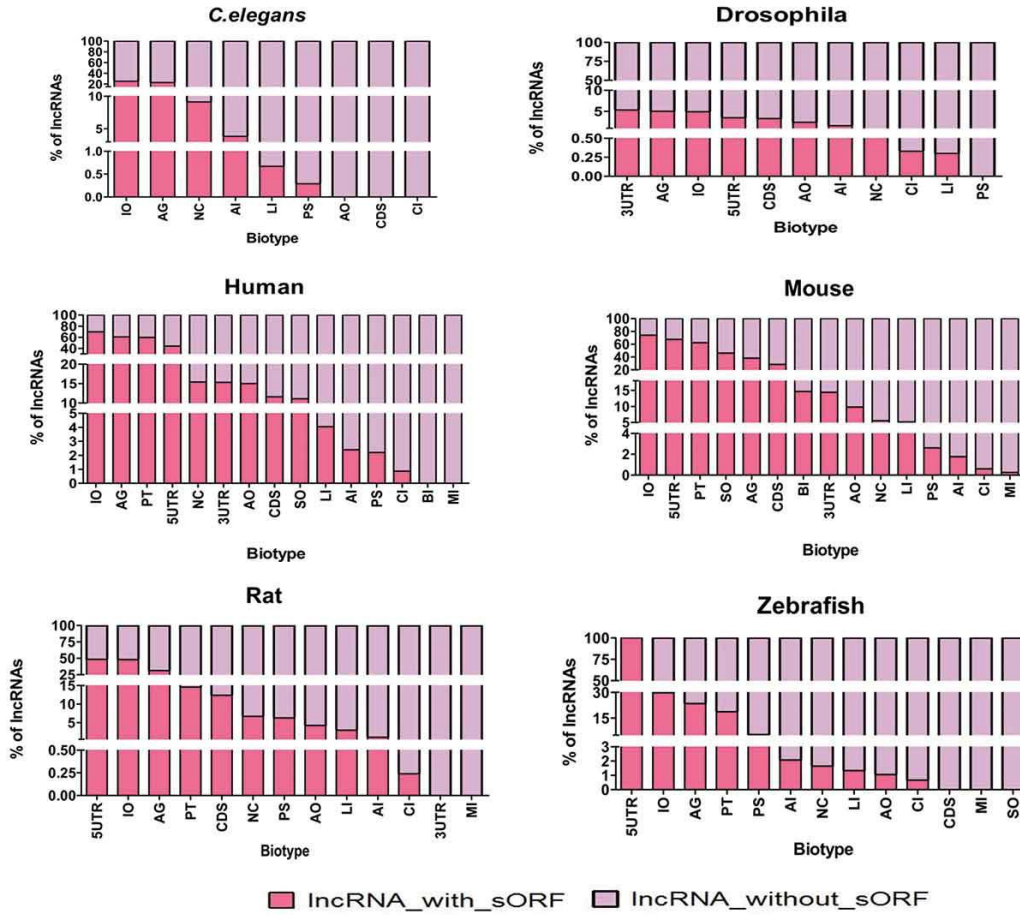
### 3.6. Bifunctional lncRNAs:

lncRNAs have been found to serve dual functions, acting both as protein-coding molecules and regulatory non-coding RNAs. This dual role was first demonstrated in *Drosophila*, where the lncRNA *tal* (tarsal less) was found to code for four micropeptides involved in embryonic development and influencing the fate of the transcription factor *Svb* (shavenbaby) [38]. Subsequent evidence of lncRNA-encoded micropeptides has been reported in various species, including Human, Mouse, Chicken, Zebrafish, Nematode, and others [39-42]. This discovery raises questions about the true non-coding identity of lncRNAs and suggests that they may function as bifunctional lncRNAs, playing both regulatory and protein-coding roles [43, 44].

Small Open Reading Frames (sORFs), initially considered as mere 'background noise' in proteomics experiments, have now been recognized for their potential to code for clinically significant micropeptides [42]. sORFs are also known for their sequence conservation [45].

The use of ribosome footprinting techniques, which can distinguish between translated and untranslated ORFs, along with stringent computational cutoff criteria [16], has enabled the identification of several thousand sORFs translated from non-coding regions of the genome. Many of these sORFs have had their coding capacity validated [16]. This evolving understanding of lncRNAs and sORFs highlights the complexity of gene regulation and protein-coding potential within the non-coding genome.

In our efforts to classify lncRNAs, we have not only utilized Coding Potential Score prediction tools to categorize them into noncoding and putatively coding types but have also sought to identify bi-functional lncRNAs. To achieve this, we have mapped sORFs within the genomic loci of lncRNAs. Information on sORFs for six species, including Human, Mouse, Rat, Fruitfly, Zebrafish, and *C.elegans*, has been sourced from [www.sorfs.org](http://www.sorfs.org) [17]. The distribution of sORF-containing lncRNAs across different biotypes is depicted in **Figure 8**.



**Figure 8:** Distribution of lncRNA Biotypes Hosting Small Open Reading Frames (sORFs)

Here are the key observations: For Human and Mouse, sORFs are predominantly mapped within the IO (Intronic Overlapping) biotype. In *C. elegans*, sORFs are most prevalent in both the IO and AG (Ambiguous) biotypes. In Rat, sORFs are found abundantly in both the IO and 5UTR biotypes. Zebrafish shows a high enrichment of sORFs within the 5'UTR biotype. *Drosophila* displays a significant presence of sORFs in all three biotypes: IO, 5UTR, and AG. Notably, a small portion of sORFs is also found within intergenic lncRNAs (LI), which are situated between two coding regions. This proportion is relatively minor but is still notable, with the highest percentages observed for Human (4.05%) and Mouse (5.17%).

Given that many annotated lncRNAs have yet to be experimentally validated, the combined approach of in silico prediction of coding potential and the incorporation of foot-printing information serves as a convincing method to identify bi-functional lncRNAs. This approach helps uncover lncRNAs that may possess both regulatory non-coding functions and the capacity to code for micro-peptides.

### 3.7. lncRNAs harbouring embedded miRNAs:

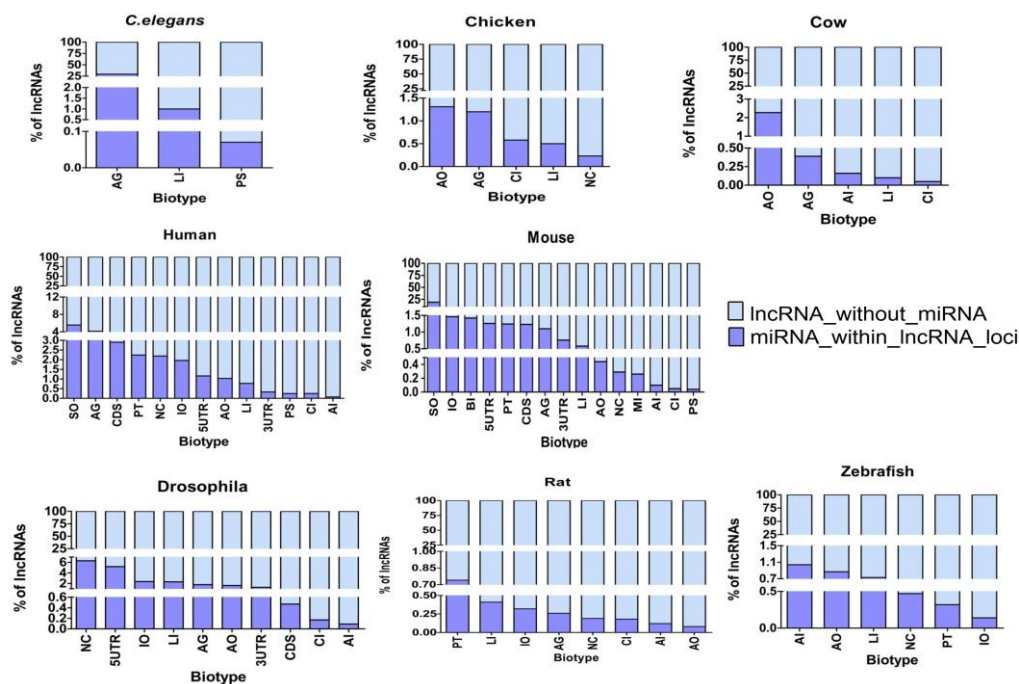
A significant number of miRNAs have been found to originate from the exonic regions of lncRNAs. A well-known example is the lncRNA H19-derived miRNA mir-675, which plays a co-dependent role in gastric cancer [46]. The biogenesis of these miRNAs is often

## Chapter 2

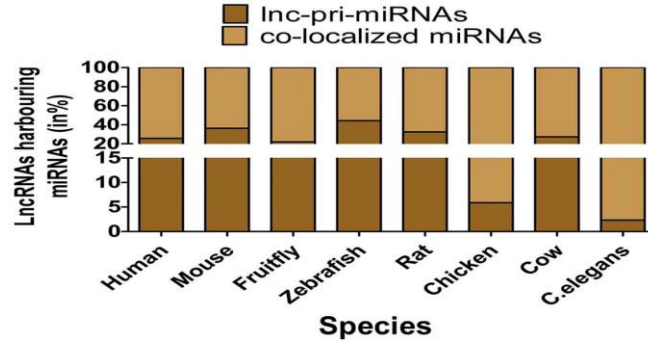
different from that of miRNAs originating from the exonic regions of protein-coding genes [47]. Unlike host protein-coding genes, where mRNA and miRNA are co-transcribed, certain cellular systems employ a microprocessor-mediated mechanism of transcriptional termination to halt the transcription of the lncRNA gene, prioritizing the transcription of the embedded miRNA [48]. These miRNAs are referred to as lnc-pri-miRNAs, and their biogenesis deviates from the canonical pathway of Pol II-associated cleavage and polyadenylation (CPA). For instance, the biogenesis of miRNA mir-122, which is involved in cholesterol metabolism and Hepatitis C virus replication, results in the terminated transcription of its host lncRNA [49]. The precise factors that determine whether the host lncRNA or its embedded miRNA predominates in the system at a given time are still under investigation.

In our previous version of LncRBase, we concentrated solely on miRNAs that originated from the same genomic loci as the lncRNAs, which we termed as co-localized miRNAs. However, in the latest version, in addition to the information presented in **Figure 9**, which includes details about co-localized miRNAs, we have expanded our data to encompass lnc-pri-miRNAs, as shown in **Figure 10**. The criteria for determining these lnc-pri-miRNAs have been thoroughly explained in the materials and methods section.

Among the species analyzed, Zebrafish exhibits the highest percentage of lnc-pri-miRNAs at 44.2%, while *C. elegans* has the lowest percentage at 2.32%. Investigating whether these lnc-pri-miRNAs follow canonical or noncanonical modes of transcription could provide valuable insights into their regulatory mechanisms and functional roles.



**Figure 9:** Percentage of lncRNAs harboring miRNA within their Loci based on biotypes.

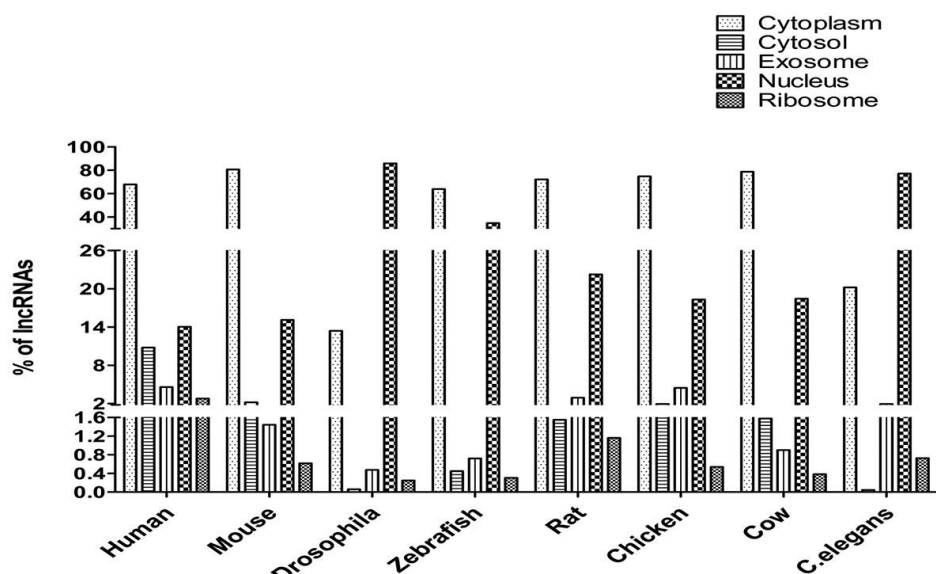


**Figure 10:** Distribution of lncRNAs Hosting lnc-pri-miRNAs and Co-localized miRNAs Across Eight Species. Co-localized miRNAs are those originating from lncRNA loci, while lnc-pri-miRNAs are a subset originating from lncRNA loci with an exact sequence match to the mapped region within those lncRNA loci. The percentage of lnc-pri-miRNAs is determined in relation to the total number of miRNAs originating from lncRNA loci.

### 3.8. Cellular localization of lncRNAs:

lncRNAs are known to exhibit diverse patterns of subcellular localization, which can involve the nucleus, cytoplasm, or a non-specific presence in both compartments [50]. This variability in localization is intricately linked with the function of lncRNAs [51, 52]. However, it's important to note that the localization remains unknown for a significant majority of lncRNAs.

In our database, we provide information regarding the predicted localization of lncRNA transcripts, as depicted in **Figure 11**. The analysis indicates cytoplasmic enrichment as the most common pattern across species, followed by nucleus and cytosol. Notably, *Drosophila* and *C.elegans* exhibit a distinct pattern, with more than 70% of lncRNAs showing nucleus enrichment. Furthermore, recent research has identified lncRNAs within exosomes, which are the smallest type of Extracellular Vesicles (EV). [53]. Exosomes play a crucial role in cell-to-cell communication, and one example is the exosomal lncRNA LINC00152, which is linked to gastric cancer [54]. Our analysis suggests approximately 5% exosome coverage for Human, Rat, and Chicken lncRNAs. Interestingly, ribosomes harbor the fewest lncRNAs across all species. Understanding the potential subcellular preference of an lncRNA transcript can serve as a valuable starting point for identifying its possible interacting partners and elucidating its functional roles in cellular processes.



**Figure 11:** Subcellular Localization of lncRNAs in the Eight Species

### 3.9. Transcription Factor binding site in lncRNA promoter regions:

In our work, we conducted an extensive scan of lncRNA promoter regions to identify the upstream regulators of lncRNA genes. To achieve this, we leveraged ChIP-Seq data analysis, which unveiled thousands of binding peaks across the genome in multiple tissues and under various disease conditions, as described in reference [29]. Subsequently, we mapped TFs within a range of 1000 kilobases both upstream and downstream of the TSS for more than 71% of lncRNA genes.

This endeavor to uncover TF binding within the lncRNA promoter region serves multiple purposes. It aids in comprehending the upstream transcriptional regulators that govern the tissue-specific expression of lncRNAs, sheds light on their modes of regulation, and can even provide insights into the association of poorly annotated lncRNAs with diseases. By identifying these regulatory interactions, we contribute to a deeper understanding of the functional roles of lncRNAs in various biological processes and disease contexts.

### 3.10. LncRNA target partners and disease association:

Extensive literature curation efforts have been undertaken to compile valuable information regarding the interacting partners of 408 Human and 41 Mouse lncRNA genes. This information encompasses the names of the interacting molecules, the mode of regulation, and references to the relevant PubMed articles. Additionally, comprehensive data on 1129 lncRNA genes associated with various diseases has been meticulously curated. This includes details on the mode of regulation, the nature of the experimental evidence, and the corresponding PubMed IDs. These curated resources will serve as a readily accessible reference for users seeking experimentally validated



information on lncRNA targets and their associations with specific diseases, allowing them to delve into the published literature for further insights and exploration.

### 4. Conclusion:

The evolution of LncRBase from its last release in 2014 represents a significant advancement in our understanding of lncRNA and their roles in diverse biological processes. Here are the key highlights and updates in the new version of LncRBase – (i) Expanded Species Coverage: LncRBase has extended its scope to include six additional species, encompassing mammals, birds, insects, nematodes, and fish. This expansion allows for a more comprehensive catalog of lncRNA transcripts, providing insights into the world of lncRNAs across diverse organisms. (ii) Incorporation of New Human and Mouse Transcripts: The latest version of LncRBase includes newly discovered lncRNA transcripts for Human and Mouse, ensuring that the database remains up-to-date with the latest research findings. Obsolete transcripts have been removed to maintain data accuracy. (iii) User-Friendly Nomenclature: LncRBase now employs a user-friendly nomenclature system for lncRNA transcripts, making it easier for users to navigate and identify specific transcripts. Additionally, new biotypes have been introduced to enhance the classification of lncRNAs. (iv) Coding Potential Assessment: The database now includes a comprehensive evaluation of lncRNA coding potential using multiple tools. Moreover, it incorporates evidence from footprinting studies, providing insights into the potential coding capacity of lncRNAs, including the production of micropeptides. (v) TFBS: In the case of Human, LncRBase V.2 offers information on TFBS within lncRNA promoter regions. This feature helps users understand the regulatory interactions that govern lncRNA expression. (vi) Tissue-Specific Expression Analysis: LncRBase now provides tissue-specific expression profiles for lncRNAs across multiple samples per tissue for all covered species. This facilitates the exploration of tissue-specific functions and expression patterns of lncRNAs. (vii) Identification of lnc-pri-miRNAs: The database includes information on lnc-pri-miRNAs, which are lncRNA transcripts hosting embedded miRNAs. This update sheds light on the potential regulatory roles of lncRNAs in miRNA biogenesis. (viii) Cellular Localization Prediction: LncRBase offers predictions regarding the subcellular localization of lncRNAs, helping users understand where these molecules operate within the cell. (ix) Disease-Associated Information: The database includes literature-curated information about the involvement of lncRNAs in various disease systems and their interacting partners. This feature provides valuable insights into the roles of lncRNAs in health and disease.

The enhancements introduced in LncRBase V.2, in addition to the existing features, position it as a comprehensive and data-rich repository of lncRNAs. Our ongoing commitment is to continually update LncRBase V.2 with newly annotated lncRNAs and incorporate additional features to ensure that it remains in sync with the rapid expansion of lncRNA research worldwide. We aspire to provide researchers with the most up-to-date and valuable resource for exploring the diverse and evolving landscape of lncRNAs, contributing to a deeper understanding of these important molecules in biology.

## Chapter2

### 5. Availability

LncRBase V.2 is accessible to users at no cost through the following web link: <http://dibresources.jcbose.ac.in/zhumur/lncrbase2/> . Users are encouraged to freely download and utilize the database's files, adhering to the guidelines outlined in the GNU Public License.

### 6. References:

1. Wilusz, J.E., H. Sunwoo, and D.L. Spector, *Long noncoding RNAs: functional surprises from the RNA world*. Genes Dev, 2009. **23**(13): p. 1494-504.
2. Rinn, J.L. and H.Y. Chang, *Genome regulation by long noncoding RNAs*. Annu Rev Biochem, 2012. **81**: p. 145-66.
3. Mercer, T.R., M.E. Dinger, and J.S. Mattick, *Long non-coding RNAs: insights into functions*. Nat Rev Genet, 2009. **10**(3): p. 155-9.
4. Cerik, S., et al., *Current Status of Long Non-Coding RNAs in Human Breast Cancer*. Int J Mol Sci, 2016. **17**(9).
5. Flicek, P., et al., *Ensembl 2012*. Nucleic Acids Res, 2012. **40**(Database issue): p. D84-90.
6. Fang, S., et al., *NONCODEV5: a comprehensive annotation database for long non-coding RNAs*. Nucleic Acids Res, 2018. **46**(D1): p. D308-D314.
7. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation*. Nucleic Acids Res, 2016. **44**(D1): p. D733-45.
8. Volders, P.J., et al., *LNCipedia 5: towards a reference set of human long non-coding RNAs*. Nucleic Acids Res, 2019. **47**(D1): p. D135-D139.
9. Zhou, K.R., et al., *ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data*. Nucleic Acids Res, 2017. **45**(D1): p. D43-D50.
10. Gao, Y., et al., *Lnc2Cancer v2.0: updated database of experimentally supported long non-coding RNAs in human cancers*. Nucleic Acids Res, 2019. **47**(D1): p. D1028-D1033.
11. Bao, Z., et al., *LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases*. Nucleic Acids Research, 2018. **47**(D1): p. D1034-D1037.
12. Ren, C., et al., *Lnc2Catlas: an atlas of long noncoding RNAs associated with risk of cancers*. Scientific Reports, 2018. **8**(1): p. 1909.
13. Chakraborty, S., et al., *LncRBase: an enriched resource for lncRNA information*. PLoS One, 2014. **9**(9): p. e108010.
14. van Heesch, S., et al., *Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes*. Genome Biol, 2014. **15**(1): p. R6.
15. Chen, L.L., *Linking Long Noncoding RNA Localization and Function*. Trends Biochem Sci, 2016. **41**(9): p. 761-772.
16. Yeasmin, F., T. Yada, and N. Akimitsu, *Micropeptides Encoded in Transcripts Previously Identified as Long Noncoding RNAs: A New Chapter in Transcriptomics and Proteomics*. Front Genet, 2018. **9**: p. 144.
17. Olexiouk, V., et al., *sORFs.org: a repository of small ORFs identified by ribosome profiling*. Nucleic Acids Res, 2016. **44**(D1): p. D324-9.
18. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool*. Nucleic Acids Res, 2004. **32**(Database issue): p. D493-6.
19. Kozomara, A. and S. Griffiths-Jones, *miRBase: integrating microRNA annotation and deep-sequencing data*. Nucleic Acids Res, 2011. **39**(Database issue): p. D152-7.
20. Geer, L.Y., et al., *The NCBI BioSystems database*. Nucleic Acids Res, 2010. **38**(Database issue): p. D492-6.



21. Olexiouk, V., W. Van Criekinge, and G. Menschaert, *An update on sORFs.org: a repository of small ORFs identified by ribosome profiling*. Nucleic Acids Res, 2018. **46**(D1): p. D497-D502.
22. Barrett, T. and R. Edgar, *Gene expression omnibus: microarray data storage, submission, retrieval, and analysis*. Methods Enzymol, 2006. **411**: p. 352-69.
23. Kent, W.J., *BLAT--the BLAST-like alignment tool*. Genome Res, 2002. **12**(4): p. 656-64.
24. Wang, L., et al., *CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model*. Nucleic Acids Res, 2013. **41**(6): p. e74.
25. Kang, Y.J., et al., *CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features*. Nucleic Acids Res, 2017. **45**(W1): p. W12-W16.
26. Li, A., J. Zhang, and Z. Zhou, *PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme*. BMC Bioinformatics, 2014. **15**: p. 311.
27. Cao, Z., et al., *The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier*. Bioinformatics, 2018. **34**(13): p. 2185-2194.
28. Pan, X.Y., et al., *Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach*. Genomics, 2011. **97**(5): p. 257-64.
29. Gheorghe, M., et al., *A map of direct TF-DNA interactions in the human genome*. Nucleic Acids Res, 2019. **47**(4): p. e21.
30. Clough, E. and T. Barrett, *The Gene Expression Omnibus Database*. Methods Mol Biol, 2016. **1418**: p. 93-110.
31. S., A., *FastQC: a quality control tool for high throughput sequence data*. . Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>, 2010.
32. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads*. 2011, 2011. **17**(1): p. 3.
33. Pertea, M., et al., *Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown*. Nat Protoc, 2016. **11**(9): p. 1650-67.
34. Wang, M., et al., *Stress-Induced Low Complexity RNA Activates Physiological Amyloidogenesis*. Cell Rep, 2018. **24**(7): p. 1713-1721 e4.
35. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses*. Genes Dev, 2011. **25**(18): p. 1915-27.
36. Kaushik, K., et al., *Dynamic expression of long non-coding RNAs (lncRNAs) in adult zebrafish*. PLoS One, 2013. **8**(12): p. e83616.
37. Takai, D. and P.A. Jones, *Comprehensive analysis of CpG islands in human chromosomes 21 and 22*. Proc Natl Acad Sci U S A, 2002. **99**(6): p. 3740-5.
38. Pueyo, J.I. and J.P. Couso, *Tarsal-less peptides control Notch signalling through the Shavenbaby transcription factor*. Dev Biol, 2011. **355**(2): p. 183-93.
39. Matsumoto, A., et al., *mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide*. Nature, 2017. **541**(7636): p. 228-232.
40. Cai, B., et al., *LncRNA-Six1 Encodes a Micropeptide to Activate Six1 in Cis and Is Involved in Cell Proliferation and Muscle Growth*. Front Physiol, 2017. **8**: p. 230.
41. Mackowiak, S.D., et al., *Extensive identification and analysis of conserved small ORFs in animals*. Genome Biol, 2015. **16**: p. 179.
42. Kondo, T., et al., *Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis*. Science, 2010. **329**(5989): p. 336-9.
43. Nam, J.W., S.W. Choi, and B.H. You, *Incredible RNA: Dual Functions of Coding and Noncoding*. Mol Cells, 2016. **39**(5): p. 367-74.
44. Ulveling, D., C. Francastel, and F. Hube, *When one is better than two: RNA with dual functions*. Biochimie, 2011. **93**(4): p. 633-44.
45. Choi, S.W., H.W. Kim, and J.W. Nam, *The small peptide world in long noncoding RNAs*. Brief Bioinform, 2018.
46. Yan, J., et al., *Long Noncoding RNA H19/miR-675 Axis Promotes Gastric Cancer via FADD/Caspase 8/Caspase 3 Signaling Pathway*. Cell Physiol Biochem, 2017. **42**(6): p. 2364-2376.

## Chapter2

47. Zhang, T., K. Nie, and W. Tam, *BIC is processed efficiently to microRNA-155 in Burkitt lymphoma cells*. Leukemia, 2008. **22**(9): p. 1795-7.
48. Diederichs, S., *Micro-terminator: 'Hasta la vista, lncRNA!'*. Nat Struct Mol Biol, 2015. **22**(4): p. 279-81.
49. Dhir, A., et al., *Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs*. Nat Struct Mol Biol, 2015. **22**(4): p. 319-27.
50. Cabili, M.N., et al., *Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution*. Genome Biol, 2015. **16**: p. 20.
51. Carlevaro-Fita, J. and R. Johnson, *Global Positioning System: Understanding Long Noncoding RNAs through Subcellular Localization*. Mol Cell, 2019. **73**(5): p. 869-883.
52. Kopp, F. and J.T. Mendell, *Functional Classification and Experimental Dissection of Long Noncoding RNAs*. Cell, 2018. **172**(3): p. 393-407.
53. Luo, J., et al., *Exosomal long non-coding RNAs: biological properties and therapeutic potential in cancer treatment*. J Zhejiang Univ Sci B, 2019. **20**(6): p. 488-495.
54. Sun, Z., et al., *Emerging role of exosome-derived long non-coding RNAs in tumor microenvironment*. Mol Cancer, 2018. **17**(1): p. 82.

## CHAPTER 3| ClinicLSNP: a database hosting genetic variants in lncRNAs for cancer patients

### Abstract:

Single Nucleotide Polymorphisms (SNPs) are common genetic alterations with implications for disease susceptibility, including cancer. Research has linked cancer risk to SNPs in long non-coding RNAs (lncRNAs), especially in breast, cervical, and ovarian cancers. While databases catalog SNP associations, none directly analyze lncRNA variants from raw clinical data. Addressing this gap, we present ClinicLSNP, part of LncRBase V.2, offering a comprehensive collection of lncRNA variants from 561 female cancer RNA-Seq datasets. ClinicLSNP, accessible at <http://dibresources.jcbose.ac.in/zhumur/lncrbase2/>, is a valuable resource for researchers and clinicians studying lncRNA-SNPs in breast, cervical, and ovarian cancers.

### 1. Introduction

Single Nucleotide Polymorphisms (SNPs) are the most common genetic alterations in the human genome, and SNPs occurring within functional regions have the potential to be linked to changes in traits and disease susceptibility, including cancer[1]. Extensive research has confirmed the association between cancer risk and lncRNAs containing SNPs[2]. These cancers encompass some of the most prevalent female cancers globally, such as breast, cervical, and ovarian cancer[3-5]. In addition to experimental approaches for identifying and analyzing SNP-related lncRNAs and their roles in different cancer types, specialized groups have compiled databases that catalog SNP associations with lncRNA loci and investigate their functional consequences in various biological contexts. At present, there are several online databases hosting information on the genomic variants within lncRNA genes, such as LincSNP 2.0[6] and lncRNASNP2[7]. LincSNP 2.0[6] identifies disease-associated SNPs in lncRNAs and their Transcription Factor Binding Sites (TFBS), and lncRNASNP2[7] contains data on disease-related GWAS and COSMIC variants in human and mouse lncRNAs and predicts the impact of variants on the loss or gain of miRNA-lncRNA interactions. In these databases, disease-associated SNPs are mapped to lncRNA loci. However, there has been no database where variants within lncRNAs are directly analyzed from raw clinical data. To address this gap, we have developed ClinicLSNP (available as part of LncRBase V.2), which offers a comprehensive collection of lncRNA variants derived from 561 female cancer-specific RNA-Seq datasets, including breast, ovarian, and cervical cancer.

In summary, ClinicLSNP, which is integrated within LncRBase V.2, offers a valuable resource that will benefit both researchers and clinicians seeking information about lncRNA-SNPs specific to breast, cervical, and ovarian cancers, along with related details. This freely accessible database can be accessed at <http://dibresources.jcbose.ac.in/zhumur/lncrbase2/>

## Chapter3

### 2. Materials and Methods

#### 2.1. Raw Data corresponding to the three cancer systems:

Raw RNA sequence data for cell lines and tissues from ovarian, breast, and cervical cancers, as well as their normal counterparts, were obtained from GEO[8] and ArrayExpress[9]. Annotated SNP data for the Human genome (GRCh37) were sourced from NCBI dbSNP (version 151).[10]A comprehensive analysis was conducted on a total of 280 raw RNA-Seq tissue datasets, comprising 281 cell line datasets, encompassing 147 distinct cell lines associated with the three types of cancers, along with 125 RNA-Seq datasets for normal tissues(comprising 32 for Ovarian, 88 for Breast, and 5 for Cervix). To enhance precision, the cancer data was further categorized based on their subtypes, with 10 for ovarian, 3 for breast, and 2 for cervical subtypes.

#### 2.2. Raw data analysis and feature mapping:

##### 2.2.1. SNP detection:

For each sample, the quality of reads was assessed using FastQC, and reads with a quality score greater than 35 were retained. Adapter sequences were trimmed using Cutadapt[11]. The paired-end raw sequence reads were then aligned to the Human reference genome (hg38) using Hisat2. The resulting BAM files were sorted and indexed using SAMtools 0.1.19[12].

Prior to variant calling, the BAM files underwent preprocessing using Opossum 0.2[13], a tool that performs quality control procedures, including the removal of duplicate reads, poorly mapped reads, and secondary alignments. It also merges overlapping reads[13]. Opossum is highly compatible with the Variant Caller Platypus, which was subsequently employed for SNP and indel detection using default parameters[14]. Platypus is known for its speed and sensitivity, similar to leading variant callers in the field. Variants with a QUALITY status of "PASSED" were selected from the output.

To ensure the retention of high-quality variants, only those with more than 5 variants containing reads (TR) were considered. If TR was less than 5 in a particular sample corresponding to a specific subtype, the variant was chosen only if it was present in more than 60% of the samples within that subtype. Custom scripts were employed to merge all variant-containing output files for each subtype, followed by the removal of redundancy. Variants from the normal tissue counterparts were then filtered out from the corresponding cancerous subtypes using custom scripts.

Subsequently, the variants were annotated using dbSNP (version 151) through the "bcftools annotate" tool[12]. Finally, the variants were matched with the lncRNA transcripts present in this new version of lncRBase using bedtools Intersect[15].

### 2.2.2. Mapping SNP association with Repeat and CGI

SNPs were mapped to various repeat elements encompassing different repeat classes and families, along with CpG islands (CGI) obtained from UCSC [16]

### 2.2.3. Mapping TFBS overlap with SNP

TFBS information from the analyzedChIP-seq data was retrieved from Unibind.uio.no [17]. Custom scripts were employed to organize the data and identify SNPs located within lncRNA loci associated with the TFBS. The resulting records include details such as the TF name, position, sequence, Data accession No., and cell line/tissue information related to the mapped SNP.

### 2.2.4. Mapping TAG SNPs

Information regarding variants associated with traits was obtained from SNPsnap[18], which is an SNP-based enrichment analysis web server. This resource includes common variants (with a maximum allele frequency of >1%) from the 1000 Genomes Project Phase 3 dataset, covering three super populations: European, East Asian, and West African. The dataset comprises a total of 1,991,6464 variants, each identified by chromosomal coordinates and rs IDs. For variants with chromosomal coordinate identifiers, genomic coordinates were converted from genome assembly hg19 to hg38 using the LiftOver utility provided by UCSC (<http://genome.ucsc.edu>). Matched SNPs are designated as 'yes' under the "TagSNP" category in the SNP search result page.

### 2.2.5. Determining structure perturbation score due to presence of SNP

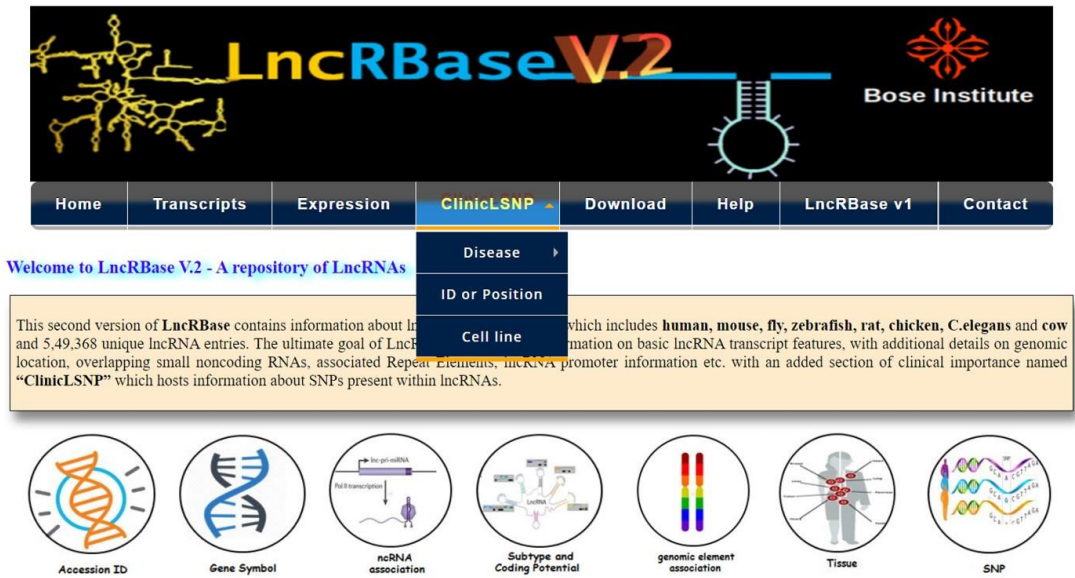
The RNAsnp tool, which can be accessed online or used as a standalone application[19], is widely utilized for predicting SNPs that disrupt RNA secondary structures. The commonly employed empirical p-value cutoff is set at  $\leq 0.2$ [19], where a lower score suggests a higher likelihood of structural effects on RNA secondary structures in the presence of a SNP.

## 2.3. Database Implementation:

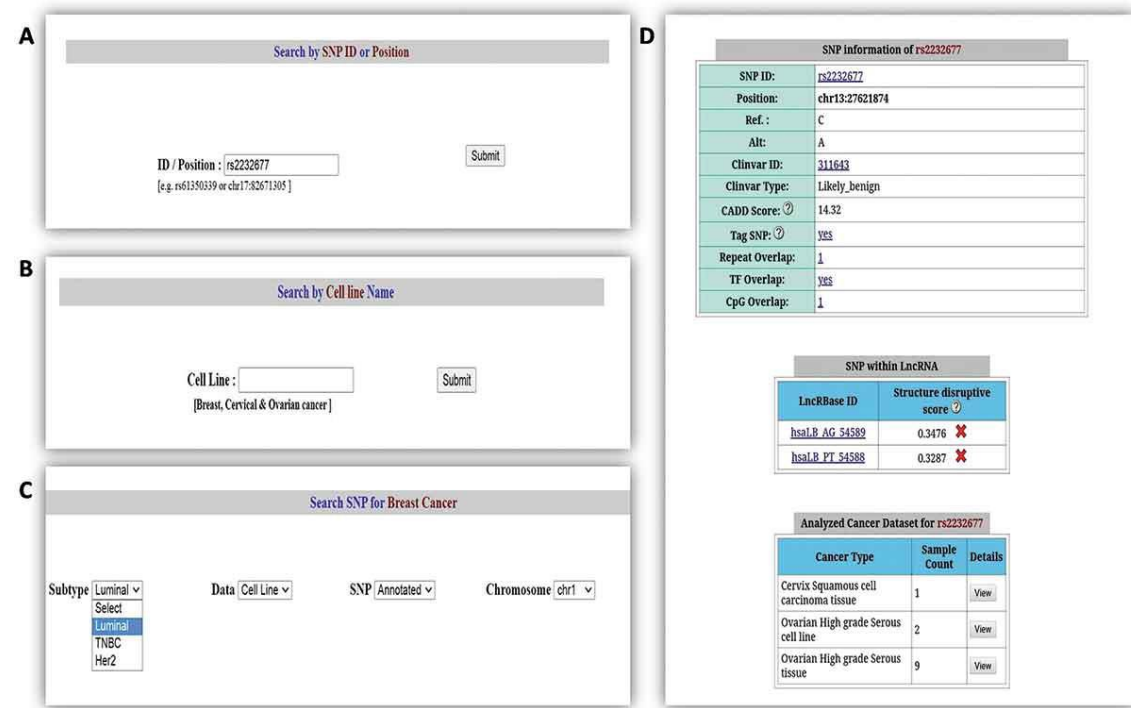
In ClinicLSNP, user query is mainly processed through simple search options, and information are displayed on the web interface after retrieving from relational databases. General Input page (**Figure 1**) displays necessary information about the SNP and

Chapter3

provides multiple options for probing into further details. Detailed Output page shows complete information about the lncRNA transcript (**Figure 2**).



**Figure1:** Several web interfaces have been designed to facilitate convenient access to ClinicLSNP: **A.** Searches using SNP ID or position.**B.** Search options are available based on cell line names for breast, cervical, and ovarian cancer.**C.** For a more focused search, there is an individual cancer subtype-specific search page.



**Figure2:** Comprehensive details regarding individual SNP IDs can be accessed through a dedicated interface.

### 2.3.1. Search options under Section “ClinicLSNP”:

#### i. Search by SNP ID/position:

Within this search option, user can input a single rsID or genomic location to retrieve comprehensive variant information, including its overlap with repeat elements, transcription factors (TFs), CGI, and trait associations. Each attribute, if present, is hyperlinked to detailed information about it. Additionally, hyperlinks to dbSNP and related Clinvar IDs (if available) are provided. Users will also find a hyperlink to the corresponding lncRNA IDs, leading to the transcript details page. Information regarding the cancer datasets where this SNP has been identified is also provided.

#### ii. Search by Cell line:

In this section, user have the ability to search for variant information within specific cell lines. The search results will provide comprehensive details about all the variants identified within the selected cell line, encompassing their locations, allele information, and corresponding lncRNA IDs. Additionally, we have included a "Download total data" option on this page, allowing users to easily download the entire dataset for further analysis or reference.

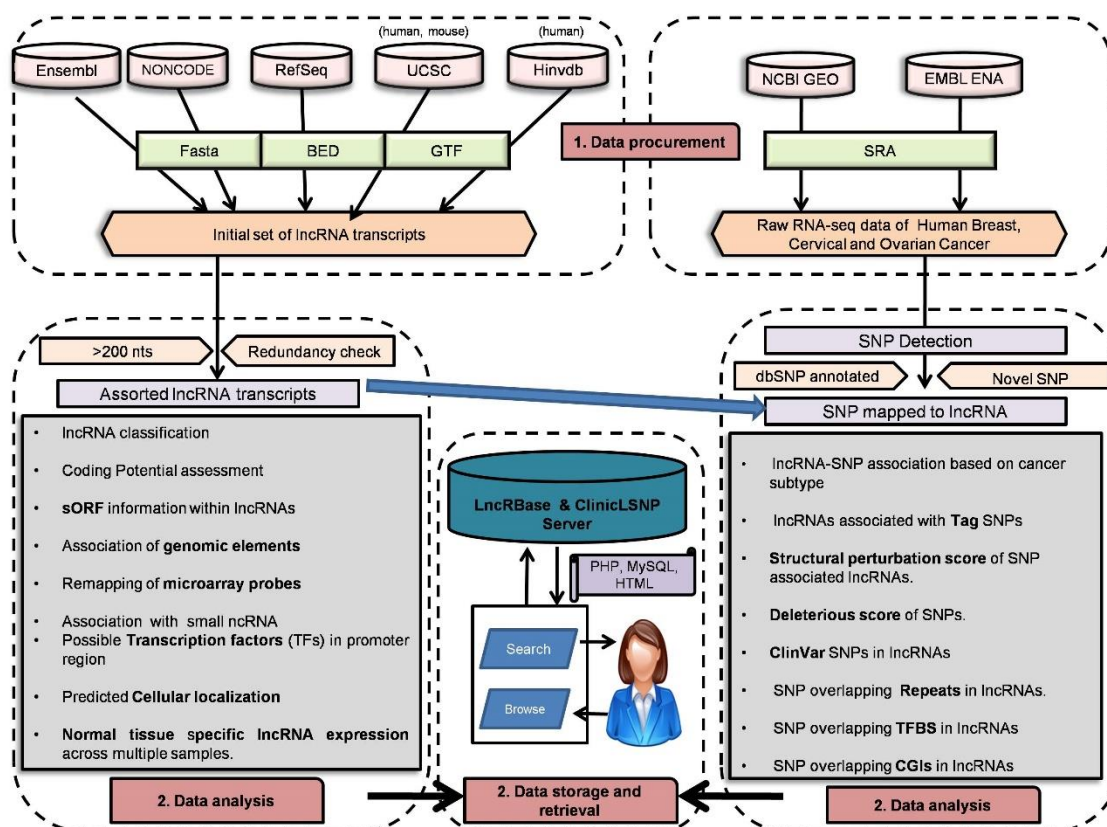
#### iii. Search by Disease:

This feature enables users to perform ClinicLSNP searches based on individual cancer types. The output results are generated according to user-provided information, including cancer subtype, data type (cell line or tissue data), variant type (novel or annotated), and genomic location. This functionality enhances the ease of searching ClinicLSNP for specific cancer-related variants.

The homepage serves as a brief overview of the entire database contents, offering users a quick tour. Detailed information on all aspects can be found on our "Information" page, accessible under the Help menu.

CliniCLSNPhas been constructed using MySQL, an open-source relational database management system. This web server operates within a Linux environment and utilizes the Apache HTTP Server, which is also free and open-source, serving as cross-platform web server software. The interface layer has been crafted using a combination of HTML, CSS, and JavaScript. To connect the database with the web interface and enable dynamic page creation, a PHP module is employed. PHP, being a server-side scripting language, facilitates the generation of dynamic pages and the retrieval of data from MySQL to produce the desired output, as depicted in **Figure 3**.





**Figure3:** The workflow of ClinicLSNP involves several key steps.

### 3. Results and Discussion:

ClinicLSNP serves as a repository of lncRNA-SNP data specifically related to prevalent female cancer systems, including breast, ovarian, and cervical cancer. To compile this data, we conducted a comprehensive analysis of 561 RNA-seq datasets, comprising 280 tissue and 281 cell line samples, across these three cancer types in humans. Our approach involved mapping variants transcriptome-wide within lncRNA loci, with a subsequent filtering step to retain cancer-specific variants. Specifically, we removed variants originating from 88 normal breast tissue samples, 32 normal ovary tissue samples, and 5 normal cervix tissue samples, ensuring that the dataset exclusively contained variants associated with these three female cancer types. These variants were then cross-referenced with existing entries in dbSNP[10] allowing us to categorize them as either novel or annotated variants based on their mapping status with existing dbSNP entries. In total, ClinicLSNP provides information on 571,886 annotated and 243,254 novel unique variants, all mapped to 172,404 lncRNA transcripts associated with these three cancer types. A summary of the results can be found in Table **Table 1**.



Cancer	Subtype	Data type	Sample analyzed	Novel SNPs	Annotated SNPs
BREAST	Her2	cell line	31	16108	55312
		tissue	22	21358	34029
	Luminal	cell line	66	27827	65820
		tissue	102	87458	131640
	TNBC	cell line	70	23774	73367
Cancer	Subtype	Data type	Sample analyzed	Novel SNPs	Annotated SNPs
OVARY	Carcinoma	cell line	11	10709	34663
	Clear cell Carcinoma	cell line	11	16425	43797
	Endometrial carcinoma	cell line	5	8081	20549
	High Grade Serous	cell line	28	25790	69916
		tissue	87	27774	141417
	Low Grade Serous	cell line	27	21613	59896
	Mucinoid carcinoma	cell line	3	2440	8841
	Primary Peritoneal	tissue	6	1801	4534
	Sarcoma	cell line	1	654	3174
	Stromal carcinoma	cell line	1	1596	5829
	Teratocarcinoma	cell line	4	2110	10488
Cancer	Subtype	Data type	Sample analyzed	Novel SNPs	Annotated SNPs
CERVIX	Adenocarcinoma	cell line	8	7655	41826
	Squamous cell carcinoma	cell line	5	9869	55921
		tissue	4	1612	13332

**Table1:** A summary of the data and results obtained from ClinicLSNP is presented.

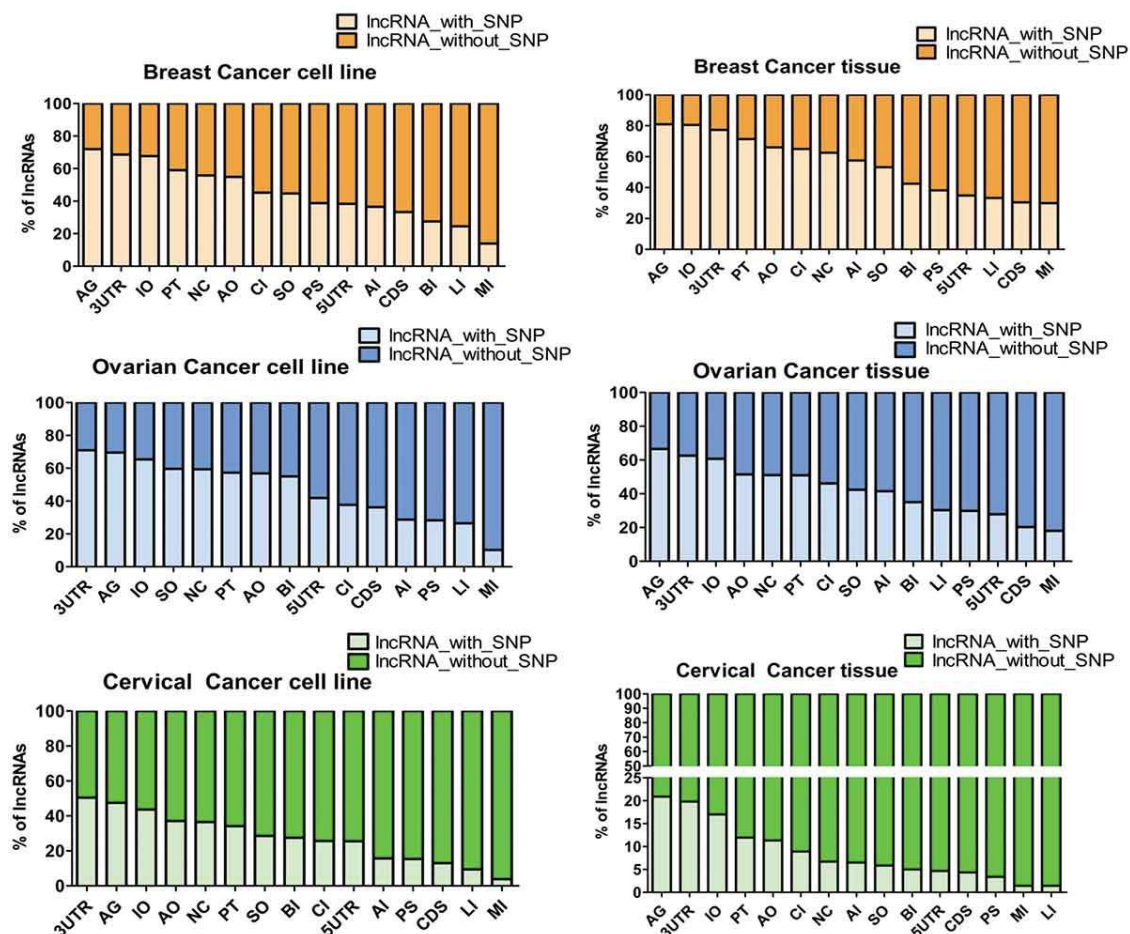
Additionally, ClinicLSNP provides valuable information about lncRNA-SNPs by offering insights into their overlap with various genomic features and functional characteristics. Specifically, the database includes details on: (i) Repeat elements, (ii) CGI, (iii) TFBS within lncRNA loci (iv) SNP localization in trait-associated LD (Linkage Disequilibrium) region, (v) predicted potentially pathogenic variants and (vi) SNP effect on lncRNA secondary structure.

### 3.1. Mapping of detected SNPs within lncRNA:

The distribution of SNPs associated with Human lncRNAs exhibits intriguing patterns across different lncRNA biotypes within breast, cervical, and ovarian cancer datasets.

## Chapter3

Despite the relatively low representation of the 3'UTR (3UTR) and Ambiguous (AG) biotypes, comprising only 0.3% and 6% of Human lncRNAs, respectively (as illustrated in **Figure4**), these biotypes notably harbor the majority of SNPs, accounting for over 60% of the SNPs identified in both cell line and tissue datasets across the three cancer types. In contrast, long intergenic non-coding RNAs (lincRNAs), which constitute the most abundant Human lncRNA biotype at approximately 40%, surprisingly exhibit a lower SNP prevalence, with less than 30% of them containing SNPs. This biotype-specific distribution of SNPs within lncRNA loci raises intriguing questions about potential associations between biotype-specific functional roles of lncRNAs and the presence of SNPs in their sequences.



**Figure 4:** The distribution of lncRNAs associated with SNPs across various biotypes.

**Abbreviations used:** NC: Non-Coding; MI: MIscRNA; SO: Sense Overlapping; PS: Pseudogenes; CI: Completely Intronic; 3UTR: 3'UTR overlapping; PT: Processed Transcript; 5UTR: 5'UTR overlapping; CDS: CDS overlapping; IO: Intron Overlapping; LI: Long Intergenic; AO: Antisense Overlapping; AI: Antisense Intronic; AG: Ambiguous; BI: Bidirectional lncRNAs.

### 3.2. Clinical relevance of the lncRNA-SNPs in the three cancer systems:

We have assessed the clinical relevance of lncRNA-SNPs identified in breast, cervical, and ovarian cancer by mapping them to known rs IDs and leveraging databases such as ClinVar[20] and SNPsnap's SNP Annotation Database[18]. Furthermore, we employed the CADD tool [21] to predict the clinical significance of both novel and annotated lncRNA variants. The outcomes of these analyses are depicted in **Figure 5**, shedding light on the potential clinical implications of these lncRNA-SNPs in the context of the three cancer systems.

### 3.2.1. Mapping with ClinVarIDs:

ClinVar database[20] serves as a repository of clinically relevant genetic variants, categorizing them into 13 distinct types based on their clinical significance. To investigate the clinical relevance of the dbSNP-annotated lncRNA variants identified in our analysis, we cross-referenced them with ClinVar IDs. The results revealed that, for breast, ovarian, and cervical cancer cell lines, approximately 2.5%, 2.3%, and 3% of lncRNA-SNPs, respectively, were mapped to ClinVar IDs. In tissue datasets, the mapping percentages for breast, ovarian, and cervical cancer lncRNA-SNPs with ClinVar IDs were approximately 2.1%, 2.6%, and 5.4%, respectively (**Figure 5**).

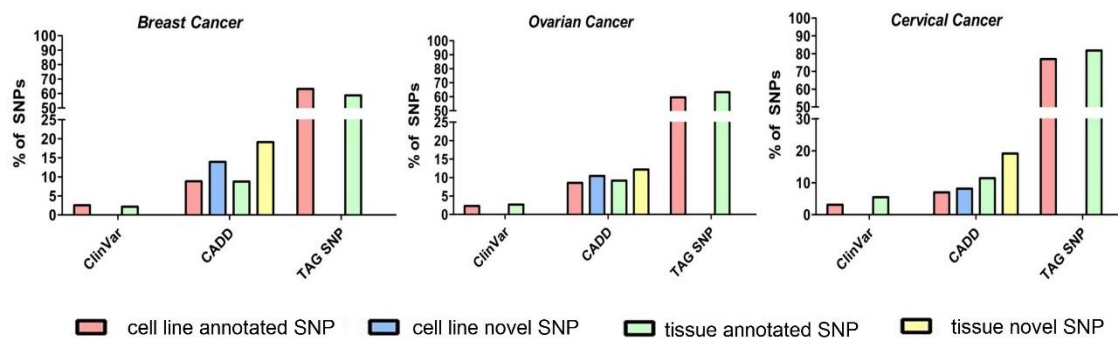
### 3.2.2. Mapping TAG SNPs:

A cluster of SNPs or haplotypes that are inherited together within a genomic region of high linkage disequilibrium often exhibits associations with diseases or phenotypic traits [22]. One representative SNP from such a cluster can be highly valuable for risk association studies, obviating the need to genotype every single variant. Remarkably, our analysis revealed that 57.9%, 58.2%, and 76.6% of the dbSNP-annotated lncRNA variants matched as TAG SNPs for breast, ovarian, and cervical cancer cell lines and tissues, respectively.

Furthermore, as depicted in the graph in **Figure 5**, over 60% of dbSNP-annotated variants were found to be associated with traits in all three disease systems, with lncRNA variants associated with cervical cancer (in both cell line and tissue datasets) exhibiting the highest enrichment.

### 3.2.3. Identification of potentially pathogenic variants:

The deleteriousness of all the variants was assessed using CADD v1.5 [21], a tool that incorporates over 60 genomic annotations to score SNPs and indels across the genome. As recommended by Rentzsch et al., 2019[21], the suggested cutoff for identifying potentially pathogenic variants falls between 10 and 15, with a scaled 'C score' greater than 10 indicating the top 10% deleterious variants, 'C score' greater than 20 signifying the top 1% deleterious variants, and so forth. In summary, 13.9%, 9.2%, and 11.7% of the variants analyzed in breast, cervical, and ovarian cancer, respectively, fall within the deleteriousness range of 15 to 60. For detailed statistics on the enrichment of annotated and novel SNPs in cell lines and tissues, please refer to **Figure 5**.



**Figure 5:** Clinical significance of lncRNA-SNPs in breast, ovarian, and cervical cancer systems: **A.** SNP Matches in ClinVar Database: SNPs that correspond to entries in the ClinVar database. **B.** CADD Scores: SNPs with notable CADD scores falling within the range of 15 to 60, indicating potential functional relevance. **C.** TAG SNPs: SNPs associated with traits based on the SNP annotation database from SNPsnap.

### 3.3. Regulatory SNPs (rSNPs) within lncRNA loci:

lncRNAs are renowned for their involvement in the transcriptional regulation of nearby or distant genes through the recruitment of chromatin modifiers or other regulatory complexes[23]. The presence of regulatory SNPs (rSNPs) can lead to the loss or gain of binding sites, disrupting interactions and impeding downstream effects. Therefore, information about rSNPs within lncRNA loci can shed light on the altered functions of these lncRNAs in transcriptional regulation. There is substantial literature evidence supporting the impact of variants on transcriptional regulation, including the following examples:

Buroker conducted a review outlining disease developments associated with variants that induce de-novo or loss of TFBS[24]. Additionally, Liu and colleagues identified genetic variants associated with breast cancer that affect transcription factor binding[25]. These studies underscore the significant role of variants in modulating transcriptional regulation and their potential impact on disease development.

In this study, the presence of rSNPs within lncRNA loci, specifically within TFBS, has been investigated. Utilizing Chip-seq datasets [17] for TFBS information, we identified 39,798, 10,633, and 28,407 such rSNPs within lncRNA loci in breast, ovarian, and cervical cancer, respectively. In essence, the maximum number of rSNPs among dbSNP-annotated lncRNA variants was found in cervical cancer, whereas those in novel variants were observed in breast cancer (Figure 6). Human Papillomavirus (HPV), which is known to cause cervical cancer, has been reported to induce mutations within immune and DNA repair-related genes during integration into the human genome[26]. Therefore, polymorphisms within regulatory regions, as identified within lncRNA loci in our analysis, could potentially be a mechanism employed by the virus to modulate the host transcriptome environment.

### 3.4. CpG-SNPs within lncRNA-loci:

CpG islands are short stretches of DNA characterized by a high frequency of the dinucleotide CG compared to other regions. These islands can undergo methylation, specifically at the cytosines in CpG dinucleotides, resulting in the formation of transcriptional repression complexes that effectively shut down gene expression [27]. CpG-SNPs, on the other hand, are point mutations occurring at CpG sites, and they can potentially impact methylation patterns by either generating or destroying CpG dinucleotides. Notably, certain CpG-SNPs like rs7766585 have been strongly associated with breast cancer risk [28], and the influence of SNPs within CGI on oncogenes and tumor suppressor genes has been well-documented [29].

In our analysis, we identified specific genomic regions within lncRNA loci, such as *chr15:100554315-100559288*, *chr21:8431968-8441142*, and *chr14:19300376-19302642*, to be enriched with SNPs and also situated within CGI. Additionally, we found that 4.2%, 4.5%, and 6.8% of lncRNAs harboring SNPs are located within CpG regions in breast, ovarian, and cervical cancer, respectively.

We have provided information regarding CGIs that overlap with SNPs within lncRNA loci, including the CGI name, length, GC percentage, CpG percentage, and CpG density. These details provide insights into the characteristics of CGIs associated with lncRNA variants.

Furthermore, the distribution of novel and annotated lncRNA variants within CGIs has been analyzed in a tissue and cell line-specific manner for breast, ovarian, and cervical cancer. The results of this analysis have been visualized in **Figure 6**, allowing for a comprehensive understanding of the distribution patterns of these variants within CGI regions across the three cancer types.

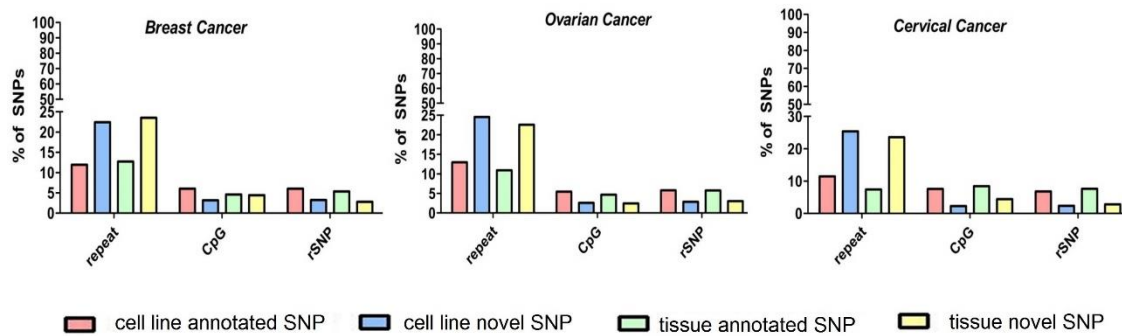
### 3.5. SNP associated with Repeat Elements:

Initially regarded as 'junk DNA,' repeat elements, which make up 50% of the genome, primarily consist of interspersed repeat sequences. These elements are recognized for their regulatory and structural functions [30]. In a study conducted by Payer et al., it was observed that 44 Alu insertion polymorphisms were in strong Linkage Disequilibrium ( $r^2 > 0.7$ ) with trait-associated SNPs, highlighting the potential functional significance of these repeat elements [31]. Additionally, lncRNAs have been shown to regulate Staufen-mediated mRNA decay by forming imperfect base pairings with 3'UTR Alu repeat elements [32].

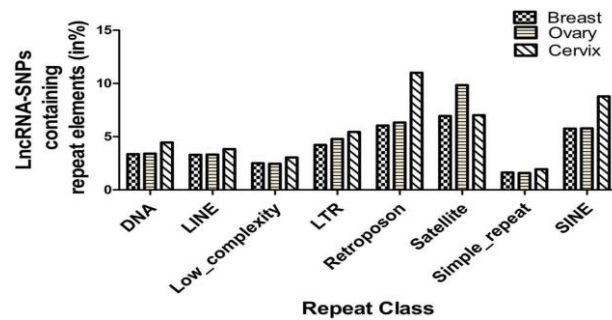
Our analysis reveals the percentage of SNPs associated with repeat elements within lncRNA loci (**Figure 6**) and the distribution of repeat classes (as a percentage) within SNPs located in lncRNA loci in the three cancer systems (**Figure 7**). In the context of these three cancers, lncRNA-SNPs exhibit the highest enrichment with SVA Retroposon and Satellite class repeats. SINE repeat elements follow closely, with others being less prominent. This enrichment of an lncRNA-SNP with a repeat element suggests the

## Chapter3

potential to influence the function of the corresponding lncRNA and warrants further investigation.



**Figure 6:** Characteristics of lncRNA locus-linked SNPs in Breast, Ovarian, and Cervical cancer systems: **A.** Repeat: SNPs overlapping with repeat elements. **B.** CpG: SNPs located within CGI. **C.** rSNP: SNPs situated within transcription factor binding sites within lncRNA loci.



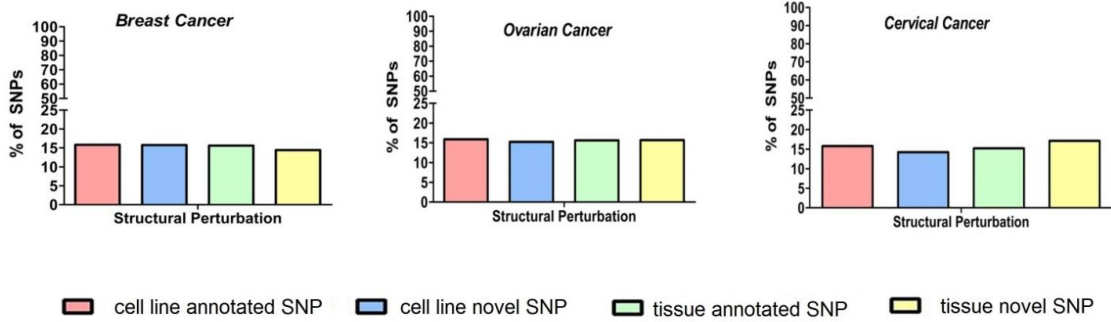
**Figure 7:** Distribution of lncRNA-SNPs associated with repeat elements varied across distinct classes of repeats.

### 3.6. Effect of SNP on lncRNA secondary structure:

The significance of lncRNA structural integrity for their biological functions has been well-established in various studies. For instance, the recruitment of LSD1 and PCR2 complexes by HOTAIR is intricately linked to its conserved secondary structure[33]. In *Drosophila*, a combinatorial point mutation within the stem-loop structure of the Rox1 lncRNA led to the inhibition of MSL complex binding, resulting in the loss of dosage compensation and male lethality[34]. Furthermore, investigations using chemical and enzymatic probing techniques to examine the conserved structures of lncRNAs have shown that both secondary and tertiary structures are highly conserved and closely associated with their biological roles[35]. Consequently, even a single-base pair alteration within lncRNAs can induce structural instability, potentially leading to functional changes. Our analysis, conducted using the RNAsnp tool[19]revealed that approximately 15.4% of SNP-containing lncRNAs in breast cancer, 15.73% in ovarian cancer, and 15.5% in cervical cancer are predicted to undergo structural disruption due to the presence of SNPs. This prediction is based on a specific base pair probability cutoff score (p-



value), indicating that a considerable proportion of lncRNAs harboring SNPs in each disease system may experience structural perturbations, as illustrated in **Figure 8**.



**Figure 8:** SNPs that disrupt the secondary structure of lncRNAs have been observed in breast, ovarian, and cervical cancer systems, with a perturbing score indicating statistical significance at p-values of 0.2 or less.

### 3.7. Case study with ClinicLSNP results:

To validate the reliability of our results, we conducted case studies involving specific lncRNA-SNPs hosted in "ClinicLSNP," as outlined in **Table 2**. The outcomes of these case studies align with experimentally validated data:

Disease System	LncRNA	SNP	Ref allele	Alt allele
Epithelial ovarian Carcinoma	HOXA11-AS	rs17427875	A	T
Breast Cancer	CCAT2	rs6983267	G	T
Cervical Cancer	HOTAIR	rs2366152	A	G

**Table 1:** Analyzing specific cases that involve lncRNA-SNPs documented within the 'ClinicLSNP' database underscores their consistency with experimentally verified findings.

- (i) In ovarian cancer, the homeobox A (HOXA) region, which plays a critical role in embryo development, is regulated by protein-coding genes. Richard et al. demonstrated that the SNP **rs17427875 (A>T)**, located within the exon of the downregulated lncRNA HOXA11-AS, is associated with a reduced risk of ovarian cancer. This suggests that HOXA11-AS may function as a tumor suppressor in this context[36].
- (ii) In breast cancer, Redis et al. investigated the expression, function, and clinical relevance of CCAT2, an lncRNA overlapping with SNP **rs6983267 (G>T)**. Their study demonstrated a correlation between this SNP and breast cancer, and our analysis supports this association[37].
- (iii) Regarding virus-induced cervical cancer, Saha et al. identified the role of SNP **rs2366152 (A>G)** in affecting the secondary structure of HOTAIR, an lncRNA involved in cellular chromatin reprogramming. Our analysis detected this SNP within a similar

## Chapter3

context, and all three of these variants were annotated as trait-associated in our study, consistent with the experimental findings[38].

Our findings align with these validated results, as we have detected the same variants in similar systems as those observed in these experiments. Importantly, all three of these variants were also annotated as trait-associated in our analysis, further substantiating their potential clinical relevance and emphasizing the reliability of our results.

### 4. Conclusion:

ClinicLSNP, as a part of LncRBase V.2, serves as a valuable repository of lncRNA variants encompassing three prominent female cancer systems. Within each cancer subtype, we have identified a comprehensive set of novel and annotated lncRNA-SNPs. The annotation process, which includes ClinVar integration for annotated IDs, provides insights into their clinical significance. The CADD score offers a means of assessing variant functionality, aiding in the identification of functional SNPs. Variants marked as 'Tag' are indicative of their association with traits within the Linkage Disequilibrium region. The presence of lncRNA variants within CGI, TFBS and STR can have substantial downstream effects. The structure disruptive score of lncRNA-SNP pairs provides insights into the potential structural perturbation efficiency of a given variant at a specific position within an lncRNA transcript. However, it's important to note that this score cannot be applied to calculate structural perturbation caused by insertions or deletions (in-dels), resulting in "NA" p-values for associations involving such variants. ClinicLSNP represents a valuable and easily accessible resource for clinically relevant lncRNA variants, making significant contributions to the functional investigation of disease-associated lncRNAs.

### 5. Availability

CliniCLSNP hosted within LncRBaseV.2is freely available at <http://dibresources.jcbose.ac.in/zhumur/lncbase2/>. The files are available for free download and can be utilized in compliance with the GNU Public License terms.

### 6. References:

1. Reich, D.E., S.B. Gabriel, and D. Altshuler, *Quality and completeness of SNP databases*. Nat Genet, 2003. **33**(4): p. 457-8.
2. Hajjari, M. and S. Rahnama, *Association Between SNPs of Long Non-coding RNA HOTAIR and Risk of Different Cancers*. Front Genet, 2019. **10**: p. 113.
3. Lin, Y., et al., *Polymorphisms of long non-coding RNA HOTAIR with breast cancer susceptibility and clinical outcomes for a southeast Chinese Han population*. Oncotarget, 2018. **9**(3): p. 3677-3689.
4. Dong, J., et al., *Long non-coding RNAs on the stage of cervical cancer (Review)*. Oncol Rep, 2017. **38**(4): p. 1923-1931.
5. Worku, T., et al., *Long Non-Coding RNAs: the New Horizon of Gene Regulation in Ovarian Cancer*. Cellular Physiology and Biochemistry, 2017. **44**(3): p. 948-966.



6. Ning, S., et al., *LincSNP 2.0: an updated database for linking disease-associated SNPs to human long non-coding RNAs and their TFBSs*. Nucleic Acids Res, 2017. **45**(D1): p. D74-D78.
7. Miao, Y.R., et al., *lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs*. Nucleic Acids Res, 2018. **46**(D1): p. D276-D280.
8. Barrett, T. and R. Edgar, *Gene expression omnibus: microarray data storage, submission, retrieval, and analysis*. Methods Enzymol, 2006. **411**: p. 352-69.
9. Kanz, C., et al., *The EMBL Nucleotide Sequence Database*. Nucleic Acids Res, 2005. **33**(Database issue): p. D29-33.
10. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation*. Nucleic Acids Res, 2001. **29**(1): p. 308-11.
11. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads*. 2011, 2011. **17**(1): p. 3.
12. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
13. Oikkonen, L. and S. Lise, *Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection*. Wellcome Open Res, 2017. **2**: p. 6.
14. Rimmer, A., et al., *Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications*. Nat Genet, 2014. **46**(8): p. 912-918.
15. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features*. Bioinformatics, 2010. **26**(6): p. 841-2.
16. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool*. Nucleic Acids Res, 2004. **32**(Database issue): p. D493-6.
17. Gheorghe, M., et al., *A map of direct TF-DNA interactions in the human genome*. Nucleic Acids Res, 2019. **47**(4): p. e21.
18. Pers, T.H., P. Timshel, and J.N. Hirschhorn, *SNPsnap: a Web-based tool for identification and annotation of matched SNPs*. Bioinformatics, 2015. **31**(3): p. 418-20.
19. Sabarinathan, R., et al., *RNAseq: efficient detection of local RNA secondary structure changes induced by SNPs*. Hum Mutat, 2013. **34**(4): p. 546-56.
20. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype*. Nucleic Acids Res, 2014. **42**(Database issue): p. D980-5.
21. Rentzsch, P., et al., *CADD: predicting the deleteriousness of variants throughout the human genome*. Nucleic Acids Res, 2019. **47**(D1): p. D886-D894.
22. International HapMap, C., *The International HapMap Project*. Nature, 2003. **426**(6968): p. 789-96.
23. Fernandes, J.C.R., et al., *Long Non-Coding RNAs in the Regulation of Gene Expression: Physiology and Disease*. Noncoding RNA, 2019. **5**(1).
24. Buroker, N.E., *Regulatory SNPs and transcriptional factor binding sites in ADRBK1, AKT3, ATF3, DIO2, TBXA2R and VEGFA*. Transcription, 2014. **5**(4): p. e964559.
25. Liu, Y., et al., *Identification of breast cancer associated variants that modulate transcription factor binding*. PLoS Genet, 2017. **13**(9): p. e1006761.
26. Joo, J., et al., *The association of integration patterns of human papilloma virus and single nucleotide polymorphisms on immune- or DNA repair-related genes in cervical cancer patients*. Scientific Reports, 2019. **9**(1): p. 13132.
27. Li, E. and Y. Zhang, *DNA methylation in mammals*. Cold Spring Harb Perspect Biol, 2014. **6**(5): p. a019133.
28. Harlid, S., et al., *A candidate CpG SNP approach identifies a breast cancer associated ESR1-SNP*. Int J Cancer, 2011. **129**(7): p. 1689-98.
29. Samy, M.D., et al., *Impact of SNPs on CpG Islands in the MYC and HRAS oncogenes and in a wide variety of tumor suppressor genes: A multi-cancer approach*. Cell Cycle, 2016. **15**(12): p. 1572-8.
30. Gemayel, R., et al., *Variable tandem repeats accelerate evolution of coding and regulatory sequences*. Annu Rev Genet, 2010. **44**: p. 445-77.

## Chapter3

31. Payer, L.M., et al., *Structural variants caused by Alu insertions are associated with risks for many human diseases*. Proc Natl Acad Sci U S A, 2017. **114**(20): p. E3984-E3992.
32. Gong, C. and L.E. Maquat, *lncRNAs transactivate STAUI-mediated mRNA decay by duplexing with 3' UTRs via Alu elements*. Nature, 2011. **470**(7333): p. 284-8.
33. Wu, L., et al., *Binding interactions between long noncoding RNA HOTAIR and PRC2 proteins*. Biochemistry, 2013. **52**(52): p. 9519-27.
34. Ilik, I.A., et al., *Tandem stem-loops in roX RNAs act together to mediate X chromosome dosage compensation in Drosophila*. Mol Cell, 2013. **51**(2): p. 156-73.
35. Li, R., H. Zhu, and Y. Luo, *Understanding the Functions of Long Non-Coding RNAs through Their Higher-Order Structures*. Int J Mol Sci, 2016. **17**(5).
36. Richards, E.J., et al., *A functional variant in HOXA11-AS, a novel long non-coding RNA, inhibits the oncogenic phenotype of epithelial ovarian cancer*. Oncotarget, 2015. **6**(33): p. 34745-57.
37. Redis, R.S., et al., *CCAT2, a novel long non-coding RNA in breast cancer: expression study and clinical correlations*. Oncotarget, 2013. **4**(10): p. 1748-62.
38. Sharma Saha, S., et al., *Identification of genetic variation in the lncRNA HOTAIR associated with HPV16-related cervical cancer pathogenesis*. Cell Oncol (Dordr), 2016. **39**(6): p. 559-572.

## CHAPTER 4| Shared lncRNA Variants in Female Cancers

### Abstract:

Long non-coding RNA (lncRNA), transcripts exceeding 200 nucleotides that do not serve as protein synthesis templates, were once considered non-functional "junk" RNA. However, they are now recognized for their diverse roles in cellular processes, including cancer. Single nucleotide polymorphisms (SNPs), prevalent genetic variations in the human genome, are often found in non-coding regions and may influence disease susceptibility, including cancer. SNPs within lncRNAs can lead to functional changes thereby influencing disease phenotypes. Breast, cervical, and ovarian cancers, prominent female malignancies, have shown genetic variations in lncRNAs associated with cancer risk. In this study, we aimed to investigate lncRNA loci harboring SNPs linked to breast, cervical, and ovarian cancers. Analyzing public datasets specific to these cancers revealed a novel SNP within the lncRNA MIR4435-2HG which is present in breast, ovary and cervical cancer. Further wet bench validation of this provides the clue towards a long-drawn influence of this SNP on MIR4435-2HG expression, emphasizing the importance of previously overlooked genetic variations in shaping lncRNA expression patterns.

### 1. Background and Objective of the study

Long non-coding RNA (lncRNA) are RNA transcripts exceeding 200 nucleotides in length that do not act as templates for protein synthesis. They are mostly transcribed by RNA polymerase II, which then undergo splicing and polyadenylation[1]. Typically expressed at lower levels, these molecules are predominantly located within the nucleus, in contrast to the more abundant mRNAs primarily found in the cytoplasm [2]. Once considered non-functional "junk" RNA, lncRNAs are now acknowledged for their diverse roles in various cellular processes. Representing a significant portion of noncoding genes in mammals and other eukaryotes, lncRNAs play diverse roles in cellular systems, including cancer[3, 4].

Single Nucleotide Polymorphisms (SNPs) represent the most prevalent genetic variations in the human genome and are often situated in functional regions, potentially influencing phenotypic traits and disease susceptibility, including cancer[5, 6]. Notably, a significant proportion of SNPs are located within non-coding regions of the genome [7], indicating their potential regulatory role in disease outcomes [8]. Many SNPs identified in Genome-Wide Association Studies (GWASs) are situated in intergenic and intronic regions [9] highlighting their crucial involvement in disease predisposition. Some GWAS SNPs have been reported to affect the expression of genes transcribed from nearby or distant loci, termed expression Quantitative Trait Loci or eQTL[7]. Furthermore, SNPs within lncRNAs can induce functional changes in these molecules, potentially leading to

## Chapter 4

alterations in phenotypes. Given the significance of structural features in lncRNA functions[10, 11], SNPs within lncRNA transcripts may influence their secondary structure, impacting their stability and functionality. Consequently, these changes can disrupt the interactions between lncRNAs and their target genes, ultimately influencing disease phenotypes[12]

Breast, cervical, and ovarian cancers stand as some of the most prevalent female cancers worldwide[13-16]. Various reports suggest the importance of genetic variations in these cancer systems. Notably, the HOXA region, a regulator of embryogenesis and ovarian carcinogenesis, features lncRNAs like HOXA10-AS, HOXA11-AS, and HOTTIP. In serous epithelial ovarian cancer (EOC), a variant within HOXA11-AS (rs17427875) has been linked to a decreased risk of EOC, hinting at a potential tumor suppressor role for this lncRNA[17]. Conversely, HOTAIR has been implicated in elevating the risk of breast cancer. In a study involving female breast cancer patients, three specific SNPs were identified in association with cancer risk. Notably, rs920778 and rs12826786 were linked to an increased risk of breast cancer, while a negative correlation was observed for the rs1899663 SNP[18].

Furthermore, specific SNPs within lncRNAs can function as expression Quantitative Trait Loci (eQTLs) for protein-coding genes. For instance, SNPs within the lncRNA AC008392.1 can act as eQTLs for CARD8, a gene linked to immune responses and apoptosis, potentially influencing the risk of virus-induced cervical cancer[19]. Similarly, ZNRD1-AS1, an lncRNA antisense to ZNRD1 implicated in immune responses against HPV infection and cervical cancer, harbors several SNPs. Notably, some of these SNPs (rs3757328, rs6940552, and rs9261204) have been associated with a decreased risk of cervical cancer[20]. The rs6983267 SNP within the CCAT2 loci has been linked to increased MYC expression and is thought to contribute to the promotion of proliferation and rapid growth of cervical squamous cell carcinoma. (SCC) compared to lower-grade carcinomas[21]. This genetic variation may contribute to the more aggressive behavior observed in certain cervical SCC cases [21].

Reports have highlighted a shared genetic architecture underlying basal-like breast tumors and high-grade serous ovarian tumors [22], suggesting a related etiology and potential therapeutic opportunities for both cancer types. Recognized risk factors linking both cancers involve gene mutations, particularly in BRCA1 and BRCA2[23, 24]. Additionally, a comprehensive panel comprising over 15 genes for hereditary Breast and Ovarian cancer (HBOC) risk has been established, consolidating them into a single, multiple-gene test[25]. Breast and ovarian cancers are occasionally grouped together due to the similar effects of estrogen or testosterone exposure [26]. COGS (Centre for Cancer Genomics epidemiology) identified 2 such regions with variants associated with these two cancers[26]. Conversely, in the case of cervical cancer, HPV infection is a necessary but not sole risk factor for its onset and progression. Studies on Heritable Cervical Cancer suggest that female offspring and siblings demonstrate a relative risk (RR) of 1.5–2.3 for developing the cancer[27]. Numerous research groups have explored the relevance of genetic variations in cervical cancers through candidate gene association and genome-

wide association studies, establishing a link between the genetic makeup of disease loci and susceptibility to malignancy[28-30]. This raises the intriguing question of whether these gynecological malignancies and breast cancer share a common network of cellular events influenced by genetic variations within the genome.

In this work, our primary aim was to scrutinize lncRNA loci harboring SNPs linked to a predisposition for breast, cervical, and ovarian cancers. Utilizing publicly available cancer datasets specific to these three distinct systems, our objective was to pinpoint SNPs and differentially regulated lncRNAs that are common across these cancer types. Our thorough analysis brought to light a previously unreported SNP within the lncRNA MIR4435-2HG, shedding new insights into its potential role in cancer development. Wet bench validation depicts the influence of the identified SNP on the expression level of MIR4435-2HG. Our study puts up the possibility of previously overlooked genetic variations in shaping the expression patterns of lncRNAs.

## 2. Materials and Methods

### 2.1. Raw Data corresponding to the three cancer systems:

The raw RNA sequence data for patients with ovarian, breast (including Luminal, TNBC, and Her2 subtypes), and cervical cancers, along with samples from healthy tissues, were acquired from NCBI GEO [31] and ArrayExpress[32]. Annotated SNP data for the Human genome (GRCh37) were obtained from NCBI dbSNP (version 151).[33]. Following an extensive data search, a total of 299 samples from Ovarian High Grade Serous, 345 from Breast Cancer (comprising 136 Luminal, 159 TNBC, and 64 Her2 subtypes), and 31 from Cervical Squamous cell carcinoma patients were successfully obtained. The details of the input data are summarized in **Table 1**.

Tissue	Cancer subtype	Cancer Tissue sample	Healthy tissue sample
Cervical	Squamous cell Carcinoma	31	31
Ovarian	Ovarian High Grade Serous	299	56
Breast	Luminal TNBC Her2	345	140

**Table1:** A summary of the input data utilized for the detection of common DE lnc-SNPs.

## Chapter4

### 2.2. Detection of Differentially Expressed LncRNAs in the three cancer systems

FastQC v.0.11.7 was employed to assess the quality of reads for each sample, and only those with a quality score exceeding 30 were preserved. Cutadapt v1.16 [34]. was utilized to trim adapter sequences. Subsequently, the paired-end raw sequence reads underwent alignment to the Human reference genome (hg38) using HISAT2 2.1.0 [35]. The resultant BAM files were organized and indexed using SAMtools 0.1.19 [36]. Subsequently, transcript assembly was performed by StringTie v1.3.4d[35] and differential analysis against their corresponding normal counterparts was conducted using Ballgown[35]. The "stat test" function within Ballgown was designed to address various concerns, including the potential for batch effects stemming from multiple data sources. Differentially regulated lncRNAs was identified with a fold change (FC) cut-off of  $\geq 1.5$  and a q-value of  $\leq 0.05$ .

### 2.3. SNP detection in patient samples:

Before variant calling, the BAM files underwent preprocessing through Opossum 0.2, [37] a tool designed for quality control procedures. This process included the elimination of duplicate or poorly mapped reads, and non specific alignments, as well as the merging of overlapping reads. Opossum is seamlessly compatible with the Variant Caller Platypus, which was then used for SNP and indel detection with default parameters [38]. Platypus is renowned for its speed and sensitivity, comparable to other top-tier variant callers in the field. Variants with a QUALITY status of "PASSED" were chosen from the output.

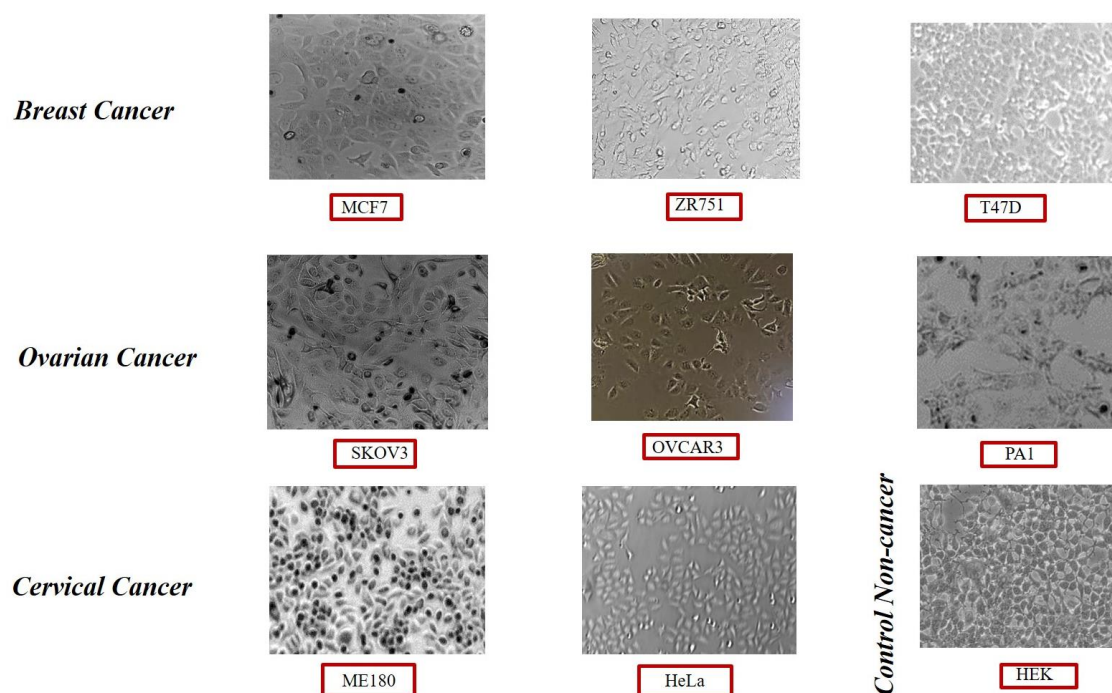
In order to uphold the selection of high-quality variants, priority was given to those with more than 5 variants containing reads (TR). If the total reads (TR) fell below 5 for a specific sample linked to a particular subtype, the variant was incorporated only if it appeared in over 60% of the samples within that subtype. Custom scripts were utilized to consolidate all output files containing variants for each subtype, and redundancy was subsequently eliminated. Following this, variants from normal tissue counterparts were filtered out from the corresponding cancerous subtypes using custom scripts. Following the variant filtering process, annotation of the variants was carried out using dbSNP (version 151) through the "bcftools (v1.9) annotate" tool [36].

### 2.4. Detection of common DE lncRNA-SNPs

The detected variants were matched with the DE lncRNA transcripts using bedtools(v2.26.0) Intersect [39]. Finally, one common LncRNA-SNP MIR4435-2HG detected as differentially expressed across the three cancer systems was selected. The lncRNAs were checked in literature for their involvement in cancer and further proceeded towards wet lab validation.

## 2.5. Culture of cell lines

For the validation of lncRNA and SNP, a selection of Breast Cancer cell lines (MCF7, T47D, and ZR751), Ovarian Cancer cell lines (SKOV3, OVCAR3, and PA1), Cervical Cancer cell lines (Hela and ME180), and the non-cancer Control Cell line HEK293 were chosen. T47D, ZR751, PA1, Hela, and HEK293 were obtained from the National Centre for Cell Science (NCCS), Pune. MCF7, OVCAR3, and SKOV3 were purchased from Cell Line Service (CLS) (<https://www.clinisciences.com/>), while ME180 was acquired from the American Type Culture Collection (ATCC). OVCAR3, ZR751 and T47D were cultured in RPMI-1640 medium, MCF7 was cultured in DMEM medium, SKOV3 was cultured in DMEM: Ham's F12 medium, Hela and PA1 were cultured in Eagle's minimal essential medium and ME180 was cultured in McCoy's 5A medium. All media included 10% fetal bovine serum (FBS, Invitrogen) and 1% penicillin/streptomycin (Invitrogen). The cells were maintained in a 5% CO<sub>2</sub> incubator at 37 °C until reaching confluency. **Figure 1** illustrates the cell condition at confluency.



**Figure1:** Culture and maintenance of Breast Cancer cell lines MCF7, ZR751 and T47D; Ovarian Cancer Cell lines SKOV3, OVCAR3 and PA1; Cervical Cancer Cell lines Hela and ME180; Non cancer cell line HEK293

## 2.6. Genotyping by the TaqMan PCR assay

Human adult normal Breast and Cervix tissue gDNA (both from single donor) were procured from Biochain (Cat# D1234086 and D1234275), while Human adult Ovarian genomic DNA extracted from Human Ovarian Surface Epithelial cells (HOSEpiC) Cat# 7319 was obtained from Sciencell. Genomic DNA from the cell lines was isolated using the HiPura mammalian genomic DNA preparation Kit (Himedia).

## Chapter4

Specifically, 600 bps of the MIR4435-2HG loci spanning the SNP was designed with the TaqMan custom genotyping design wizard (Assay ID: ANNMAMD). TaqMan PCR and genotyping analyses were conducted on the Applied Biosystems 7500 Real-Time PCR System following the manufacturer's instructions. The reaction mixtures were amplified with 0.5 µl of 20X primer-probe mix, 5 µl of 2X TaqMan Universal Master Mix (Applied Biosystems), 1 µl of gDNA (10 ng/µl), and 3.5 µl of ddH<sub>2</sub>O, the total volume being 10 µl. PCR cycling conditions are as follows: one cycle at 95 °C for 10 min; 40 cycles at 95 °C for 15 s and 58 °C for 1 min. The results were analyzed on the Applied Biosystems 7500 Real Time PCR System using the allelic discrimination assay program.

### 2.7. Real-time PCR analysis:

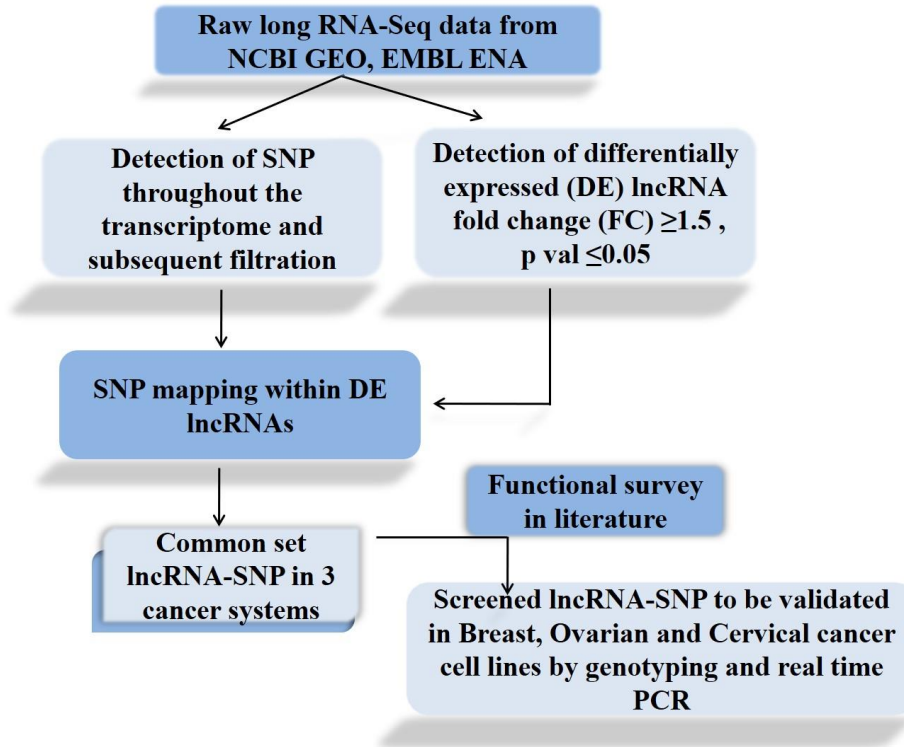
Normal Breast, Ovary, and Cervical tissue RNA were purchased from Biochain (Cat No# R1234086-50, C605167, and R1234275-50-A509136, respectively). Total RNA from cell lines was extracted using the RNeasy® Mini kit (Qiagen), and complementary DNA (cDNA) synthesis was carried out using the Verso cDNA Synthesis Kit (Thermo Fisher Scientific) following the manufacturer's instructions.

Subsequently, qPCR was performed using the SYBR Green PCR Master Mix (Applied Biosystems) on the Applied Biosystems 7500 Real-Time PCR System. Each experiment was performed in triplicates. The transcript expression levels were calculated using the 2- $\Delta\Delta C_t$  method, normalized to the expression of the endogenous control (18s). Primer sequences were designed using NCBI primer blast and Sigma Oligo Evaluator, and they were ordered from Sigma-Aldrich. The primer sequences used in the study, purchased from Merck, are summarized in **Table 2**. The entire workflow of the study is illustrated in **Figure2**.

Primer	Sequence
MIR4435-2HG_Foward	GACATTCCAGACAAGCGGTG
MIR4435-2HG_Reverse	CCCAGTTATTCAGGGAGAGGC
18s_Foward	GCGGCGTTATTCCCATGAC
18s_Reverse	GCTATCAATCTGTCAATCCTGTCC

**Table2:** Primer sequences of the lncRNA MIR4435-2HG and endogenous control 18s used for this study





**Figure2:** Workflow of the study of shared lncRNA variants in female cancers

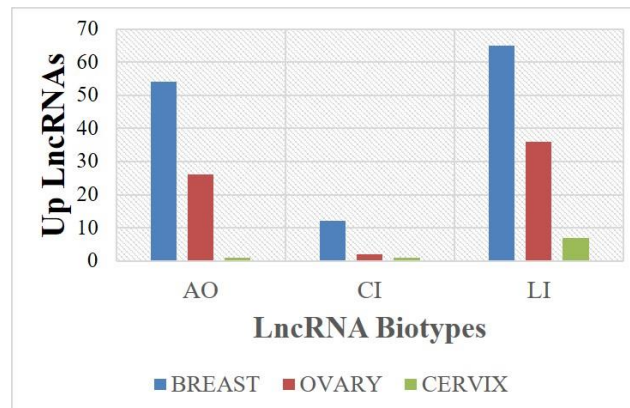
### 3. Results and Discussion:

#### 3.1. Analysis of Breast, Cervical and Ovarian Cancer datasets revealed only one common DE lncRNA-SNP across the three systems:

We conducted a comprehensive analysis of RNA sequencing data from breast, cervical, and ovarian cancers, focusing on transcriptome-wide variant mapping within lncRNA loci in these female cancer systems. Subsequently, we matched these variants with existing entries in the dbSNP database. In parallel, we performed differential expression (DE) analysis for each cancer type, comparing cancer samples with their respective normal counterparts. We then mapped the annotated variants within their loci, specifically examining the upregulated set of SNP harboring lncRNAs and their biotype distribution. **Table 3** presents the number of differentially expressed lncRNA-SNPs for each cancer, while **Figure 3** illustrates their biotype distribution.

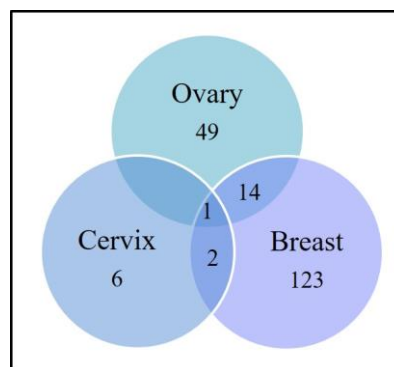
Cancer Tissues	UP lncRNA-SNPs
Breast	140
Cervical	9
Ovarian	64

**Table 3:** Upregulated lncRNAs harbouring SNP in the three cancer systems



**Figure 3:** The biotype distribution of lncRNAs associated with SNPs across various biotypes

Our hypothesis revolves around the shared gene regulatory circuits governed by SNP-associated lncRNAs in the breast, ovarian, and cervical cancers. In our investigation for a common set of differentially expressed SNP-containing lncRNAs across these three cancer types, we identified only a single significant lncRNA-SNP. It's important to note that the limited availability of cervical cancer datasets compared to that for breast and ovarian cancer is a constraint in this study. This limitation underscores the need for additional cervical cancer datasets in the future, which could enhance our ability to identify a more comprehensive set of commonly shared lncRNA-SNPs. **Figure 4** visually depicts the Venn diagram illustrating the common differentially expressed lncRNA-SNPs across the three cancer systems.



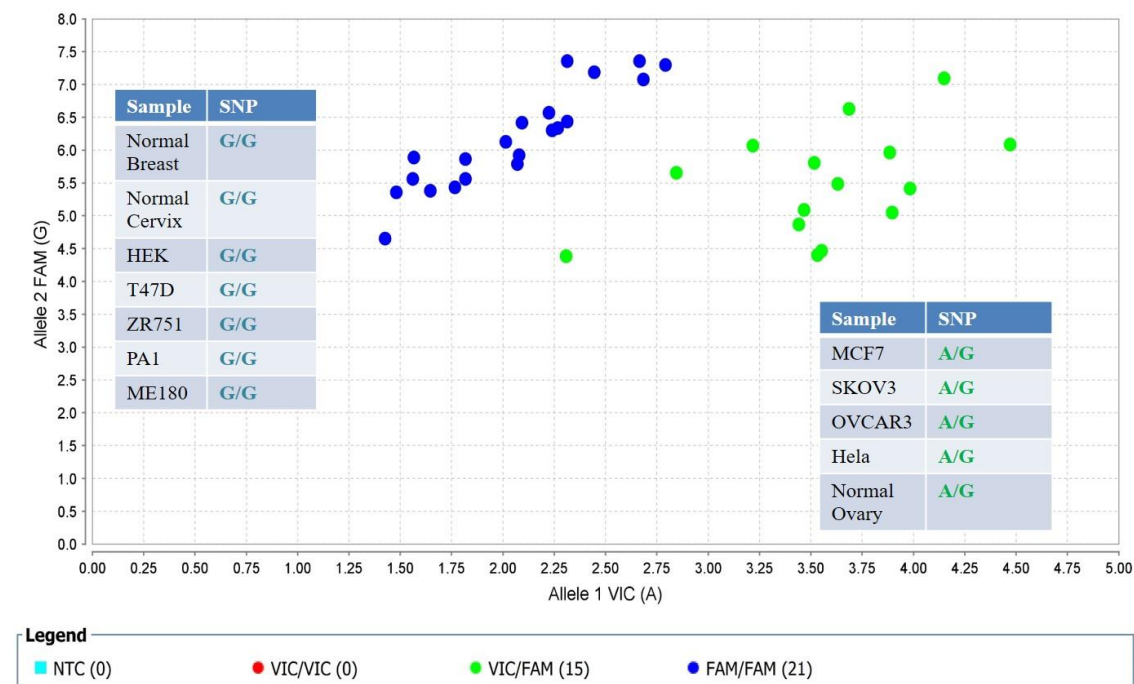
**Figure 4: A.** Common DE lncRNA harboring SNPs among the three systems. Only one significant lncRNA-SNP could be found commonly shared among Breast, Cervical and Ovarian Cancer

### 3.2. **MIR4435-2HG has reports of regulation in Breast, Cervical and Ovarian cancer but not the SNP:**

Our analysis identified MIR4435-2HG as the shared lncRNA hosting SNP. A literature review revealed the participation of MIR4435-2HG in a range of both cancerous and non-cancerous disorders[40]. Also known as LINC00978, this lncRNA located in the 2q13 region of chr 2, comprises of ten exons. It generates 108 transcripts through alternative splicing. Among them transcript ENST00000409569 was found to be enriched in our analysis. Functioning as an onco-lncRNA, MIR4435-2HG exerts its effects through various mechanisms. These include hindering apoptosis[41], acting as a microRNA (miRNA) sponge[42], fostering cell proliferation[43], promoting cell invasion and migration[44, 45], and enhancing epithelial-to-mesenchymal transition (EMT)[46]. MIR4435-2HG also has the capability to modulate multiple signalling pathways[47-49]. Consequently, its involvement can contribute to the progression of tumors. In recent studies on breast, cervical, and ovarian cancers, MIR4435-2HG has been implicated as a competing endogenous lncRNA[50-52]. Despite the available reports, no specific work reporting the detailed role of any functional variant within the MIR4435-2HG loci has been found. rs1045267 was identified as a commonly shared variant. However, the in-silico detection method used to identify genomic variants did not provide information about whether the variant exists in a homozygous or heterozygous state. To address this gap, we initiated a validation process for the SNP rs1045267 (A/G) located within MIR4434-2HG in the three cancer systems. This validation aims to confirm the presence of the SNP and determine its allelic status (homozygous or heterozygous) within the context of MIR4434-2HG in these cancer types.

### 3.3. **TaqMan genotyping assay validated the presence of heterozygous and homozygous recessive allele of the variant in Cancer and Control systems**

The outcomes derived from the TaqMan genotyping assay, as illustrated in **Figure 4**, indicate the presence of both homozygous recessive (GG) and heterozygous variants (AG) of rs1045267. Specifically, the GG allele was detected in normal breast tissue, normal cervix tissue, the non-cancer control cell line HEK, breast cancer cell lines ZR751 and T47D, ovarian cancer cell line PA1, and cervical cancer cell line ME180. In contrast, the heterozygous genotype was discerned in the breast cancer cell line MCF7, ovarian cancer cell lines SKOV and OVCAR3, cervical cancer cell line HeLa, as well as genomic DNA derived from normal human ovarian surface epithelial cells.



**Figure 4:** TaqMan genotyping assay revealed the presence of heretozygous AG allele (green) and homozygous recessive GG allele (blue) in Control tissues and cancer cell lines.

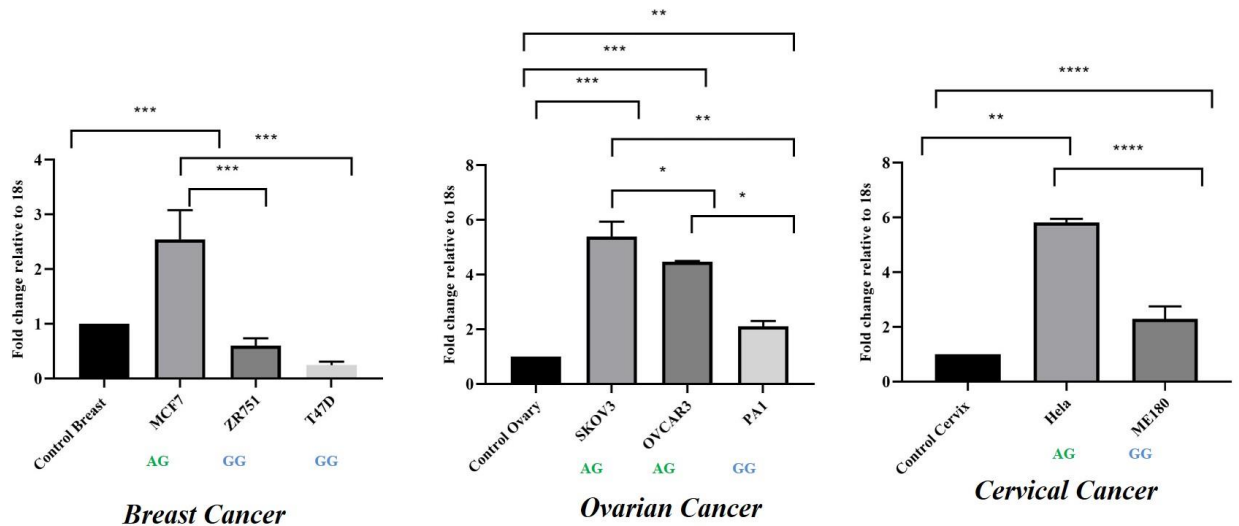
We did not detect the presence of the major allele in any of the samples. Upon closer examination of the variant's minor allele frequency, we observed a predominance of the minor allele over the major allele, particularly in the European population. The cancer cell lines used for validation, with the exception of Hela (African American) and HEK293 (of unknown parenthood), are of Caucasian origin. This could account for the prevalent minor allele observed in the majority of the samples. Additionally, the breast and cervical genomic DNA purchased from Biochain originates from Caucasian and Asian populations, respectively. However, we were unable to find information on the allele frequency of the SNP in the Asian population.

**3.4. Real-Time PCR analysis revealed the upregulated expression of lncRNA MIR4435-2HG in the cancer cell line carrying the heterozygous allele of rs1045267**

It has been previously documented that functional genetic variants within lncRNA loci can influence the expression of the corresponding lncRNA. For instance, individuals carrying the TT genotype of HOTAIR rs920778 exhibit an elevated risk of Gastric Cancer (GC) in Jinan and Huaian populations compared to CC carriers, with higher HOTAIR expression observed among T allele carriers. In the Chinese Han population, individuals with the AG genotype of rs619586 experience a reduced risk of Breast Cancer (BC). Furthermore, the expression of MALAT1 in individuals with AG and GG genotypes of rs619586 was notably lower than those with the AA genotype. Another study conducted

on a Chinese population revealed a significant association between the minor alleles of rs4759314 and susceptibility to Pancreatic Cancer (PC). Individuals carrying the minor alleles of rs4759314 showed notably elevated HOTAIR RNA levels in prostate cancer tissues compared to those carrying the major alleles. Therefore, it is imperative to verify whether the expression of the MIR4435-2HG lncRNA is influenced by the presence of the heterozygous or homozygous form of the SNP in various cell lines of breast, ovarian, and cervical cancer. We confirmed the differential expression of the lncRNA MIR4435-2HG in the three cancer systems by comparing them to their respective normal counterparts, as depicted in **Figure 5**. Using 18s as the endogenous control, q-PCR analysis of lncRNA MIR4435-2HG in cancer cell lines and their normal counterparts demonstrated a statistically significant upregulated expression of the heterozygous genotype (AG) harboring lncRNA MIR4435-2HG compared to those harboring the minor allele (GG).

In breast cancer, the expression of MIR4435-2HG is higher in MCF7 (allele AG) compared to ZR751 (allele GG) and T47D (GG). Similarly, for ovarian cancer, its expression in SKOV3 (AG) and OVCAR3 (AG) is higher than in PA1 (GG). In cervical cancer, the expression in HeLa (AG) is higher than in ME180 (GG). Therefore, it is observed that the presence of the heterozygous allele of SNP rs10425267 in MIR4435-2HG, unlike in the normal counterpart, could influence the upregulated expression of this lncRNA in different female cancer systems.



**Figure 5:** qPCR analysis revealed upregulated expression of ENST00000409569 (MIR4435-2HG) in different cell lines of Breast, Ovarian and Cervical cancer harbouring allele AG compared to those cell lines harbouring allele GG

To fortify these findings and extend their applicability, it is imperative to conduct further validation across an expanded array of breast, ovarian, and cervical cancer cell lines. By encompassing a more comprehensive set of cancer types and incorporating diverse

## Chapter4

cellular contexts, this extended validation will not only bolster the association between the identified SNP and MIR4435-2HG expression but also refine our understanding of its regulatory role across different female cancer systems.

### 4. Conclusion:

In this study, our primary objective was to investigate lncRNA loci containing SNPs associated with a predisposition to the development of breast, cervical, and ovarian cancers. Leveraging publicly available cancer datasets for these three distinct systems, our goal was to identify lncRNA-SNPs and differentially regulated lncRNAs shared across these cancer types. Our comprehensive analysis unveiled a previously unreported SNP within the lncRNA MIR4435-2HG, shedding light on its potential role in cancer development. This finding represents a novel contribution to the understanding of the genetic factors influencing cancer susceptibility. To validate our computational findings, we conducted wet bench experiments, confirming the presence of SNP within the 3 cancer systems and provide clue regarding the impact of the identified SNP on the expression level of MIR4435-2HG. This wet bench validation not only substantiates our computational predictions but also provides valuable new insights into the intricate regulatory mechanisms governing lncRNA expression through genetic variants. In essence, our study highlights the importance of previously unrecognized genetic variations within lncRNAs and its influence on its expression level. These findings enhance our understanding of the complex interplay between genetic factors and lncRNA regulation in these three cancer systems.

### 5. References:

1. Mercer, T.R., M.E. Dinger, and J.S. Mattick, *Long non-coding RNAs: insights into functions*. Nat Rev Genet, 2009. **10**(3): p. 155-9.
2. Rashid, F., A. Shah, and G. Shan, *Long Non-coding RNAs in the Cytoplasm*. Genomics, Proteomics & Bioinformatics, 2016. **14**(2): p. 73-80.
3. Wilusz, J.E., H. Sunwoo, and D.L. Spector, *Long noncoding RNAs: functional surprises from the RNA world*. Genes Dev, 2009. **23**(13): p. 1494-504.
4. Rinn, J.L. and H.Y. Chang, *Genome regulation by long noncoding RNAs*. Annu Rev Biochem, 2012. **81**: p. 145-66.
5. Reich, D.E., S.B. Gabriel, and D. Altshuler, *Quality and completeness of SNP databases*. Nat Genet, 2003. **33**(4): p. 457-8.
6. Erichsen, H.C. and S.J. Chanock, *SNPs in cancer research and treatment*. Br J Cancer, 2004. **90**(4): p. 747-51.
7. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits*. Nat Rev Genet, 2009. **10**(4): p. 241-51.
8. Kumar, V., et al., *Human disease-associated genetic variation impacts large intergenic non-coding RNA expression*. PLoS Genet, 2013. **9**(1): p. e1003201.
9. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits*. Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.
10. Wang, C., et al., *lncRNA Structural Characteristics in Epigenetic Regulation*. Int J Mol Sci, 2017. **18**(12).
11. Graf, J. and M. Kretz, *From structure to function: Route to understanding lncRNA mechanism*. BioEssays, 2020. **42**(12): p. 2000027.

12. Gong, J., et al., *lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse*. Nucleic Acids Res, 2015. **43**(Database issue): p. D181-6.
13. Yoneda, A., et al., *Breast and ovarian cancers: a survey and possible roles for the cell surface heparan sulfate proteoglycans*. J Histochem Cytochem, 2012. **60**(1): p. 9-21.
14. Tao, Z., et al., *Breast Cancer: Epidemiology and Etiology*. Cell Biochem Biophys, 2015. **72**(2): p. 333-8.
15. Bray, F., et al., *Global cancer transitions according to the Human Development Index (2008-2030): a population-based study*. Lancet Oncol, 2012. **13**(8): p. 790-801.
16. Jemal, A., et al., *Cancer statistics, 2009*. CA Cancer J Clin, 2009. **59**(4): p. 225-49.
17. Liu, Z., et al., *Over-expressed long noncoding RNA HOXA11-AS promotes cell cycle progression and metastasis in gastric cancer*. Mol Cancer, 2017. **16**(1): p. 82.
18. Hassanzarei, S., et al., *Genetic polymorphisms of HOTAIR gene are associated with the risk of breast cancer in a sample of southeast Iranian population*. Tumour Biol, 2017. **39**(10): p. 1010428317727539.
19. Yin, J., et al., *Expression Quantitative Trait Loci for CARD8 Contributes to Risk of Two Infection-Related Cancers--Hepatocellular Carcinoma and Cervical Cancer*. PLoS One, 2015. **10**(7): p. e0132352.
20. Guo, L., et al., *Expression quantitative trait loci in long non-coding RNA ZNRD1-AS1 influence cervical cancer development*. Am J Cancer Res, 2015. **5**(7): p. 2301-7.
21. Łażniak, S., et al., *The association of CCAT2 rs6983267 SNP with MYC expression and progression of uterine cervical cancer in the Polish population*. Arch Gynecol Obstet, 2018. **297**(5): p. 1285-1292.
22. Koboldt, D.C., et al., *Comprehensive molecular portraits of human breast tumours*. Nature, 2012. **490**(7418): p. 61-70.
23. Hennessy, B.T., R.L. Coleman, and M. Markman, *Ovarian cancer*. The Lancet, 2009. **374**(9698): p. 1371-1382.
24. Mehrgou, A. and M. Akouchekian, *The importance of BRCA1 and BRCA2 genes mutations in breast cancer development*. Med J Islam Repub Iran, 2016. **30**: p. 369.
25. Malhotra, H., et al., *Genetic Counseling, Testing, and Management of HBOC in India: An Expert Consensus Document from Indian Society of Medical and Pediatric Oncology*. JCO Glob Oncol, 2020. **6**: p. 991-1008.
26. Bahcall, O., *Shared susceptibility loci for breast, prostate and ovarian cancers*. Nature Genetics, 2013.
27. Ramachandran, D. and T. Dörk, *Genomic Risk Factors for Cervical Cancer*. Cancers (Basel), 2021. **13**(20).
28. Wang, S., et al., *Association of 42 SNPs with genetic risk for cervical cancer: an extensive meta-analysis*. BMC Medical Genetics, 2015. **16**(1): p. 25.
29. Madeleine, M.M., et al., *Comprehensive analysis of HLA-A, HLA-B, HLA-C, HLA-DRB1, and HLA-DQB1 loci and squamous cell cervical cancer risk*. Cancer Res, 2008. **68**(9): p. 3532-9.
30. Hosono, S., et al., *HLA-A alleles and the risk of cervical squamous cell carcinoma in Japanese women*. J Epidemiol, 2010. **20**(4): p. 295-301.
31. Barrett, T. and R. Edgar, *Gene expression omnibus: microarray data storage, submission, retrieval, and analysis*. Methods Enzymol, 2006. **411**: p. 352-69.
32. Kanz, C., et al., *The EMBL Nucleotide Sequence Database*. Nucleic Acids Res, 2005. **33**(Database issue): p. D29-33.
33. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation*. Nucleic Acids Res, 2001. **29**(1): p. 308-11.
34. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads*. 2011, 2011. **17**(1): p. 3.
35. Pertea, M., et al., *Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown*. Nature Protocols, 2016. **11**(9): p. 1650-1667.

## Chapter4

36. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
37. Oikkonen, L. and S. Lise, *Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection*. Wellcome Open Res, 2017. **2**: p. 6.
38. Rimmer, A., et al., *Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications*. Nat Genet, 2014. **46**(8): p. 912-918.
39. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features*. Bioinformatics, 2010. **26**(6): p. 841-2.
40. Ghasemian, M., et al., *Long non-coding RNA MIR4435-2HG: a key molecule in progression of cancer and non-cancerous disorders*. Cancer Cell International, 2022. **22**(1): p. 215.
41. Luo, P., et al., *LncRNA MIR4435-2HG mediates cisplatin resistance in HCT116 cells by regulating Nrf2 and HO-1*. PLoS One, 2020. **15**(11): p. e0223035.
42. Li, X., Y. Ren, and T. Zuo, *Long noncoding RNA LINC00978 promotes cell proliferation and invasion in non-small cell lung cancer by inhibiting miR-6754-5p*. Mol Med Rep, 2018. **18**(5): p. 4725-4732.
43. Kong, Q., et al., *The lncRNA MIR4435-2HG is upregulated in hepatocellular carcinoma and promotes cancer cell proliferation by upregulating miRNA-487a*. Cellular & Molecular Biology Letters, 2019. **24**(1): p. 26.
44. Wu, D., et al., *LncRNA AWPPH accelerates the progression of non-small cell lung cancer by sponging miRNA-204 to upregulate CDK6*. Eur Rev Med Pharmacol Sci, 2020. **24**(8): p. 4281-4287.
45. Zhang, H., et al., *LncRNA MIR4435-2HG promotes cancer cell migration and invasion in prostate carcinoma by upregulating TGF- $\beta$ 1*. Oncol Lett, 2019. **18**(4): p. 4016-4021.
46. Gao, L.F., et al., *Inhibition of MIR4435-2HG on Invasion, Migration, and EMT of Gastric Carcinoma Cells by Mediating MiR-138-5p/Sox4 Axis*. Front Oncol, 2021. **11**: p. 661288.
47. Zhang, Q., et al., *LINC00978 promotes hepatocellular carcinoma carcinogenesis partly via activating the MAPK/ERK pathway*. Bioscience Reports, 2020. **40**(3): p. BSR20192790.
48. Ghasemian, M., et al., *Long noncoding RNA LINC00978 acts as a potential diagnostic biomarker in patients with colorectal cancer*. Experimental and Molecular Pathology, 2021. **122**: p. 104666.
49. Sabbadini, F., et al., *The Multifaceted Role of TGF- $\beta$  in Gastrointestinal Tumors*. Cancers (Basel), 2021. **13**(16).
50. Liu, A.N., et al., *LncRNA AWPPH and miRNA-21 regulates cancer cell proliferation and chemosensitivity in triple-negative breast cancer by interacting with each other*. J Cell Biochem, 2019. **120**(9): p. 14860-14866.
51. Zhu, L., et al., *LncRNA MIR4435-2HG triggers ovarian cancer progression by regulating miR-128-3p/CKD14 axis*. Cancer Cell International, 2020. **20**(1): p. 145.
52. Wang, R., et al., *Knockdown of MIR4435-2HG Suppresses the Proliferation, Migration and Invasion of Cervical Cancer Cells via Regulating the miR-128-3p/MSI2 Axis in vitro*. Cancer Manag Res, 2020. **12**: p. 8745-8756.



## **CHAPTER 5| Developing LncRNA-SNP based Breast and Ovarian cancer risk prediction models and its effect on gene regulation**

### **Abstract:**

Long non-coding RNAs (lncRNAs) and single nucleotide polymorphisms (SNPs) within them play crucial roles in cancer susceptibility and disease outcomes. Breast and Ovarian cancers, characterized by genetic heterogeneity, present significant challenges for precise diagnosis and treatment. Despite recent advancements in personalized medicine, including lncRNA-SNP markers into cancer risk detection panels remains limited. In this work, we put forward lncRNA-SNP regulated gene expression-based Breast and Ovarian Cancer risk probability models for predicting the risk factor in case of patients with abnormal breast or ovary conditions that increases the risk of getting these cancers. Notably, our approach accounts for the tissue-specificity of lncRNAs as well the benefit for individuals with predisposing conditions. Additionally, pathway analysis revealed the involvement of the regulatory genes targeted by the lncRNA harbouring the SNP in key cancer-regulating pathways. TaqMan genotyping and qPCR are being performed to confirm the presence of selected lncRNA-SNPs in ovarian and breast cancer cell lines along with the upregulated expression of the lncRNA transcripts. These findings highlight previously overlooked genetic variants within lncRNA loci and their regulatory impact on disease outcomes, providing insights into personalized cancer diagnosis and treatment strategies.

### **1. Background and Objective of the study**

Long non-coding RNA (lncRNA) molecules, surpassing 200 nucleotides in length, are predominantly transcribed by RNA polymerase II, undergoing splicing and polyadenylation processes. [1]. Primarily situated within the nucleus, these molecules exhibit multifaceted roles in cellular mechanisms, notably implicated in cancer progression [2, 3]. Single Nucleotide Polymorphisms (SNPs), often found within non-coding genomic regions, exert significant influence on disease susceptibility, particularly in cancer contexts. [4, 5].

SNPs may influence the expression of disease-associated lncRNAs[6], impact splice sites leading to the generation of alternative splice variants with modified functionality. Additionally, they can bring alterations in secondary structures such as hairpin loops, which have the potential to modify RNA folding and influence interactions with target proteins or transcripts[7].

It is essential to note that certain clinical conditions in breast and ovary increase the risk of breast and ovarian cancer respectively. Atypical Hyperplasia and Multiple Papillomas, can increase the risk of Breast cancer if left untreated or poorly managed[8-11]. Similarly,

## Chapter 5

patients with Endometriosis (where tissue similar to the lining of the uterus grows outside the uterus, often found on the ovaries) or Polycystic Ovarian Syndrome (PCOS) are at increased risk of developing Ovarian cancer[12-15], influenced by various factors including age, reproductive history, and lifestyle. Discrete works have also identified shared genetic mutation, copy number alteration, protein expression and active signalling pathways[16, 17] which underscore the complex interplay between benign conditions and the development of Cancers. In such scenarios, tissue biopsy is done to check the presence of any malignancy within it. This work focuses towards looking into the presence of tissue specific lncRNA-SNPs along with interacting gene expression within these biopsy samples which provides a much precise and well resolved view regarding cancer risk in such patients. This will further help for early cancer diagnosis in such disease susceptible cases.

Here it is important to mention that the focus mainly has been on coding gene SNPs. Known risk factors for both breast and ovarian cancer include gene mutations, notably in BRCA1 and BRCA2[18, 19]. The inherent heterogeneity of these cancers has been repeatedly emphasized, and metabolic diversity, brought about by genetic alterations, along with their impact on interacting partners is evident among tumors originating from the same tissue[20]. This heterogeneity plays a crucial role in determining therapeutic susceptibilities and has the potential to predict clinical outcomes. Recent advancements include not only germline DNA screening for BRCA1/2 mutations but also tumor-level assessment, often through Next-Generation Sequencing (NGS) for personalized treatment[21]. Accurate characterization of tumors and immune microenvironments through transcriptome sequencing has become essential for effective personalized cancer treatment. With the advent of tissue specific lncRNAs, lncRNA-SNP markers are getting increasing attention nowadays with the individual efforts coming up to establish connections between the impact of SNP in lncRNA loci with Breast and Ovarian carcinoma[22-25]. But there lies a notable gap in incorporating lncRNAs-SNP markers into cancer risk detection panels mainly in case of such patients having aberrant breast and ovarian clinical conditions. During our prior research, as outlined in Chapter 4, we observed a significant overlap of non-coding SNPs between Breast and Ovarian cancer. Thus, our current aim is to specifically select lncRNA-SNP markers that exclusively influence breast and ovarian cancer.

In this study, the work has been done in two phases: First, utilizing publicly available gene expression datasets of cancer patients, we first aimed to identify such exclusive lncRNA-SNPs in Breast and Ovarian Cancer systems. We next developed individual Breast and Ovarian cancer risk prediction models using machine learning (ML) techniques. We incorporated the expression status of genes in both types of cancer datasets and their respective normal or non-cancerous counterparts, regulated by the presence or absence of these variants. With the final set of 14 lncRNA-SNPs (4 specific to Breast Cancer and 10 specific to Ovarian cancer), each model demonstrated a high precision in predicting the probability of cancer risk or non-cancerous status. This has been followed by wet lab validation of most important SNPs along with the expression validation of the corresponding lncRNAs. Our findings highlight the significance of

previously overlooked genetic variants within lncRNA loci and their regulatory influence on interacting partners, ultimately dictating cancer risk outcomes.

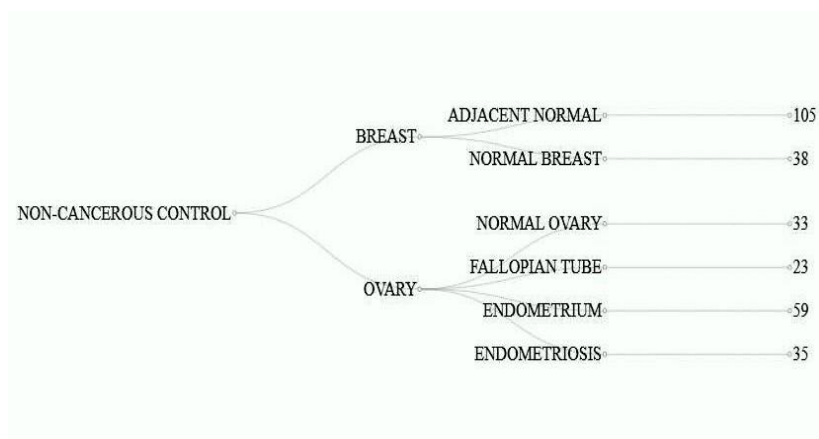
## 2. Materials and Methods

### 2.1. Raw Data corresponding to Breast and Ovarian cancer systems and their normal counterparts:

The raw RNA sequence data for patients with Ovarian epithelial and Breast cancers (including Luminal, TNBC, and Her2 subtypes), along with their respective healthy or adjacent normal counterparts, were sourced from NCBI GEO [26] and ArrayExpress[27]. Additionally, endometriosis samples were collected to serve as non-cancerous controls for training the Ovarian cancer models. Annotated SNP data for the Human genome (GRCh37) were retrieved from NCBI dbSNP (version 151).[28]. After an extensive data search, a total of 367 Breast carcinoma samples, 38 Normal Control and 105 adjacent Normal controls were successfully acquired. For Ovarian cancer, a total of 308 samples for Ovarian High Grade Serous, along with 56 non-cancerous controls of Ovarian and Fallopian tube tissue of epithelial origin, were obtained. For model training of Ovarian cancer specific SNPs, further 59 samples of healthy endometrial cells (HEMT) and 35 samples of Endometriosis (EMT) were also considered. The datasets were divided into Training, Validation (comprising randomly selected samples from the same GEO and ArrayExpress datasets used for Training), and Testing (involving samples from entirely separate datasets not utilized during training to prevent any bias). All the patient datasets categorized under Training and Validation were utilized for selection of distinct SNPs between Breast and Ovarian cancer. **Table 1** provides a summary of the details pertaining to the input data. **Figure 1** shows the distribution of non-cancerous tissue control samples.

Cancer	Patient tissue samples			Non cancerous tissue samples		
	Training	Validation	Testing	Training	Validation	Testing
BREAST	277	68	22	113	20	10
OVARY	249	50	22	119	21	10

**Table1:** A summary of the input data utilized for the entire work.



**Figure1:** Tissue samples considered as non-cancerous control counterparts for Breast and Ovarian Cancer respectively. For Ovarian Cancer, Healthy endometrial cells and Endometriosis were considered for model training.

## 2.2. Pre-processing and SNP detection throughout transcriptomic data

FastQC v.0.11.7 was employed for assessment of read quality for each sample, retaining only those with a quality score exceeding 30. Adapter sequences were then trimmed using Cutadapt v1.16 [29]. Subsequently, the paired-end raw sequence reads were aligned to the Human reference genome (hg38) through HISAT2 2.1.0 [30]. The resulting BAM files underwent organization and indexing with SAMtools 0.1.19[31]. Before initiating variant calling, the BAM files underwent preprocessing using Opossum 0.2, [32] a tool tailored for quality control procedures. This preprocessing step included the removal of duplicate or poorly mapped reads, and secondary alignments, as well as the merging of overlapping reads. Opossum seamlessly integrates with the Variant Caller Platypus[33], which was then employed for SNP and indel detection with default parameters. Platypus is renowned for its speed and sensitivity, placing it on par with other leading variant callers in the field. Variants with a QUALITY status of "PASSED" and more than 5 reads retaining the variants (TR) were selected from the output. Following the variant filtering process, annotation of the variants was executed using dbSNP (version 151) through the "bcftools(v1.9) annotate" tool [31].

## 2.3. Selection of exclusive breast and ovarian lncRNA-SNPs

A SNP matrix file was generated incorporating all the cancer patient datasets categorized under Training and Validation to identify exclusive SNPs in each of these two cancers. The chi-square test was employed with a significance threshold value of  $p \leq 5e-12$ . The selected SNPs underwent further filtering based on the following criteria:

- Mapping SNPs within lncRNAs loci not overlapping with any protein coding region using bedtools(v2.26.0) Intersect [34].
- Determining whether the SNP is capable of disrupting the lncRNA secondary structure. This assessment was conducted using RNAsnp[35], which predicts SNP effects on local RNA secondary structure, with a p-

value threshold  $< 0.2$  indicating significant structural change. c. Filtering out the common variants and retaining only low-frequency variants, which are more prone to be linked with disease risk [36].

#### **2.4. Detection of lncRNA-SNP expression as well as differentially expressed transcripts in the two cancer systems against their normal counterparts**

Using the BAM files, StringTie v1.3.4d [30] has been used to assemble transcripts. For lncRNAs, transcript information have been retrieved from lncRBaseV.2[37] which offers a comprehensive non-redundant list of lncRNAs from various sources. The lncRNAs within which the selected SNPs reside are checked for their expression and filtered out. Annotations for protein-coding genes were retrieved from Gencode[38].

Differential analysis (DE) of patient samples was performed using Ballgown[30]. In the context of Breast cancer, both healthy Breast biopsy tissue and non-involved tissue adjacent to the tumor were utilized as control samples. Conversely, for Ovarian cancer, healthy epithelial tissue originating from the Ovaries and fallopian tubes was considered as the control group. The stat test function within Ballgown addressed various considerations, including potential batch effects arising from multiple data sources. Differentially regulated mRNAs were identified, applying a fold change (FC) cut-off of  $\geq 2$ , along with a significance threshold of  $p\text{-value} \leq 0.05$ . The DE status of the selected lncRNAs is also checked with a FC cut-off of  $\geq 1.5$  and  $p\text{-value} \leq 0.05$ .

#### **2.5. Screening interacting gene partners of lncRNA-SNPs through correlation analysis, RNA-RNA interaction information and ce(competitive endogenous)-RNA network analysis**

Cis and trans correlation analyses were conducted using cancer datasets to identify robust lncRNA target genes. For cis-pairs, a DE coding gene located with 20kb loci of an lncRNA was considered. Trans co-expression analysis between DE mRNA and lncRNA was conducted using Pearson's correlation coefficient, employing a screening criterion of  $R^2$  (Spearman coefficient)  $\geq 0.7$  and  $p\text{-value} \leq 0.05$ . This analysis was performed using the rcorr function from the Hmisc library in R, which can be accessed at <https://cran.r-project.org/web/packages/Hmisc>.

Additionally, in order to explore the tissue-specific functions of lncRNAs, "guilt by association" method has been adopted[39, 40]. We accessed datasets containing 1569 Breast cancer samples and 507 Ovarian cancer samples from the TCGA database through the GDC portal (<https://portal.gdc.cancer.gov/>). Employing the Spearman method in R with a threshold of Spearman coefficient  $> 0.5$  and  $p \leq 0.01$ , we conducted correlation analyses between the lncRNAs and all 19,962 coding genes within the respective cancer samples. Following this, we scrutinized the differential expression status of the correlated

## Chapter5

genes in our analysis, including them only if they showed significant differential expression in the corresponding cancers with respect to their control group.

StarBaseV2.0[41] has been considered to collect lncRNA-mRNA interacting pairs identified from high-throughput sequencing data of RNA-RNA interactome, such as LIGR-Seq[42], PARIS[43], SPLASH[44].

Following the ceRNA theory, which posits that lncRNAs can act as endogenous "sponges" to regulate mRNA expression by sequestering miRNAs, we constructed a lncRNA-miRNA-mRNA network. Human miRNA sequences were obtained from miRBase[45]. Point mutations were introduced at the selected lncRNA loci corresponding to the mapped SNPs to investigate the creation of new miRNA binding sites (8mers, 7mer-m8, and 7mer-A1) using TargetScan Release 8.[46] The roles of these miRNAs were explored in the literature, and DE mRNAs with potential target sites within the 3' UTR region for these miRNAs were identified.

By consolidating the outputs from the four aforementioned approaches, a distinct set of interacting partner genes have been identified for each screened SNP associated lncRNA(lncRNA-SNP) . Their gene expression pattern have been used as the backbone to serve as the feature set for developing the prediction model for cancer risk prediction.

### 2.6. Machine Learning based analysis and developing cancer risk prediction model

By considering the expression value of the genes associated with specific lncRNA-SNPs as features (as shown to be screened from the previous steps), we have incorporated the simplest generalized linear model in forms of logistic regression to segregate the normal samples from the cancerous one. We have developed separate models for breast and ovarian cancer risk prediction. The workflow has been executed in the following way:

a. **Initial Model Building:** Based on the expression levels of the specific set of genes interacting with each lncRNA harboring an SNP, we developed an initial logistic regression model using training data inclusive of all features. Utilizing the resulting machine learning model, we then extracted and separated the importance of selected features into positive and negative categories, which were subsequently forwarded for the feature selection stage.

b. **Feature Selection by Combinatorial Method:** Given the positive and negative important features, responsible for segregation of cancer samples from the normal ones, they are mixed to select the best set of features which yields maximum accuracy upon validation and unknown test data. For individual lncRNA-SNP, following strategy has been incorporated to extract the best feature sets: (i) Initially, positive and negative features are sorted according to their importance score (ii) Based on the combination of positive and negative features, individual feature sets were constructed. (iii) Finally, we have incorporated greedy search-based feature selection strategy to select subset of

features from original feature sets in an iterative manner such that the performance of individual machine learning models are optimized.

**c. Final Model Building:** Using the best set of features yielded from the previous stage we have built the final logistic regression model. Initially all the un-normalized features are normalized and standardized using the equation (1) where **mean** denotes **average** and **std** denotes **standard deviation**. Finally, the normalized features are trained using ML based logistic regression model.

$$\text{Normalized\_X} = \frac{X - \text{mean}(X)}{\text{std}(X)} \quad (1)$$

The above sets of operations have been executed individually for each lncRNA-SNP associated with Breast and Ovarian Cancer. We have utilized the python based Scikit-learn (<https://scikit-learn.org/stable/index.html>) tool to create ML model.

The entire code for the model is provided below:

### Preprocess Dataset

We utilized Pandas to generate training, validation, and unknown test datasets from raw SNP information. Within these datasets, the columns ALT\_BREAST and REF\_BREAST denote the alternate and reference SNP alleles for breast cancer samples, while ALT\_NB and REF\_NB represent the alternate and reference SNP alleles for normal breast samples.

```
import pandas as pd

train_df1=pd.read_csv('Input_Dataset/rs2366152_ALT_BREAST_SELECTED_ENST_TRAIN_details.txt',sep='\t')
train_df1.drop(['transcriptNames'],axis=1,inplace=True)
train_df1['SNP']=1
train_df1['Target']=1

train_df2=pd.read_csv('Input_Dataset/rs2366152_REF_BREAST_SELECTED_ENST_TRAIN_details.txt',sep='\t')
train_df2.drop(['transcriptNames'],axis=1,inplace=True)
train_df2['SNP']=0
train_df2['Target']=1

train_df3=pd.read_csv('Input_Dataset/rs2366152_ALT_NB_SELECTED_ENST_TRAIN_details.txt',sep='\t')
train_df3.drop(['transcriptNames'],axis=1,inplace=True)
train_df3['SNP']=1
train_df3['Target']=0

train_df4=pd.read_csv('Input_Dataset/rs2366152_REF_NB_SELECTED_ENST_TRAIN_details.txt',sep='\t')
train_df4.drop(['transcriptNames'],axis=1,inplace=True)
```

## Chapter5

```
train_df4['SNP']=0
train_df4['Target']=0

train_final=pd.concat([train_df1,train_df2,train_df3,train_df4])
train_final=train_final.sample(frac=1,random_state=369)
train_final.reset_index(drop=True,inplace=True)
print(train_final.shape,train_final.Target.value_counts())
train_final.head(3)

val_df1=pd.read_csv('Input_Dataset/rs2366152_ALT_BREAST_SELECTED
_ENST_TEST_details.txt',sep='\t')
val_df1.drop(['transcriptNames'],axis=1,inplace=True)
val_df1['SNP']=1
val_df1['Target']=1

val_df2=pd.read_csv('Input_Dataset/rs2366152_REF_BREAST_SELECTED
_ENST_TEST_details.txt',sep='\t')
val_df2.drop(['transcriptNames'],axis=1,inplace=True)
val_df2['SNP']=0
val_df2['Target']=1

val_df3=pd.read_csv('Input_Dataset/rs2366152_ALT_NB_SELECTED_ENS
T_TEST_details.txt',sep='\t')
val_df3.drop(['transcriptNames'],axis=1,inplace=True)
val_df3['SNP']=1
val_df3['Target']=0

val_df4=pd.read_csv('Input_Dataset/rs2366152_REF_NB_SELECTED_ENS
T_TEST_details.txt',sep='\t')
val_df4.drop(['transcriptNames'],axis=1,inplace=True)
val_df4['SNP']=0
val_df4['Target']=0

val_final=pd.concat([val_df1,val_df2,val_df3,val_df4])
val_final=val_final.sample(frac=1,random_state=369)
val_final.reset_index(drop=True,inplace=True)
print(val_final.shape,val_final.Target.value_counts())
val_final.head(3)

test_df1=pd.read_csv('Input_Dataset/rs2366152_ALT_BREAST_SELECTE
D_ENST_UNTRAINED_details.txt',sep='\t')
test_df1.drop(['transcriptNames'],axis=1,inplace=True)
test_df1['SNP']=1
test_df1['Target']=1

test_df2=pd.read_csv('Input_Dataset/rs2366152_REF_BREAST_SELECTE
D_ENST_UNTRAINED_details.txt',sep='\t')
```



```
test_df2.drop(['transcriptNames'],axis=1,inplace=True)
test_df2['SNP']=0
test_df2['Target']=1

test_df3=pd.read_csv('Input_Dataset/rs2366152_ALT_NB_SELECTED_EN
ST_UNTRAINED_details.txt',sep='\t')
test_df3.drop(['transcriptNames'],axis=1,inplace=True)
test_df3['SNP']=1
test_df3['Target']=0

test_df4=pd.read_csv('Input_Dataset/rs2366152_REF_NB_SELECTED_EN
ST_UNTRAINED_details.txt',sep='\t')
test_df4.drop(['transcriptNames'],axis=1,inplace=True)
test_df4['SNP']=0
test_df4['Target']=0

test_final=pd.concat([test_df1,test_df2,test_df3,test_df4])
test_final=test_final.sample(frac=1,random_state=369)
test_final.reset_index(drop=True,inplace=True)
print(test_final.shape,test_final.Target.value_counts())
test_final.head(3)

train_final.to_csv('Final_Dataset/Train_Datset.txt',index=False,
sep='\t')
val_final.to_csv('Final_Dataset/Val_Datset.txt',index=False,sep=
'\t')
test_final.to_csv('Final_Dataset/Test_Datset.txt',index=False,se
p='\t')
```

### **Initial Creation of Logistic Regression Model**

To establish the initial combination of feature sets, we employed a logistic regression model trained with the training data. Subsequently, we utilized scikit-learn to obtain the feature importance associated with the expression values of various genes, as outlined in the following snippet:

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import
StandardScaler,MinMaxScaler,MaxAbsScaler
from sklearn.linear_model import LogisticRegression
from sklearn.utils.class_weight import compute_class_weight

train_df=pd.read_csv('Final_Dataset/Train_Datset.txt',sep='\t')

class_weights = compute_class_weight(class_weight='balanced',
                                     classes =
np.unique(train_df.Target.values),
```

## Chapter5

```
y=train_df.Target.values)

class_weights_dict=dict()
class_weights_dict[0]=class_weights[0]
class_weights_dict[1]=class_weights[1]

train_np=train_df.values
X_train=train_np[:, :-1]
Y_train=train_np[:, -1].astype(np.int8)

stdscaler=StandardScaler()
X_train=stdscaler.fit_transform(X_train)

model =
LogisticRegression(random_state=11, class_weight='balanced')
model.fit(X_train, Y_train)

feat_imp = pd.DataFrame(zip(train_df.columns[:-
1], model.coef_[0]*np.power(10, 2)), columns=['Feature', 'Value'])
feat_imp.sort_values(['Value'], inplace=True, ascending=False)
feat_imp.to_csv('Final_Dataset/LR_Feat_Imp.txt', index=False)
```

Where, `compute_class_weight` manages imbalance condition within training data.

### **Blindfold LR Feature Selection**

Initially, the feature importance scores extracted from the previous module are segregated into positive and negative values. Subsequently, for each combination of positive features, negative features are systematically combined and used to train logistic regression models, which are then validated with the validation data. The feature set that maximizes both accuracy and F1-score is selected for final analysis, as illustrated in the following snippet:

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import
StandardScaler, MinMaxScaler, MaxAbsScaler
from sklearn.linear_model import LogisticRegression
from tqdm import tqdm
from sklearn.metrics import accuracy_score

feat_imp=pd.read_csv('Final_Dataset/LR_Feat_Imp.txt')

feat_imp=feat_imp[feat_imp.Feature!='SNP']

feat_imp_pos=feat_imp[feat_imp.Value>0]
feat_imp_neg=feat_imp[feat_imp.Value<0]
```

```
feat_imp_pos.sort_values(['Value'],inplace=True,ascending=True)
feat_imp_neg.sort_values(['Value'],inplace=True,ascending=False)

print(feat_imp_pos.shape,feat_imp_neg.shape)
feat_imp_neg.head(3)

train_df=pd.read_csv('Final_Dataset/Train_Dataset.txt',sep='\t')
val_df=pd.read_csv('Final_Dataset/Val_Dataset.txt',sep='\t')

final=pd.DataFrame(columns=['Pos','Neg','Train_Score','Valid_Score'])
count=0

for pos in tqdm(range(0,feat_imp_pos.shape[0]+1)):
    for neg in range(1,feat_imp_neg.shape[0]+1):

        feat_list=['Target']+list(feat_imp_pos.head(pos).Feature.values)
        +list(feat_imp_neg.head(neg).Feature.values)+['SNP']
        train_df_sub=train_df[feat_list]
        val_df_sub=val_df[feat_list]

        train_np=train_df_sub.values
        X_train=train_np[:,1:]
        Y_train=train_np[:,0].astype(np.int8)

        val_np=val_df_sub.values
        X_val=val_np[:,1:]
        Y_val=val_np[:,0].astype(np.int8)

        stdscaler=StandardScaler()
        stdscaler.fit(X_train)
        X_train= stdscaler.transform(X_train)
        X_val= stdscaler.transform(X_val)

        model =
LogisticRegression(random_state=1,class_weight='balanced',max_iter=200)

        model.fit(X_train,Y_train)

        final.loc[count,'Pos']=pos
        final.loc[count,'Neg']=neg

        final.loc[count,'Train_Score']=model.score(X_train,Y_train
)
```

## Chapter5

```
final.loc[count, 'Valid_Score']=model.score(X_val,Y_val)
count=count+1
```

### **Final Logistic Regression Model**

Once the feature sets were selected, we proceeded to create the final logistic regression model based on these chosen features. This model underwent validation using both the validation dataset and the unknown test set to ensure its robustness and efficacy in predicting outcomes.

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import
StandardScaler,MinMaxScaler,MaxAbsScaler
from sklearn.linear_model import LogisticRegression
from sklearn.utils.class_weight import compute_class_weight
from sklearn.metrics import f1_score
from sklearn.metrics import
confusion_matrix,classification_report,f1_score,roc_auc_score,au
c,roc_curve,
import pickle

feat_imp=pd.read_csv('Final_Dataset/LR_Feat_Imp.txt')

feat_imp=feat_imp[feat_imp.Feature!='SNP']

feat_imp_pos=feat_imp[feat_imp.Value>0]
feat_imp_neg=feat_imp[feat_imp.Value<0]

feat_imp_pos.sort_values(['Value'],inplace=True,ascending=True)
feat_imp_neg.sort_values(['Value'],inplace=True,ascending=False)

print(feat_imp_pos.shape,feat_imp_neg.shape)
feat_imp_neg.head(3)

#Where pos and neg are number of features with positive and
negative values as determined by initial logistic regression
model

pos=137
neg=53
feat_list=['Target']+list(feat_imp_pos.head(pos).Feature.values)
+list(feat_imp_neg.head(neg).Feature.values)+['SNP']

train_df=pd.read_csv('Final_Dataset/Train_Datset.txt',sep='\t')
val_df=pd.read_csv('Final_Dataset/Val_Datset.txt',sep='\t')
test_df=pd.read_csv('Final_Dataset/Test_Datset.txt',sep='\t')
train_df.shape,val_df.shape,test_df.shape
```

```
train_df_sub=train_df[feat_list]
val_df_sub=val_df[feat_list]
test_df_sub=test_df[feat_list]

train_np=train_df_sub.values
X_train=train_np[:,1:]
Y_train=train_np[:,0].astype(np.int8)
Y_train[:3]

val_np=val_df_sub.values
X_val=val_np[:,1:]
Y_val=val_np[:,0].astype(np.int8)
Y_val[:3]

test_np=test_df_sub.values
X_test=test_np[:,1:]
Y_test=test_np[:,0].astype(np.int8)
Y_test[:3]

stdscaler=StandardScaler()

X_train=stdscaler.fit_transform(X_train)
X_val=stdscaler.transform(X_val)
X_test=stdscaler.transform(X_test)

model =
LogisticRegression(random_state=1,class_weight='balanced',max_iter=200)
model.fit(X_train,Y_train)

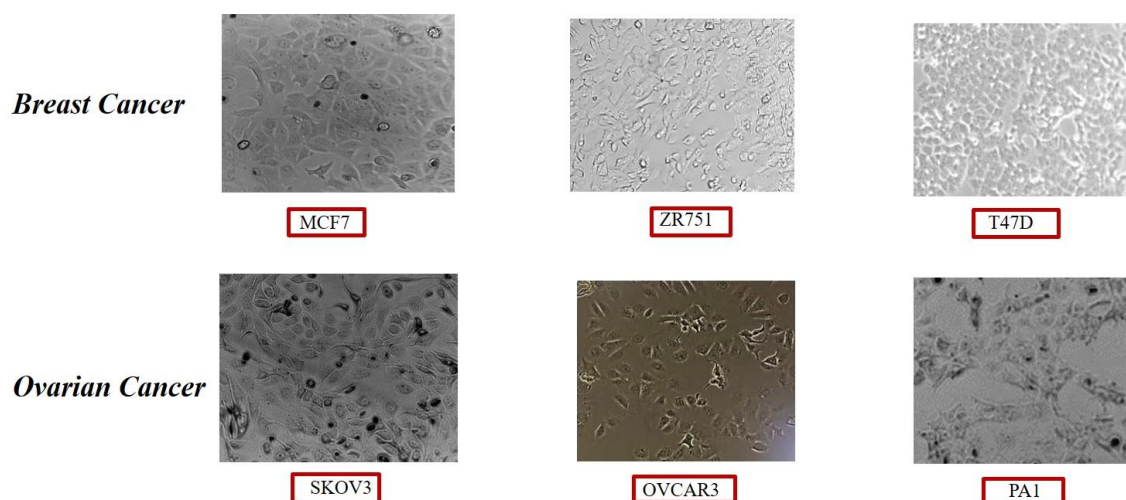
print(str(model.score(X_train,Y_train)),f1_score(Y_train,model.predict(X_train)))
print(str(model.score(X_val,Y_val)),f1_score(Y_val,model.predict(X_val)))
print(str(model.score(X_test,Y_test)),f1_score(Y_test,model.predict(X_test)))

pickle.dump(stdscaler,open('Models/rs2366152_stdscaler.sav','wb'))
pickle.dump(model,open('Models/rs2366152_LR.sav','wb'))
```

## Chapter5

### 2.7. Culture of Breast and Ovarian Cancer cell lines

The selected set of DE lncRNAs and their corresponding SNPs were chosen for experimental validation through wet bench procedures. Breast Cancer cell lines (MCF7, T47D, and ZR751) and Ovarian Cancer cell lines (SKOV3, OVCAR3, and PA1) were specifically selected for this validation. T47D, ZR751, and PA1 cell lines were sourced from the National Centre for Cell Science (NCCS), Pune, while MCF7, OVCAR3, and SKOV3 were acquired from Cell Line Service (CLS). OVCAR3, ZR751, and T47D were cultured in RPMI-1640 medium, MCF7 was cultured in DMEM medium, and SKOV3 was cultured in DMEM: Ham's F12 medium. All culture media contained 10% fetal bovine serum (FBS, Invitrogen) and 1% penicillin/streptomycin (Invitrogen). The cells were maintained in a 5% CO<sub>2</sub> incubator at 37 °C until reaching confluency. **Figure 2** illustrates the cellular condition at confluency.



**Figure2:** Culture and maintenance of Ovarian Cancer Cell line SKOV3, OVCAR3 and PA1.

### 2.8. Genotyping by the TaqMan PCR assay

Human adult normal Breast and Ovarian genomic DNA extracted from Human Ovarian Surface Epithelial cells (HOSEpiC) (Cat# D1234086 and 7319) was obtained from Sciencell (<https://sciencellonline.com/>). Genomic DNA from the specified cell lines was isolated using the HiPura mammalian genomic DNA preparation Kit provided by Himedia.

The Ready-to-use TaqMan® SNP Genotyping Assay mix for the selected SNPs was directly purchased (Assay ID: C\_\_\_3130439\_10). TaqMan PCR and genotyping analyses were performed on the Applied Biosystems 7500 Real-Time PCR System following the manufacturer's instructions. The reaction mixtures were amplified with 0.5 µl of 20X primer-probe mix, 5 µl of 2X TaqMan Universal Master Mix (Applied Biosystems), 1 µl of gDNA (10 ng/µl), and 3.5 µl of ddH<sub>2</sub>O, the total volume being 10 µl. PCR cycling

conditions are as follows : one cycle at 95 °C for 10 min; 40 cycles at 95 °C for 15 s and 58 °C for 1 min. The results were analyzed on the Applied Biosystems 7500 Real Time PCR System using the allelic discrimination assay program.

## 2.9. Real-time PCR analysis:

Normal Breast and Ovarian tissue RNA was procured from Biochain (Cat # R1234086-50 and C605167). Total RNA extraction from cell lines was carried out using the RNeasy® Mini kit (Qiagen), and complementary DNA (cDNA) synthesis was performed using the Verso cDNA Synthesis Kit (Thermo Fisher Scientific) following the manufacturer's instructions.

We aim to check the differential expression of the DE lncRNA-SNP as well as a selected list of feature genes corresponding to each lncRNA-SNP. Subsequently, qPCR for the selected lncRNAs was conducted using the SYBR Green PCR Master Mix (Applied Biosystems) on the Applied Biosystems 7500 Real-Time PCR System. Each experiment was performed in triplicate. Transcript expression levels were calculated using the 2- $\Delta\Delta C_t$  method, normalized to the expression of the endogenous control (18s). Primer sequences were designed using NCBI primer blast and Sigma Oligo Evaluator, and they were ordered from Sigma-Aldrich. The primer sequences used in the study, purchased from Merck, are summarized in **Table 2**. The entire workflow of the study is illustrated in **Figure 3**

Primer	Sequence
LINC00621 _Forward	GTTCTGGAGGCTGGGAAGTC
LINC00621 _Reverse	TTGCCCTTCCACCTTCTTCC
HOTAIR _Forward	CCAGAGAACGCTGGAAAAACCTG
HOTAIR _Reverse	GGAGATGATAAGAAGAGCAAGGAA
18s _Forward	GCGGCGTTATTCCCATGAC
18s _Reverse	GCTATCAATCTGTCAATCCTGTCC
FOXP1 _Forward	TAACGGTTCAGCCATCCAGAA
FOXP1 _Reverse	GGTTTATGAGATGCCACTGTTG
CFL1 _Forward	TTCCGGAAACATGGCCTC
CFL1 _Reverse	CTCCTCTGGCGTTGAAGACT
UQCRQ _Forward	CTACAGCTTGTACCGTTTCG
UQCRQ _Reverse	AACACTACAACTGCGGCAC
LAPTM4B _Forward	ATTTTATTGAGTGCCCTGGCTG
LAPTM4B _Reverse	CTGCGCGTTGCTTGTACG
PRKX _Forward	GAGATGCTTTCGGGGTTTCC
PRKX _Reverse	TTGTTCTGTCAACCACGAGC
CANT1 _Forward	GTCGTCTACCAGATCGAAGGCA
CANT1 _Reverse	TACAGACGCTCGTCCTTCACTG
S100PBP _Forward	GGGAGCTTTGTGCCTTGATGGA
S100PBP _Reverse	TCCTGTCAACCCACTGAGTCAG
TMTC1 _Forward	TGGTCTCTGATGAGGTGTCTG
TMTC1 _Reverse	TCCACAGGCGGAAATACAAAAT

**Table2:** Sequences of lncRNA and gene primers used for this study

### 3. Results and Discussion:

#### 3.1. Selection of low frequency variants exclusive in Breast and Ovarian Cancer patients:

After variant calling and filtration across patient transcriptomes, a chi-square test with a stringent significance threshold of  $p \leq 5e-12$  identified a total of 5777 distinct SNPs predominant in either Breast or Ovarian Cancer patients. Subsequent mapping of these SNPs within lncRNA loci, non-overlapping with any protein-coding region, followed by the selection of variants capable of disrupting the secondary structure of corresponding lncRNAs, reduced the number to 270.

Exploring allele frequencies across diverse populations, we observed that most SNPs were common variants, prevalent in the population. This prevalence could be attributed to natural selection, where certain SNP alleles confer resistance to diseases or offer survival advantages. Additionally, random fluctuations in allele frequencies over time may contribute to their increased prevalence, despite lacking selective advantages.

While common variants are implicated in disease susceptibility, our focus shifted towards *rare* or *low-frequency variants* in disease risk association studies. These variants hold significant interest due to their potential to exert a larger effect on disease risk compared to common variants. They exhibit population specificity and are less likely to be in linkage disequilibrium with nearby markers, thereby directly tagging specific functional variants or genes. Consequently, we excluded common variants from our study, allowing us to concentrate on less-frequent variants potentially pivotal in elucidating the genetic factors underlying disease susceptibility.

#### 3.2. Addressing Imbalanced Case-Control Data and Insignificant Odds Ratios through Machine Learning-Based Approach

After filtration of common variants which left us with 21 SNPs, we categorized them as Breast Cancer specific and Ovarian cancer specific based on the number of affected patients in each cancer type. We then sought to investigate the association between these SNPs and their respective cancers using Odds Ratios (ORs). However, we encountered a significant challenge with imbalanced case-control data. The number of true healthy controls were 10 times and 5 times less than case numbers in Breast and Ovarian cancer respectively.

This imbalance reduced the statistical power to detect associations, especially given the small sample size of controls. Consequently, wider confidence intervals around the estimated ORs led to larger p-values, rendering potentially influential variants statistically insignificant[47]. Despite obtaining  $OR > 1$  for 15 SNPs (4 specific to Breast Cancer and 11 to Ovarian Cancer), the associated p-values remained insignificant in many cases.



Recognizing the limitations of traditional analytical approaches in addressing this complexity, we turned to ML as a promising solution[47]. ML algorithms offer resampling techniques such as oversampling the minority class (controls) or under-sampling the majority class (cases) to balance the dataset, mitigating bias towards the majority class and enhancing the accuracy of SNP significance analysis. Additionally, employing metrics such as precision, recall, F1-score, or the area under the Receiver Operating Characteristic (ROC) curve allows for a comprehensive evaluation of model performance, particularly in the context of imbalanced datasets.

### **3.3. LncRNA-SNP mediated effects on gene regulation for improved model efficiency**

With the aim of creating separate models for breast and ovarian cancer risk prediction we utilized gene expression data of those genes that interacts with the lncRNAs harbouring SNPs. This served as features to construct the cancer risk prediction model for the selected lncRNA-SNPs. Instead of solely correlating gene expression changes with the presence or absence of SNPs, we adopted a strategy where features were selected based on the lncRNA within which the SNP resides. This approach allows us to incorporate both the SNP itself and its effect mediated through the function of the associated lncRNAs into the prediction model. We investigated the presence of DE coding genes within the vicinity of the lncRNAs, also identified regulatory players through guilt by association studies. Recognizing ceRNA as one of the well-established mechanisms of lncRNA regulation of coding genes, we also carried out analyses to incorporate this aspect into our study. Additionally, physically interacting lncRNA-mRNA pairs from RNA-RNA interactome data were integrated, considering their DE status in patient datasets.

Based on available datasets, we incorporated endometrial samples with and without endometriosis as non-cancerous controls. For Breast cancer analysis, we included adjacent normal tissue datasets. While adjacent normal tissue may harbor the same SNPs as that by cancerous tissue, differences in gene expression patterns of the interacting partners can arise due to tumor-induced changes, epigenetic modifications, the tumor microenvironment, clonal selection, and post-transcriptional regulation. We believe that considering these aspects will enhance the robustness of our prediction model and its applicability in real-world clinical scenarios.

Following feature selection, we generated four gene expression matrices to serve as input for each lncRNA-SNP. These matrices were structured to capture distinct sample groups: (a) Control/Adjacent Normal/ Non-cancerous samples containing the SNP of interest, (b) Control/ Adjacent normal/ Non-cancerous samples lacking the SNP, (c) Cancer samples with the SNP, and (d) Cancer samples without the SNP. Each matrix represents a specific combination of genetic and disease status, facilitating the training and evaluation of separate predictive models for each scenario. By organizing the data in this manner, we aim to discern the unique gene expression patterns associated with the presence or absence of the SNP in both control and cancer samples. To address the imbalanced

## Chapter 5

classification problem, we specified the `class_weight` parameter as "balanced", allowing class-specific weights to be automatically calibrated inversely according to their frequency. Performance metrics associated with the Breast and Ovarian cancer models have been detailed in **Table 3** and **Table 4** respectively.

SNP_ID	lncRNA	Validation Data			Test Data		
		Accuracy	F1-Score	ROC-AUC	Accuracy	F1-Score	ROC-AUC
rs2366152	HOTAIR	89.66	93.13	95.51	87.5	91.67	82.73
rs3803699	CEROX1	87.36	91.47	90.87	90.63	93.62	93.64
rs62484970	AZGP1P1	87.36	91.2	97.45	87.5	91.3	92.73
rs7091441	LINC00993	88.51	92.06	96.52	87.5	91.67	95

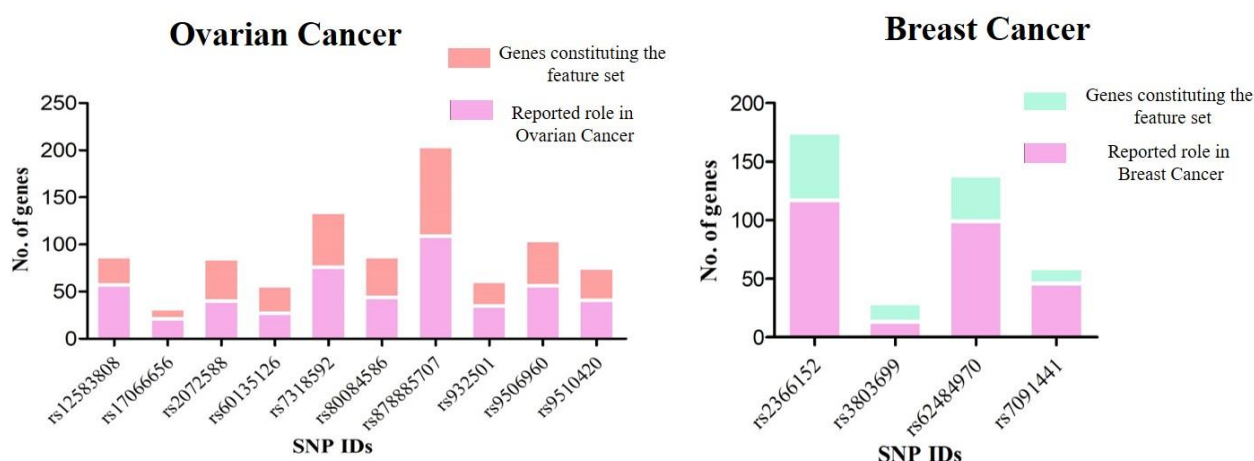
**Table 3:** Performance metric for Breast Cancer

SNP_ID	lncRNA	Validation Data			Test Data		
		Accuracy	F1-Score	ROC-AUC	Accuracy	F1-Score	ROC-AUC
rs932501	AL079338.1	95.77	96.91	99.81	84.38	87.18	96.36
rs2072588	HAGLROS	97.18	97.96	99.71	81.25	84.21	89.09
rs7318592	LINC00621	97.18	98	99.05	90.63	92.68	99.09
rs9506960	LINC00621	95.77	96.91	98.86	87.5	90	98.64
rs9510420	LINC00621	91.55	93.75	97.14	87.5	90	94.09
rs12583808	LINC00621	98.59	98.99	99.81	81.25	84.21	92.73
rs60135126	LINC00621	95.77	96.91	99.43	87.5	90	91.82
rs17066656	TPT1-AS1	90.14	92.63	96.95	84.38	87.18	92.27
rs80084586	BAALC-AS1	95.77	96.97	98.86	93.75	95.45	95
rs878885707	MDN1 Antisense	92.96	94.74	98.76	90.63	92.68	96.82
rs879083236	AF279873.1	94.37	95.83	98.29	96.88	97.67	99.55
rs1007064349	NACAP1	95.77	97.09	98.86	90.63	93.02	92.27

**Table 4:** Performance metric for Ovarian Cancer

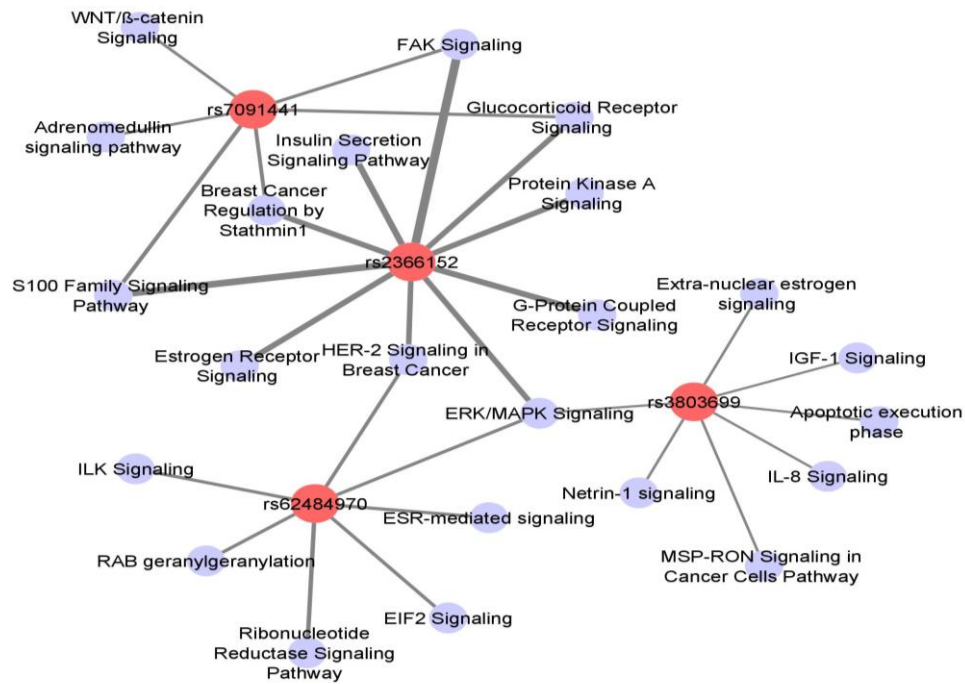
### 3.4. Pathway and gene set enrichment of the genes constituting the feature set

The genes finally selected (constituting the feature set) for each lncRNA-SNP were then checked for their functional involvement in the corresponding cancers using QIAGEN Ingenuity Pathway Analysis (Qiagen IPA) (<https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-ipa/>) For all the 4 Breast lncRNA-SNPs and 10 among 12 Ovarian lncRNA-SNPs, more than 50% of the feature gene set revealed their involvement in their corresponding cancers. For the two Ovary specific lncRNA-SNPs, rs1007064349 and rs879083236, only a few of the interacting genes were found to be involved in Ovarian cancer (based on literature evidence) and hence were not considered further. Figure 4 represents the final no. of interacting genes constituting the feature set for each lncRNA-SNP in Breast and Ovarian cancer as well as the proportion of it having their reported role in each of these cancers.

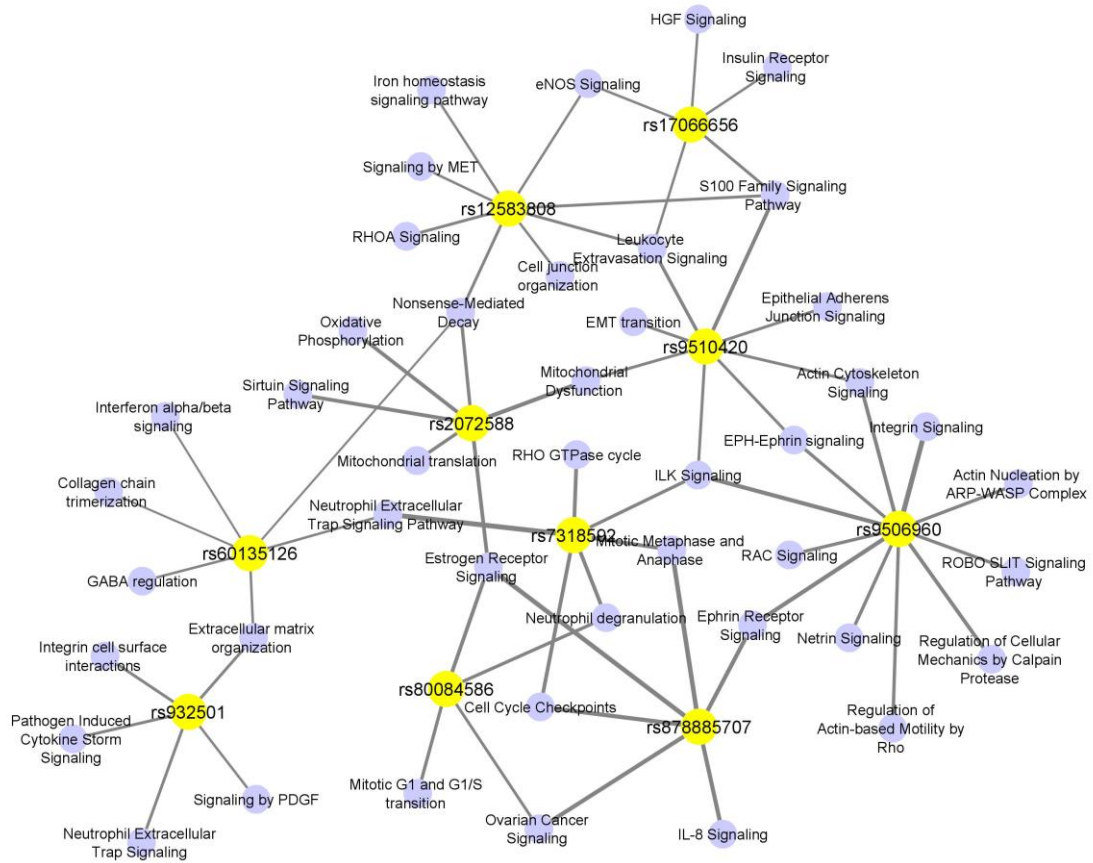


**Figure 4:** The number of genes constituting the feature sets for lncRNA-SNPs models of (a) Ovary and (b) Breast cancers. The pink region indicates genes having reported role in the respective cancers.

We proceeded to analyze the top significant pathways (with p-values < 0.05) associated with each lncRNA-SNP, which revealed their involvement in key cancer-regulating pathways. In the context of Breast cancer, multiple models exhibited enrichment in top cancer pathways such as ERK/MAPK Signaling, FAK Signaling, Glucocorticoid Receptor Signaling, HER-2 Signaling in Breast Cancer, and the S100 Family Signaling Pathway. Conversely, for Ovarian cancer models, pathways like Estrogen Receptor Signaling, ILK Signaling, Leukocyte Extravasation Signaling, Nonsense-Mediated Decay, and the S100 Family Signaling Pathway were enriched. **Figures 5a** and **5b** present the cytoscape network illustrating the top enriched cancer pathways for each SNP-lncRNA model in Breast and Ovarian cancers respectively.



**Figure 5a:** Gene set enrichment analysis revealed top significant cancer pathways related to Breast cancer specific lncRNA-SNPs. Breast cancer specific lncRNA-SNPs are indicated in red, pathways in blue and thickness of the edge reveals the no. of genes involved in each pathway



**Figure 5b:** Gene set enrichment analysis revealed top significant cancer pathways related to Ovarian cancer lncRNA-SNPs. Ovarian cancer specific lncRNA-SNPs are indicated in yellow, pathways in blue and thickness of the edge reveals the no of genes involved in each pathway

### 3.5. Screening the most promising lncRNA-SNPs for wet lab validation

Reviewing our final selection of ovarian lncRNA-SNPs, we made an intriguing observation: LINC00621 harbors five out of the twelve chosen SNPs. This clustering of SNPs within the locus of LINC00621 sparked our curiosity. Upon delving deeper, our literature survey unveiled LINC00621 as a novel onco-lncRNA in lung cancer [48]. However, its involvement in ovarian cancer has not been explored. Leveraging our previous work (Chapter3) where we analyzed RNA-seq data from ovarian cancer cell lines for variant detection, we investigated the status of LINC00621 SNPs in the ovarian cancer cell lines at our disposal. We discovered that rs9510420 was present in all SKOV3 cell lines and in our laboratory-generated PA1 cell line dataset, but notably absent in OVCAR3. Consequently, we selected LINC00621-rs9510420 and for wet lab validation.

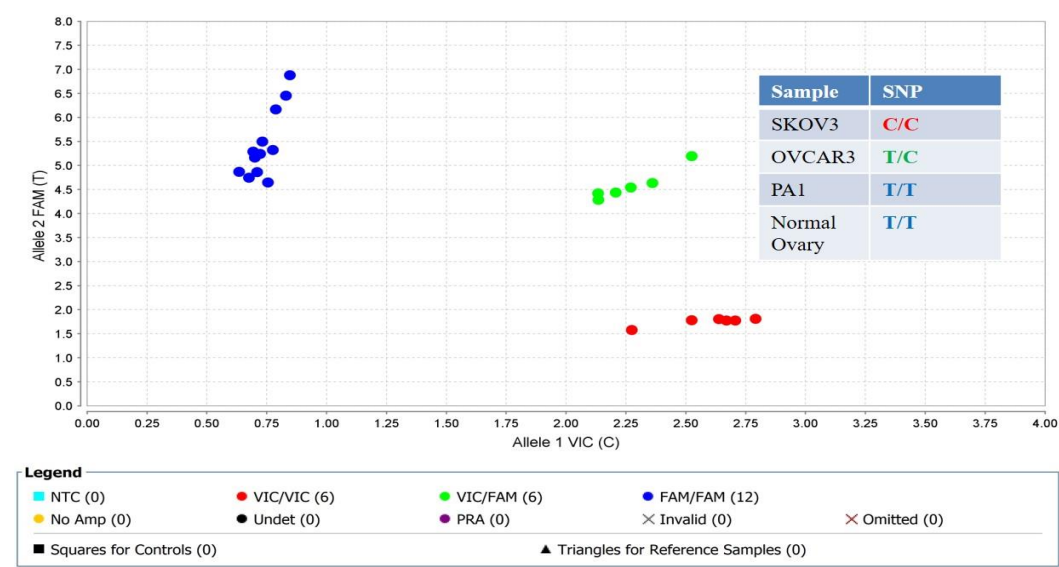
For breast cancer, we selected HOTAIR, a well-recognized lncRNA implicated in dysregulated functions within breast cancer [49]. rs2366152, residing within HOTAIR loci, has previously been associated with Colorectal Cancer susceptibility in the Iranian population[50] and Cervical Cancer susceptibility in the Polish population[51]. However,

Chapter5

its involvement in breast cancer remains unexplored in existing literature. Through variant analysis of Breast Cancer cell line transcriptome data, we detected the presence of this SNP in cell lines MCF7 and T47D. Hence, we proceeded with HOTAIR-rs2366152 for validation.

3.6. Validating the presence of the ovarian cancer specific SNP rs9510420 harboured by the lncRNA LINC00621 and qRT-PCR validation of the corresponding lncRNA

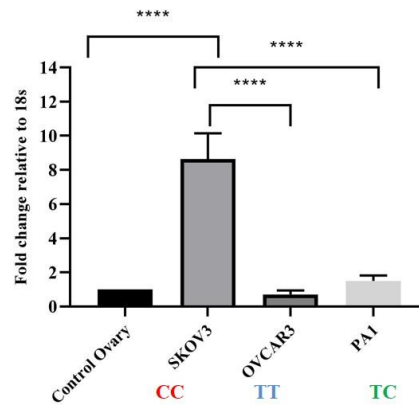
The outcomes derived from the TaqMan genotyping assay, as illustrated in **Figure 6**, indicate the presence of homozygous dominant allele (CC) of rs9510420 in Control Ovary and OVCAR3 line, homozygous recessive (TT) allele in SKOV3 and heterozygous variants (CT) in PA1.



**Figure 6:** TaqMan genotyping assay revealed the presence of homozygous dominant TT allele (blue) in Normal Ovary and OVCAR3, homozygous recessive CC allele (red) in SKOV3 and heterozygous TC allele (green) in PA1 cell line.

We carried out the differential expression of the lncRNA LINC00621 in the 3 Ovarian cancer cell lines by comparing them with their corresponding normal counterpart, as illustrated in **Figure 7**. Utilizing 18s as the endogenous control, our q-PCR analysis of cancer cell lines with their normal counterpart revealed a statistically significant increase in the expression of the lncRNA in the SKOV3 cell line harboring the SNP (CC). Additionally, we observed a marginal increase in its expression in PA1 carrying the heterozygous allele (TC) and a slight decrease in OVCAR3 carrying the homozygous dominant allele (TT), though these findings did not reach statistical significance in our analysis.

Here, we note an increase in the expression levels of LINC00621 in the cell line carrying SNP rs10425267, implying the likelihood of this genetic variant impacting the expression pattern of the lncRNA within the ovarian cancer system.



**Figure 7:** qPCR analysis revealed upregulated expression of LINC00621 in cell line of Ovarian cancer harbouring SNP rs10425267.

### 3.7. Validating the presence of the breast cancer specific SNP rs2366152 harboured by the lncRNA HOTAIR and qRT-PCR validation of the corresponding lncRNA

Currently, the process of validating rs2366152 within HOTAIR and its effect on the lncRNA expression in different Breast cancer cell lines is underway using the same experimental designs.

Additionally, we plan to validate a subset of genes from the gene feature list associated with the two lncRNA-SNPs. This validation aims to assess their differential regulation patterns in cell lines with and without the SNP, aligning with our in-silico predictions.

To strengthen these findings and broaden their relevance, it is crucial to carry out additional validation across a wider range of Ovarian cancer cell lines. By including a more diverse array of cancer types and incorporating various cellular contexts, this expanded validation will not only reinforce the connection between the identified SNP and LINC00621 expression but also enhance our comprehension of its regulatory function within the Ovarian system.

## 4. Conclusion:

In conclusion, our study sheds light on the intricate interplay between lncRNA-SNPs, and their role as early markers in Breast and Ovarian cancers. Through meticulous analysis of publicly available cancer datasets and ML techniques, we identified and

## Chapter5

characterized Breast and Ovary specific lncRNA-SNP associations that hold promise as potential markers for cancer risk prediction. By focusing on less-frequent variants within lncRNA loci, we aimed to elucidate their regulatory roles and their impact on disease susceptibility. Addressing the challenge of imbalanced case-control data, we employed ML algorithms to enhance the accuracy of SNP significance analysis. Furthermore, we developed separate predictive models for each SNP, incorporating gene expression data and considering the regulatory influence of associated lncRNAs. Through pathway analysis, we uncovered the involvement of key cancer-regulating pathways, highlighting the potential mechanisms by which lncRNAs and SNPs contribute to disease pathogenesis. The model also offers particular value for individuals with predisposing conditions, facilitating timely interventions. Notably, our findings underscore the importance of personalized treatment approaches tailored to the molecular profiles of individual tumors. By integrating wet bench validation, we have tried to strength our conclusion regarding the possible diagnostic implication of these lncRNA-SNPs. The integration of transcriptomic analysis into clinical practice holds promise for improving cancer risk stratification and guiding personalized treatment approaches. Overall, our study puts forward the importance of lncRNA-SNPs towards inducing Breast and Ovarian cancers, laying the groundwork for improved diagnostic and therapeutic interventions.

### 5. References:

1. Mercer, T.R., M.E. Dinger, and J.S. Mattick, *Long non-coding RNAs: insights into functions*. Nat Rev Genet, 2009. **10**(3): p. 155-9.
2. Wilusz, J.E., H. Sunwoo, and D.L. Spector, *Long noncoding RNAs: functional surprises from the RNA world*. Genes Dev, 2009. **23**(13): p. 1494-504.
3. Rinn, J.L. and H.Y. Chang, *Genome regulation by long noncoding RNAs*. Annu Rev Biochem, 2012. **81**: p. 145-66.
4. Reich, D.E., S.B. Gabriel, and D. Altshuler, *Quality and completeness of SNP databases*. Nat Genet, 2003. **33**(4): p. 457-8.
5. Erichsen, H.C. and S.J. Chanock, *SNPs in cancer research and treatment*. Br J Cancer, 2004. **90**(4): p. 747-51.
6. Kulkarni, S., et al., *CCR5AS lncRNA variation differentially regulates CCR5, influencing HIV disease outcome*. Nat Immunol, 2019. **20**(7): p. 824-834.
7. Aznaourova, M., et al., *Disease-Causing Mutations and Rearrangements in Long Non-coding RNA Gene Loci*. Front Genet, 2020. **11**: p. 527484.
8. Al Sarakbi, W., et al., *Breast papillomas: current management with a focus on a new diagnostic and therapeutic modality*. Int Semin Surg Oncol, 2006. **3**: p. 1.
9. Hartmann, L.C., et al., *Atypical hyperplasia of the breast--risk assessment and management options*. N Engl J Med, 2015. **372**(1): p. 78-89.
10. Hartmann, L.C., et al., *Understanding the Premalignant Potential of Atypical Hyperplasia through Its Natural History: A Longitudinal Cohort Study*. Cancer Prevention Research, 2014. **7**(2): p. 211-217.
11. Ali-Fehmi, R., et al., *Clinicopathologic analysis of breast lesions associated with multiple papillomas*. Human Pathology, 2003. **34**(3): p. 234-239.
12. Brilhante, A.V., et al., *Endometriosis and Ovarian Cancer: an Integrative Review (Endometriosis and Ovarian Cancer)*. Asian Pac J Cancer Prev, 2017. **18**(1): p. 11-16.
13. Daniilidis, A. and K. Dinas, *Long term health consequences of polycystic ovarian syndrome: a review analysis*. Hippokratia, 2009. **13**(2): p. 90-2.



14. Carmina, E. and R.A. Lobo, *Polycystic Ovary Syndrome (PCOS): Arguably the Most Common Endocrinopathy Is Associated with Significant Morbidity in Women*. The Journal of Clinical Endocrinology & Metabolism, 1999. **84**(6): p. 1897-1899.
15. Murakami, K., et al., *Endometriosis-Associated Ovarian Cancer: The Origin and Targeted Therapy*. Cancers (Basel), 2020. **12**(6).
16. Throwba, H., et al., *The epigenetic correlation among ovarian cancer, endometriosis and PCOS: A review*. Critical Reviews in Oncology/Hematology, 2022. **180**: p. 103852.
17. Kader, T., et al., *The genetic architecture of breast papillary lesions as a predictor of progression to carcinoma*. NPJ Breast Cancer, 2020. **6**: p. 9.
18. Hennessy, B.T., R.L. Coleman, and M. Markman, *Ovarian cancer*. The Lancet, 2009. **374**(9698): p. 1371-1382.
19. Mehrgou, A. and M. Akouchekian, *The importance of BRCA1 and BRCA2 genes mutations in breast cancer development*. Med J Islam Repub Iran, 2016. **30**: p. 369.
20. Kim, J. and R.J. DeBerardinis, *Mechanisms and Implications of Metabolic Heterogeneity in Cancer*. Cell Metab, 2019. **30**(3): p. 434-446.
21. Santana Dos Santos, E., et al., *HRness in Breast and Ovarian Cancers*. Int J Mol Sci, 2020. **21**(11).
22. Liu, Z., et al., *Over-expressed long noncoding RNA HOXA11-AS promotes cell cycle progression and metastasis in gastric cancer*. Mol Cancer, 2017. **16**(1): p. 82.
23. Hassanzarei, S., et al., *Genetic polymorphisms of HOTAIR gene are associated with the risk of breast cancer in a sample of southeast Iranian population*. Tumour Biol, 2017. **39**(10): p. 1010428317727539.
24. Saeedi, N. and S. Ghorbian, *Analysis of clinical important of LncRNA-HOTAIR gene variations and ovarian cancer susceptibility*. Molecular Biology Reports, 2020. **47**(10): p. 7421-7427.
25. Khorshidi, H.R., et al., *ANRIL Genetic Variants in Iranian Breast Cancer Patients*. Cell J, 2017. **19**(Suppl 1): p. 72-78.
26. Barrett, T. and R. Edgar, *Gene expression omnibus: microarray data storage, submission, retrieval, and analysis*. Methods Enzymol, 2006. **411**: p. 352-69.
27. Kanz, C., et al., *The EMBL Nucleotide Sequence Database*. Nucleic Acids Res, 2005. **33**(Database issue): p. D29-33.
28. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation*. Nucleic Acids Res, 2001. **29**(1): p. 308-11.
29. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads*. 2011, 2011. **17**(1): p. 3.
30. Pertea, M., et al., *Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown*. Nature Protocols, 2016. **11**(9): p. 1650-1667.
31. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
32. Oikonen, L. and S. Lise, *Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection*. Wellcome Open Res, 2017. **2**: p. 6.
33. Rimmer, A., et al., *Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications*. Nat Genet, 2014. **46**(8): p. 912-918.
34. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features*. Bioinformatics, 2010. **26**(6): p. 841-2.
35. Sabarinathan, R., et al., *The RNAsnp web server: predicting SNP effects on local RNA secondary structure*. Nucleic Acids Res, 2013. **41**(Web Server issue): p. W475-9.
36. Kido, T., et al., *Are minor alleles more likely to be risk alleles?* BMC Med Genomics, 2018. **11**(1): p. 3.
37. Das, T., et al., *LncRBase V.2: an updated resource for multispecies lncRNAs and ClinicLSNP hosting genetic variants in lncRNAs for cancer patients*. RNA Biol, 2021. **18**(8): p. 1136-1151.
38. Frankish, A., et al., *GENCODE 2021*. Nucleic Acids Res, 2021. **49**(D1): p. D916-D923.
39. Guo, S., et al., *Novel Breast-Specific Long Non-coding RNA LINC00993 Acts as a Tumor Suppressor in Triple-Negative Breast Cancer*. Front Oncol, 2019. **9**: p. 1325.

## Chapter5

40. Lefever, S., et al., *decodeRNA- predicting non-coding RNA functions using guilt-by-association*. Database (Oxford), 2017. **2017**.
41. Li, J.H., et al., *starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data*. Nucleic Acids Res, 2014. **42**(Database issue): p. D92-7.
42. Sharma, E., et al., *Global Mapping of Human RNA-RNA Interactions*. Mol Cell, 2016. **62**(4): p. 618-26.
43. Lu, Z., J. Gong, and Q.C. Zhang, *PARIS: Psoralen Analysis of RNA Interactions and Structures with High Throughput and Resolution*. Methods Mol Biol, 2018. **1649**: p. 59-84.
44. Chaung, K., et al., *SPLASH: A statistical, reference-free genomic algorithm unifies biological discovery*. Cell, 2023. **186**(25): p. 5440-5456.e26.
45. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature*. Nucleic Acids Res, 2006. **34**(Database issue): p. D140-4.
46. Grimson, A., et al., *MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing*. Molecular Cell, 2007. **27**(1): p. 91-105.
47. Dai, X., et al., *Statistical Learning Methods Applicable to Genome-Wide Association Studies on Unbalanced Case-Control Disease Data*. Genes (Basel), 2021. **12**(5).
48. Wei, J., et al., *FOXA1-induced LINC00621 promotes lung adenocarcinoma progression via activating the TGF- $\beta$  signaling pathway*. Thorac Cancer, 2023. **14**(21): p. 2026-2037.
49. Cantile, M., et al., *Long Non-Coding RNA HOTAIR in Breast Cancer Therapy*. Cancers (Basel), 2020. **12**(5).
50. Eivazi, N., et al., *Association of HOTAIR rs2366152 and rs1899663 polymorphisms with colorectal cancer susceptibility in Iranian population: A case-control study*. J Clin Lab Anal, 2023. **37**(9-10): p. e24931.
51. Łażniak, S., et al., *Role of rs2366152 single-nucleotide variant located in the long noncoding RNA HOTAIR gene in the cervical cancer susceptibility in a Polish population*. J Appl Genet, 2023.

## CHAPTER 6| General discussion and future perspectives

The impact of single nucleotide polymorphisms (SNPs) within coding genes on their translational function is a well-established concept. However, in recent times, a paradigm shift has occurred, shedding light on the significance of SNPs within non-coding regions of the genome, especially within long non-coding RNAs (lncRNAs), which do not encode proteins. This has raised questions about their functional importance. While the role of SNPs in female cancers, such as breast and ovarian cancers, is widely recognized, there is limited information regarding non-coding SNPs and their effects on these cancers. This thesis seeks to unravel the distribution of such SNPs across lncRNA loci in the context of the three female cancer systems prevalent worldwide. Furthermore, it aims to discern whether there are common or distinct regulatory patterns among these cancers, guided by the interplay between lncRNA-SNP pairs. This research delves into uncharted territories, potentially advancing our understanding of the molecular underpinnings of female cancers and the significance of non-coding SNPs in these complex diseases.

The initial phase of this work is dedicated to the comprehensive study of the lncRNA repertoire across eight different species, including human and mouse. With the initial release of LncRBase in 2014, we witnessed a substantial surge in the annotation of lncRNAs in various species beyond just human and mouse. Consequently, there arose a compelling need to amalgamate our understanding of lncRNAs, encompassing their genomic and sequence attributes, as well as their distribution throughout the genome.

In this upgraded version of LncRBase, we have not only retained the features of the first version but have also enriched it with information regarding the probable subcellular location of lncRNAs, their micro-peptide coding potential, embedded miRNA details, transcription factor binding sites, tissue-specific expression profiles, and meticulously curated data regarding their target genes and associations with diseases. The update of LncRBase was an undertaking that involved a comprehensive annotation and characterization of lncRNAs. Looking ahead, there are exciting prospects to expand this knowledge to include even more species. Furthermore, exploring the structural classification of lncRNAs including circular RNAs, emerging as novel and increasingly recognized components of the lncRNAs class is a promising avenue that can significantly influence our understanding of their functionality and regulatory roles.

Building upon the foundation laid by LncRBase V.2, my next work centres on investigating the presence of SNPs within the loci of lncRNAs in the three most prevalent female cancers: breast, cervical, and ovarian. These SNPs are meticulously identified through the analysis of patient transcriptomes and cancer cell lines, with precise mapping within the lncRNA loci. My exploration also takes into account the specific subtypes within each cancer.

Moreover, I've delved into an in-depth analysis of the disease-causing potential of these variants, their capacity to influence the secondary structure of the lncRNA, and any potential overlap with genomic elements. These efforts culminate in the creation of ClinicLSNP, which serves as a valuable resource housing information pertaining to cancer-specific lncRNA variants and their associated data. In the future, our goal is to further expand this work by including insights into the impact of SNP-lncRNA interactions on their target genes. This may be in the form of ce-

lncRNA or through other relevant mechanisms, and there is a need for more comprehensive and detailed study in this area.

The next phase of work focuses on exploring common lncRNA-SNP that play a guiding role in shaping the tumor landscape of the three prevalent female cancers. Patient transcriptomic data from public resources such as GEO and EMBL has been employed to detect both upregulated lncRNAs and SNPs within the lncRNA loci. However, with the progress of the work, only one significant common lncRNA-SNP shared among these three cancer systems could be identified. Notably, the lncRNA MIR4435-2HG, is a well-established oncogenic lncRNA in cancer, including the three systems of our interest. However, the role of the SNP within this lncRNA has not been previously investigated. Through wet lab validation in two or more cell lines specific to each cancer type, it was observed that the presence of a heterozygous allele of SNP rs10425267 has the potential to influence the elevated expression of lncRNA MIR4435-2HG in all three cancer systems.

It is important to acknowledge that the cervical cancer dataset was relatively sparse, with fewer detected lncRNAs compared to breast and ovarian cancer. While several elevated lncRNAs common to breast and ovarian cancer were identified, I was unable to find the same overlaps with cervical cancer. With the anticipation of more datasets becoming available in the future, the likelihood of detecting additional lncRNAs harbouring SNPs, commonly shared between cervical carcinoma and breast and ovarian carcinoma is expected to increase.

In the concluding section of my thesis, I strived to put forward cancer risk prediction models based on gene expression changes influenced by the presence or absence of lncRNA-SNPs in case of patients with abnormal breast or ovary conditions that increases the risk of getting these cancers. Statistical approach was adopted along with existing biological knowledge to select crucial lncRNA-SNP feature and the models were trained using both cancerous and non-cancerous samples to ensure robustness and generalizability. Unfortunately, it was not possible to include cervical cancer in the analysis due to non-availability of adequate dataset needed for training the prediction model. It is important to mention that this study is based on publicly available data from various regions and ethnicities. To enhance the relevance of this work, it would be valuable to acquire a substantial amount of region-specific patient data. Conducting ethnicity-based studies can provide insights into variations unique to specific populations, offering a deeper understanding of the influencing factors that may not be evident elsewhere. Furthermore, it is noteworthy that there is a scarcity of research specifically focusing on Indian variants of Breast and Ovarian cancer specifically present in the lncRNA loci. Considering the genetic diversity among populations, investigating lncRNA-SNPs influencing disease progression within the Indian population holds significant promise. Understanding the unique genetic signatures and molecular mechanisms underlying cancer susceptibility in Indian individuals could provide invaluable insights to develop personalized therapy tailored for this population.

In summary, these studies have provided me with a comprehensive understanding of the distribution of SNPs within the lncRNA loci across the three cancer systems, their genomic and corresponding clinical characteristics, their potential to impact lncRNA expression, and the mechanisms through which they interact with other partners. To strengthen the conclusions, further genotyping assays of these SNPs on clinical samples will be worth doing so as to harness its applicative potential on a clinical setting.

## APPENDIX

### PEER-REVIEWED PUBLICATIONS (Sorted by Date)

(# equal contribution)

1. Debsharma S, Pramanik S, Bindu S, Mazumder S, **Das T**, Pal U, Saha D, De R, Nag S, Banerjee C, Maiti N C, Ghosh Z, Bandyopadhyay U: *NSAID targets SIRT3 to trigger mitochondrial dysfunction and gastric cancer cell death*. iScience, 2024
2. Birari P, Mal Soumyav Majumder D, Sharma A K, Kumar M, **Das T**, Ghosh Z, Jana K, Gupta U D, Kundu M, Basu J: *Nur77 influences immunometabolism to regulate the release of proinflammatory cytokines and the formation of lipid bodies during Mycobacterium tuberculosis infection of macrophages*. Pathogens and Disease, 2023. 81
3. Ghosh B #, **Das T** #, Das G #, Chowdhury N #, Bagchi A, Ghosh Z: *Mapping Drug-gene Interactions to Identify Potential Drug Candidates Targeting Envelope Protein in SARS-CoV-2 Infection*. Current Bioinformatics 2023, 18(9):760-773.
4. Debsharma S, Pramanik S, Bindu S, Mazumder S, **Das T**, Saha D, De R, Nag S, Banerjee C, Siddiqui AA et al: *Honokiol, an inducer of sirtuin-3, protects against non-steroidal anti-inflammatory drug-induced gastric mucosal mitochondrial pathology, apoptosis and inflammatory tissue injury*. Br J Pharmacol 2023, 180(18):2317-2340.  
  
Das G, **Das T**, Parida S, Ghosh Z. LncRTPred: Predicting RNA–RNA mode of interaction mediated by lncRNA. IUBMB Life. 2024; 76(1): 53–68.
5. Sarkar A #, **Das T** #, Das G #, Ghosh Z: *MicroRNA mediated gene regulatory circuits leads to machine learning based preliminary detection of acute myeloid leukemia*. Computational Biology and Chemistry 2023, 104:107859.
6. Acharya U, **Das T**, Ghosh Z, Ghosh A: *Defense Surveillance System at the Interface: Response of Rice Towards Rhizoctonia solani During Sheath Blight Infection*. Molecular Plant-Microbe Interactions® 2022, 35(12):1081-1095.
7. Debnath S, Sarkar A, Mukherjee DD, Ray S, Mahata B, Mahata T, Parida PK, **Das T**, Mukhopadhyay R, Ghosh Z et al: *Eriodictyol mediated selective targeting of the TNFR1/FADD/TRADD axis in cancer cells induce apoptosis and inhibit tumor progression and metastasis*. Translational Oncology 2022, 21:101433.

8. Ganguly P, Roy D, **Das T**, Kundu A, Cartieaux F, Ghosh Z, DasGupta M: *The Natural Antisense Transcript DONE40 Derived from the lncRNA ENOD40 Locus Interacts with SET Domain Protein ASHR3 During Inception of Symbiosis in Arachis hypogaea*. Molecular Plant-Microbe Interactions® 2021, 34(9):1057-1070.
9. Das G #, **Das T** #, Chowdhury N #, Chatterjee D #, Bagchi A, Ghosh Z: *Repurposed drugs and nutraceuticals targeting envelope protein: A possible therapeutic strategy against COVID-19*. Genomics 2021, 113(1, Part 2):1129-1140.
10. Mondal S, Bhattacharya N, **Das T**, Ghosh Z, Khatua S: *Implication of Statistical Methods on Patient Data: An Approach for Cancer Survivability Prediction*. In: Intelligent Healthcare: Applications of AI in eHealth. edn. Edited by Bhatia S, Dubey AK, Chhikara R, Chaudhary P, Kumar A. Cham: Springer International Publishing; 2021: 57-80.
11. **Das T**, Deb A, Parida S, Mondal S, Khatua S, Ghosh Z: *LncRBase V.2: an updated resource for multispecies lncRNAs and ClinicLSNP hosting genetic variants in lncRNAs for cancer patients*. RNA biology 2021, 18(8):1136-1151.

**TROYEE DAS**[troyee@jcbose.ac.in](mailto:troyee@jcbose.ac.in) ; [dastrovee1993@gmail.com](mailto:dastrovee1993@gmail.com)

Department of Biological Sciences  
Bose Institute  
EN-80, Sector V  
Kolkata - 700091  
West Bengal, India

J3/1204  
Block K, SP Sukhobrishti Ln.  
Action Area III, New Town  
Kolkata -700135  
West Bengal, India  
(+91) 9674511772

**CURRENT POSITION**

<b>Bioinformatics Centre, Bose Institute</b>	Kolkata, India
<b>Senior Research Fellow</b>	2017-Present
Pursuing Ph.D in Dept. Life Science and Biotechnology, Jadavpur University, India	

**EDUCATION**

<b>University of Calcutta</b>	
M.Sc, Microbiology. First Class, University Rank 1	2016
<b>University of Calcutta</b>	
B.Sc, Microbiology. First Class, University Rank 10	2014

**RESEARCH EXPERIENCE**

<b>Dept. of Biological Sciences, Bose Institute Research Fellow under Dr. Zhumur Ghosh</b>	Kolkata, India
<i>Genomic variants with Long Non-Coding RNA Loci: Role in Cancer</i>	2017-Present
<b>Biomedical genomic Centre, Kolkata under Prof. Nitai P. Bhattacharyaa</b>	Kolkata, India
<i>Molecular basis of interactions of organs/tissues based on protein present in the organs and their biological function</i>	Summer, 2015

## GRANTS AND REWARDS

---

### **Qualified Jadavpur university Ph.D course work examination**

Department of Science and Technology, Government of India

### **Senior Research Fellowship**

Council of Scientific and Industrial Research, Government of India

### **Qualified Bose Institute Ph.D course work examination**

### **Junior Research Fellowship**

Council of Scientific and Industrial Research, Government of India

### **Qualified National Eligibility Test (NET) for Research and Lecturership – Life Sciences All India Rank 47**

University Grants Commission, Government of India

### **Qualified National Eligibility Test (NET) for Research and Lecturership – Life Sciences All India Rank 72**

University Grants Commission, Government of India

### **Qualified Graduate Aptitude Test in Engineering (GATE) – Life Sciences**

**Eligible for DST Inspire fellowship, University 1<sup>st</sup> Rank holder in Microbiology**

## PAERTICIPATION IN CONFERENCES

---

### **Recent Trends in Natural Science 2023**

Bose institute

Kolkata, India

November 2023

### **Recent Trends in Natural Science 2022**

Bose institute

Kolkata, India

November 2022

### **2022 Cold Spring Harbor meeting: Regulatory & Non-coding RNAs**

Cold Spring Harbor Laboratory

New York, USA

May 2022

(Attended Online)

### **41st Annual Conference of IACR- An International symposium on: Cancer and Stem cells,**

Amity University

Noida, India

March, 2022

(Attended Online)

### **10th RNA Group Meet**

Rajiv Gandhi Centre of Biotechnology

Trivandrum, India

May, 2019

### **International Symposium on Systems, Synthetic & Chemical Biology**

Bose Institute

Kolkata, India

December, 2017

## PARTICIPATION IN WORKSHOPS

---

### **National workshop on Bioinformatics: AI in Healthcare**

Bose Institute

Kolkata, India

January, 2024

### **One Day National Workshop on Plant Bioinformatics**

Bose Institute

Kolkata, India

December, 2023

### **5 Day Intensive Hands-on Workshop: CRISPR-Cas9 based precise genome editing**

CSIR-IGIB

New Delhi, India

March, 2018



## CONFERENCE PRESENTATIONS

---

Das T., and Ghosh Z. Role of lncRNA-SNP in Cancer, **Recent Trends in Natural Sciences 2023**, Bose Institute, 26-29 November, 2023

Das T., Ghosh B., Das G. and Ghosh Z. Repurposed Drug candidates against COVID-19, **Recent Trends in Natural Sciences 2022**, Bose Institute, 26-29 November, 2022

Das T., and Ghosh Z. Genetic variants within lncRNA loci: Role in Female Cancers, **2022 Cold Spring Harbor meeting: Regulatory & Non-coding RNAs**, CSHL, USA, 17-20 May, 2022.

Das T., and Ghosh Z. Role of lncRNA-SNPs in female cancers, **41st Annual Conference of IACR- An International symposium on: Cancer and Stem cells**, Amity University, Noida, 2-5 March, 2022.

Das T., Parida S. and Ghosh Z. Genetic variants within lncRNA candidate loci: Their role in Cancer, **10th RNA Group Meet**, Rajiv Gandhi Centre of Biotechnology, Trivandrum, 2-4 May, 2019.

## PEER-REVIEWED PUBLICATIONS (sorted by date)

---

(# equal contribution)

1. Debsharma S, Pramanik S, Bindu S, Mazumder S, **Das T**, Pal U, Saha D, De R, Nag S, Banerjee C, Maiti N C, Ghosh Z, Bandyopadhyay U: *NSAID targets SIRT3 to trigger mitochondrial dysfunction and gastric cancer cell death*. iScience, 2024
2. Birari P, Mal Soumyav Majumder D, Sharma A K, Kumar M, **Das T**, Ghosh Z, Jana K, Gupta U D, Kundu M, Basu J: *Nur77 influences immunometabolism to regulate the release of proinflammatory cytokines and the formation of lipid bodies during Mycobacterium tuberculosis infection of macrophages*. Pathogens and Disease, 2023. 81
3. Ghosh B #, **Das T** #, Das G #, Chowdhury N #, Bagchi A, Ghosh Z: *Mapping Drug-gene Interactions to Identify Potential Drug Candidates Targeting Envelope Protein in SARS-CoV-2 Infection*. Current Bioinformatics 2023, 18(9):760-773.
4. Debsharma S, Pramanik S, Bindu S, Mazumder S, **Das T**, Saha D, De R, Nag S, Banerjee C, Siddiqui AA et al: *Honokiol, an inducer of sirtuin-3, protects against non-steroidal anti-inflammatory drug-induced gastric mucosal mitochondrial pathology, apoptosis and inflammatory tissue injury*. Br J Pharmacol 2023, 180(18):2317-2340.
5. Das G, **Das T**, Parida S, Ghosh Z. lncRTPred: *Predicting RNA–RNA mode of interaction mediated by lncRNA*. IUBMB Life. 2024; 76(1): 53–68.

6. Sarkar A #, **Das T** #, Das G #, Ghosh Z: *MicroRNA mediated gene regulatory circuits leads to machine learning based preliminary detection of acute myeloid leukemia*. Computational Biology and Chemistry 2023, 104:107859.
7. Acharya U, **Das T**, Ghosh Z, Ghosh A: *Defense Surveillance System at the Interface: Response of Rice Towards Rhizoctonia solani During Sheath Blight Infection*. Molecular Plant-Microbe Interactions® 2022, 35(12):1081-1095.
8. Debnath S, Sarkar A, Mukherjee DD, Ray S, Mahata B, Mahata T, Parida PK, **Das T**, Mukhopadhyay R, Ghosh Z et al: *Eriodictyol mediated selective targeting of the TNFR1/FADD/TRADD axis in cancer cells induce apoptosis and inhibit tumor progression and metastasis*. Translational Oncology 2022, 21:101433.
9. Ganguly P, Roy D, **Das T**, Kundu A, Cartieaux F, Ghosh Z, DasGupta M: *The Natural Antisense Transcript DONE40 Derived from the lncRNA ENOD40 Locus Interacts with SET Domain Protein ASHR3 During Inception of Symbiosis in Arachis hypogaea*. Molecular Plant-Microbe Interactions® 2021, 34(9):1057-1070.
10. Das G #, **Das T** #, Chowdhury N #, Chatterjee D #, Bagchi A, Ghosh Z: *Repurposed drugs and nutraceuticals targeting envelope protein: A possible therapeutic strategy against COVID-19*. Genomics 2021, 113(1, Part 2):1129-1140.
11. Mondal S, Bhattacharya N, **Das T**, Ghosh Z, Khatua S: *Implication of Statistical Methods on Patient Data: An Approach for Cancer Survivability Prediction*. In: Intelligent Healthcare: Applications of AI in eHealth. edn. Edited by Bhatia S, Dubey AK, Chhikara R, Chaudhary P, Kumar A. Cham: Springer International Publishing; 2021: 57-80.
12. **Das T**, Deb A, Parida S, Mondal S, Khatua S, Ghosh Z: *LncRBase V.2: an updated resource for multispecies lncRNAs and ClinicLSNP hosting genetic variants in lncRNAs for cancer patients*. RNA biology 2021, 18(8):1136-1151.

RESEARCH PAPER



## LncRBase V.2: an updated resource for multispecies lncRNAs and ClinicLSNP hosting genetic variants in lncRNAs for cancer patients

Troyee Das<sup>a</sup>, Aritra Deb<sup>a</sup>, Sibun Parida<sup>a</sup>, Sudip Mondal<sup>b</sup>, Sunirmal Khatua<sup>b</sup>, and Zhumur Ghosh<sup>a</sup>

<sup>a</sup>Division of Bioinformatics, Bose Institute, Kolkata, India; <sup>b</sup>Department of Computer Science and Engineering, University of Calcutta, Kolkata, India

### ABSTRACT

The recent discovery of long non-coding RNA as a regulatory molecule in the cellular system has altered the concept of the functional aptitude of the genome. Since our publication of the first version of LncRBase in 2014, there has been an enormous increase in the number of annotated lncRNAs of multiple species other than Human and Mouse. LncRBase V.2 hosts information of 549,648 lncRNAs corresponding to six additional species besides Human and Mouse, viz. Rat, Fruitfly, Zebrafish, Chicken, Cow and *C. elegans*. It provides additional distinct features such as (i) Transcription Factor Binding Site (TFBS) in the lncRNA promoter region, (ii) sub-cellular localization pattern of lncRNAs (iii) lnc-pri-miRNAs (iv) Possible small open reading frames (sORFs) within lncRNA. (v) Manually curated information of interacting target molecules and disease association of lncRNA genes (vi) Distribution of lncRNAs across multiple tissues of all species. Moreover, we have hosted ClinicLSNP within LncRBase V.2. ClinicLSNP has a comprehensive catalogue of lncRNA variants present within breast, ovarian, and cervical cancer inferred from 561 RNA-Seq data corresponding to these cancers. Further, we have checked whether these lncRNA variants overlap with (i) Repeat elements, (ii) CGI, (iii) TFBS within lncRNA loci (iv) SNP localization in trait-associated Linkage Disequilibrium (LD) region, (v) predicted the potentially pathogenic variants and (vi) effect of SNP on lncRNA secondary structure. Overall, LncRBase V.2 is a user-friendly database to survey, search and retrieve information about multi-species lncRNAs. Further, ClinicLSNP will serve as a useful resource for cancer specific lncRNA variants and their related information. The database is freely accessible and available at <http://dibresources.jcbiose.ac.in/zhumur/LncRbase2/>.

### ARTICLE HISTORY

Received 10 June 2020  
Revised 8 September 2020  
Accepted 4 October 2020

### KEYWORDS

Long non-coding RNA;  
lncRNA variant; clinicLSNP;  
female cancer; sORFs; sub-  
cellular localization; lncRbase

### Introduction

Long non-coding RNAs (lncRNAs) constitute a group of non-coding RNAs (ncRNAs) which are associated with diverse functions in the cellular system [1,2]. Having length  $\geq 200$ nts, they often show similarity to that of messenger RNAs (mRNAs) with respect to their mode of biogenesis, having a 5' cap, polyA tail, multiple exons, but lacking the coding capability [3]. Their recent discovery as a regulatory molecule in the cellular system has altered the concept of the functional aptitude of the genome, a major portion of which was once considered to be junk. The diverse role of lncRNA in the regulation of transcription, inactivation of X chromosome, genomic reprogramming, miRNA scavenging, nuclear-cytoplasmic trafficking, RNA-splicing, and apoptosis [4] has been experimentally proved. But this represents the tip of an iceberg. Owing to the advancement of next-generation sequencing technology, the number of identified lncRNAs are multiplying rapidly. Validation by standard wet bench techniques reinforces the claim of these *in silico* findings. Existing databases like Ensembl [5], NONCODE [6], Refseq [7], LNCipedia [8] have provided information on lncRNAs for multiple species. Ensembl has genebuild pipeline to classify their lncRNA transcripts into several biotypes.

NONCODE enlisted the length, sequence, coding probability score, expression profile of Human and Mouse lncRNAs, and predicted the secondary structure of Human lncRNAs. LNCipedia [8] records lncRNA isoform names, location, coding probability score and locus conservation of lncRNA transcripts. There are other dedicated databases hosting detailed information on discrete features of lncRNAs. CHIPBase v2.0 [9] categorized regulatory molecules in the promoter region of Ensembl enlisted lncRNA transcripts only. Databases like Lnc2Cancer [10], LncRNADisease2 [11], Lnc2Catlas [12] catalogues experimentally validated lncRNA disease relationships.

Despite such extensive work on lncRNAs, there remained certain domains such as their association with other ncRNA molecules or overlap with genomic elements which could serve as inherent regulatory sites. Our idea to publish the first version of LncRBase [13] was to probe deeper into these lesser-explored domains of lncRNA regulome. This time we aim to delve further to study their upstream regulators by predicting regulatory binding sites in the lncRNA promoter region, which could unveil much about its interacting partners. We have provided the probability of the presence of these lncRNAs within different cellular compartments. With recent evidence revealing the

dependence of lncRNA function on its sub-cellular localization, [14], prognosticating sub-cellular localization could provide meaningful insights into their potential functions. The ribosome footprinting assays indicating the micro peptide coding capability of some of these lncRNAs along with their coding probability prediction *in silico* could serve as a better way to designate them as putatively coding or noncoding. Besides, a readily available exhaustive catalogue of tissue-specific lncRNA expression of multiple species will serve as another useful resource.

This new LncRBase V.2 is an expanded version of LncRBase and includes several distinct features (as provided in Table 1) which are as follows: (i) Our analysis has been expanded to 6 new species beyond Human and Mouse. (ii) We have elucidated their regulatory mechanism by predicting Transcription Factor Binding Site (TFBS) in the lncRNA promoter region, which could unveil much about its upstream regulators. (iii) Evidence regarding the influence of events like sub-cellular localization pattern of lncRNAs on their functionality [15] has instigated us to predict the localization of a lncRNA which dictates the activating and deactivating functions of lncRNAs. (iv) Ribosome footprinting assays indicate the micro peptide coding capability of some of the noncoding molecules [16]. Hence we have mapped the possible small open reading frames (sORFs) within lncRNA loci from footprinting studies which is a better way to confirm the status of the molecule as putatively coding or noncoding [17]. (v) We have also provided manually curated information of interacting target molecules and disease involvement of lncRNA genes (vi) Besides, LncRBase V.2 is a readily available source of tissue-specific expression of lncRNAs corresponding to multiple species.

Single Nucleotide Polymorphisms (SNPs) represent the most frequent genetic changes in the Human genome, and SNPs occurring within functional regions has a possibility to be associated with phenotypic changes and disease susceptibility [18] like cancer. Literature evidence confirms the connection of

cancer risk to SNP containing lncRNAs too [19]. These cancers include the most prevalent female cancers worldwide like breast, cervical and ovarian cancer [20–22]. Apart from such experimental approaches to identify and analyse SNP-associated lncRNAs and their functional implications in different cancer systems, dedicated groups have catalogued databases of SNP association with lncRNA loci and analyzed their functional consequences in different biological systems. At present, there are several online databases hosting information on the genomic variants within lncRNA genes, such as LincSNP 2.0 [23] and lncRNASNP2 [24]. LincSNP 2.0 [23] distinguished disease-associated SNPs in lncRNAs and their TFBS. lncRNASNP2 [24] hosts information on disease-associated GWAS and COSMIC variants in human and mouse lncRNAs and predicts the effect of variants on loss/gain of miRNA-lncRNA interactions. In these cases, the available disease-associated SNPs have been mapped to lncRNA loci. However, there is no such database where variants mapped to lncRNA loci are analyzed directly from raw clinical data. We have developed ClinicLSNP (which is being hosted here as a part of LncRBase V.2), which has a comprehensive collection of lncRNA variants inferred from 561 female cancer-specific RNA-Seq data which includes breast, ovarian, and cervical cancer.

Overall, LncRBase V.2 is a user-friendly database to survey, search and retrieve information on multi-species lncRNAs. ClinicLSNP (hosted within LncRBase V.2) will serve as a useful resource for both researchers and clinicians to retrieve information on breast, cervical and ovarian cancer specific lncRNA-SNPs and their related information. The database is freely accessible and available at <http://dibroseources.jcbose.ac.in/zhumur/lncrbase2/>

## Results and discussion

LncRBase V.2 (the updated version of LncRBase) hosts information on 549,368 lncRNAs corresponding to eight species viz. Human, Mouse, Rat, Fruitfly, Cow, Chicken, Zebrafish and *C.elegans*. This new version has several added features (apart from the coverage of species) which add on to the importance of this database.

### Distribution of lncRNA subtypes based on their genomic location

LncRBase V.2 hosts species-specific classification of lncRNA transcripts into sixteen distinct, both newly classified as well as already existing biotypes. Human and Mouse lncRNAs show the distribution across all the seventeen biotypes while that for species like Cow and *C. elegans*, lncRNAs show the distribution across ten biotypes. These are shown below along with the nomenclatures used for them in Table 2:

In LncRBase V.2, certain new biotypes have been added (marked in italics) and some changes have been incorporated in the abbreviation of existing biotypes that is present in LncRBase to make it more user-friendly (marked in bold). Biotype Ambiguous ORF (abbreviated as AG) present in the previous version has been removed in this new update as it is no longer supported by Ensembl 90 which has been used in this work.

**Table 1.** Comparison between the contents of LncRBase V.2 and LncRBase

Database content	LncRBase	LncRBase V.2
Number of Species	2 (Human, Mouse)	8 (Human, Mouse, Fly, Zebrafish, Rat, Chicken, Cow and <i>C.elegans</i> )
Transcript entries	216,562	549,368
Coding potential Score	CPAT	CPAT1.2.3 (Human, Mouse, Fly, Zebrafish), CPC2 (all species) and PLEK (all species)
sORF within lncRNA loci	No	Yes (Human, Mouse, Fly, Zebrafish, Rat, <i>C.elegans</i> )
Association with Repeat elements	Yes	Yes (all species)
Co localized miRNAs	Yes	Yes (all species)
lnc-pri-miRNAs	No	Yes (all species)
Co localized piRNAs	Yes	Yes (Human, Mouse, Fly, Rat, <i>C.elegans</i> )
Predicted Sub cellular Localization	No	Yes (all species)
CGI in lncRNA promoter	Yes	Yes (all species)
TFBS within lncRNA promoter	No	Yes (Human)
Tissue specific expression (multiple datasets, all species)	Yes	dataset)
lncRNA disease association (literature curated)	No	Yes (Human)
lncRNA target genes (literature curated)	No	Yes (Human)

**Table 2.** Nomenclature for biotype wise distribution of Long non-coding RNAs. New biotypes introduced in this version of the database are italicized. Changed abbreviation of existing biotypes that is present in LncRBase are marked in bold.

No. of biotypes	Genomic location of the lncRNA	Name of the biotype	Nomenclature for the biotype specific lncRNA
1.	<b>LncRNAs overlapping with 3'UTR exonic region in the sense strand.</b>	<b>3'UTR overlapping lncRNA</b>	<b>3UTR</b>
2.	<b>LncRNAs overlapping any 5'UTR exon in the sense strand</b>	<b>5'UTR overlapping lncRNA</b>	<b>5UTR</b>
3.	LncRNAs overlapping any CDS exon	CDS overlapping lncRNA	CDS
4.	Intergenic (linc) lncRNAs transcribed from in between two gene loci	Intergenic (linc) lncRNA	LI
5.	<b>LncRNAs intersecting any exon of a protein-coding locus on the opposite strand</b>	<b>Antisense Overlapping lncRNA</b>	<b>AO</b>
6.	LncRNAs residing within introns of a coding gene, but do not intersect any exons	Completely Intronic lncRNA	CI
7.	<b>Antisense lncRNAs completely overlapping with an intron in the opposite strand</b>	<b>Antisense Intronic lncRNA</b>	<b>AI</b>
8.	Intron Overlapping lncRNA splice variants of a gene, contain intronic sequence	Intron Overlapping lncRNA	IO
9.	Pseudogene transcripts having homology to protein coding transcripts but containing disrupted coding sequence and an active homologous gene can be found at another locus	Pseudogene	PS
10.	Sense Overlapping lncRNAs containing a coding gene in its intron on the same strand(Ensembl annotated)	Sense Overlapping lncRNA	SO
11.	Processed Transcripts not containing an ORF (Ensembl annotated)	Processed Transcripts	PT
12.	Miscellaneous RNA(miscRNA) from the Ensembl transcript dataset(Ensembl annotated)	miscRNA	MI
13.	<i>A non-coding locus that originates from within the promoter region of a protein-coding gene, with transcription proceeding in the opposite direction on the other strand (Ensembl annotated)</i>	<i>Bidirectional promoter lncRNA</i>	<i>BI</i>
14.	<i>lncRNAs present in mitochondria</i>	<i>Mitochondrial lncRNA</i>	<i>Mito_M</i>
15.	<i>Unspliced lncRNAs that are several kb in size (Ensembl annotated)</i>	<i>Macro lncRNA</i>	<i>Macro</i>
16.	<i>lncRNAs falling under more than one biotype</i>	<i>Ambiguous lncRNA</i>	<i>AG</i>
17.	Non coding transcripts not falling in any of the above mentioned categories	Non coding	NC

Fig. 1 depicts the biotype wise distribution of lncRNAs for all species across their genome. It reveals the abundance of intergenic lncRNAs within mammalian systems as well as in Chicken and Zebrafish. For *Drosophila*, the majority of the lncRNAs are overlapping with more than one genomic location and are hence classified as Ambiguous (AG) type. Further, most of the lncRNAs in *C.elegans* are classified as Noncoding (NC) type since they could not be classified to specific genomic loci based on existing genomic annotation.

### Distribution of repeat elements within lncRNA loci

Similar to the first version of LncRBase, repeat elements belonging to different repeat classes have been mapped to lncRNA loci for all species. The distribution of 8 predominant repeat classes viz. SINE, LINE, DNA, Simple repeat, low complexity, LTR, Satellite and SVA Retroposon (Human-specific) across different lncRNA biotypes have been represented in Supplementary Fig. S1. Fruitfly, Zebrafish and Mammalian lncRNAs are found to be enriched with low complexity repeats. The physiological importance of low complexity sequences within lncRNAs can be found in the literature. For eg, Low complexity CU/AG repeats within ribosomal intergenic noncoding RNA(rIGSRNA) facilitates the formation of amyloid bodies [25]. *C.elegans* lncRNAs and Chicken lncRNAs are found to be enriched with Simple repeat and LINE repeat elements.

### The abundance of piwil interacting RNAs (piRNAs) within lncRNA loci

Genomic location of Human, Mouse, Rat, *Drosophila* and *C. elegans* piRNAs that are available in NCBI, have been mapped to the respective lncRNA loci. 55.7% of Human piRNAs, 51.5% of Mouse piRNAs, 4.5 % of Rat piRNAs, 11.5% of

Fruitfly piRNAs and 31.4% of *C.elegans* piRNAs have been found to overlap with their respective lncRNA loci. The distribution of piRNAs within the loci of different biotype specific lncRNAs for these five species is provided in Supplementary Fig. S2. Interestingly, a considerable percentage of Human and Mouse piRNAs overlap within the newly identified lncRNA Mito\_M biotype, although a very small percentage of Human and Mouse lncRNAs belongs to this lncRNA biotype(as shown in Fig. 1).

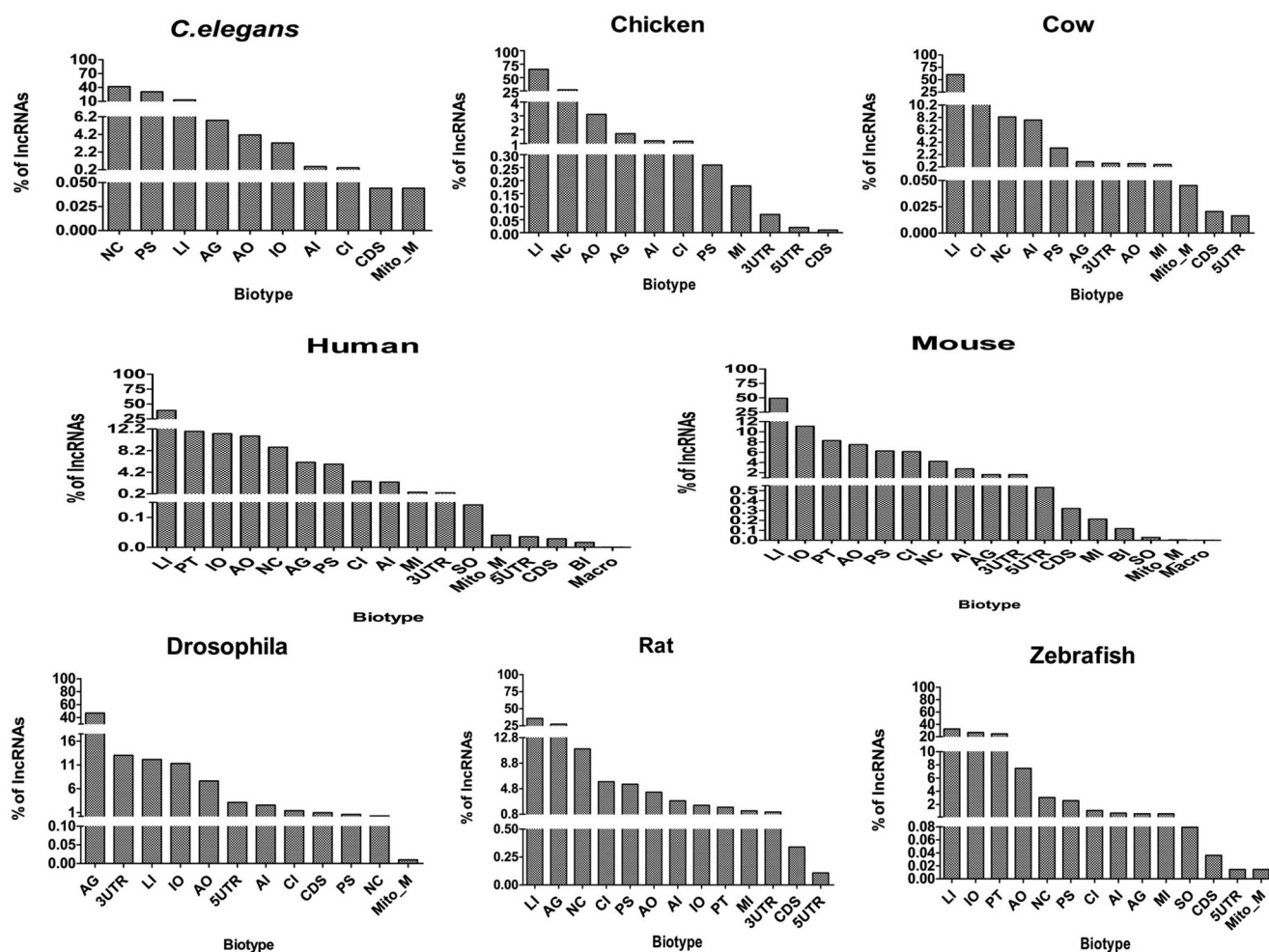
### CGI association with the lncRNA promoter region

We have checked the distribution of CpG island (CGI) within the lncRNA promoter region spanning 1000 bases upstream and downstream of Transcription Start Site (TSS) as represented in Supplementary Fig. S3. The overall scenario indicates only a minority of lncRNA promoters to be CGI associated (as low as 2% for Cow and highest for Fruitfly~23%). However, among the various biotypes, promoters of 5UTR lncRNA transcripts are mostly observed to be CGI associated with high CpG content. Association of CGI along with tissue specificity of lncRNAs has also been seen to have a major impact on various diseases. lncRNAs like MALAT1 and NEAT1 having an established role in multiple disorders and carcinoma of different tissues are found to be having CGI associated promoters shown in our analysis. In LncRBase V.2, we have explored the relationship between CGI association and lncRNA tissue specificity which we have discussed in the next section.

### Tissue-specific distribution of lncRNAs

lncRNAs are mostly known to be tissue-specific [26,27]. Further, a significant portion of the tissue-specific gene promoters is not CGI associated [28]. This goes on par with the





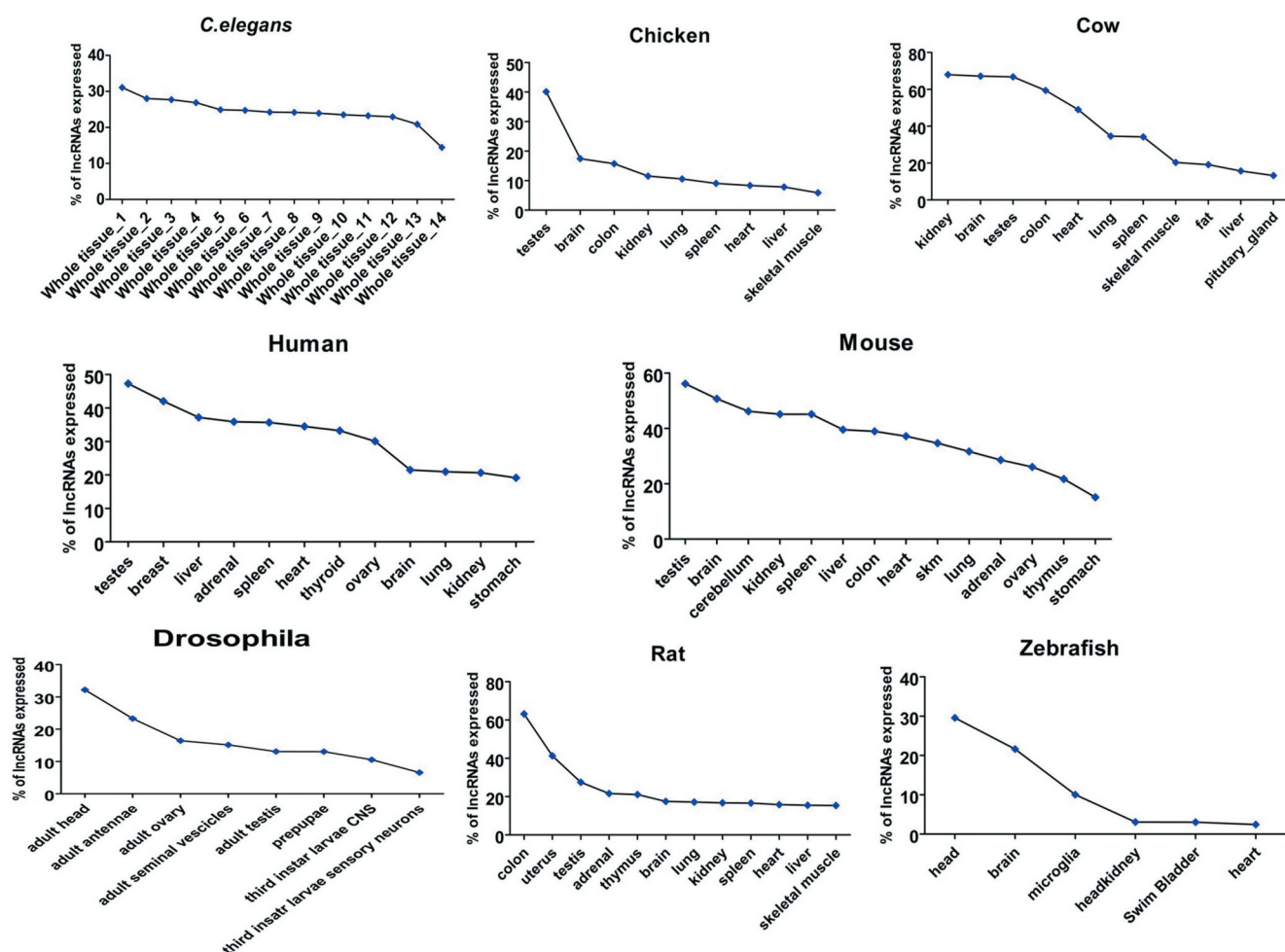
**Figure 1.** Biotypes specific distribution of lncRNAs.

Abbreviations used: NC Non Coding, MI MiscRNA, SO Sense Overlapping, PS Pseudogenes, CI Completely Intronic, 3UTR 3'/UTR overlapping, PT Processed Transcript, 5UTR 5'/UTR overlapping, CDS CDS overlapping, IO Intron Overlapping, LI Long Intergenic, AO Antisense Overlapping, AI Intronic Antisense, AG Ambiguous, BI Bidirectional lncRNAs, Mito\_M Mitochondrial lncRNAs, Macro Macro lncRNAs.

result of our CpG island analysis where it has been observed for the majority of lncRNA gene promoters which do not lie within CGI regions (as discussed in Section 1). Hence, a clear picture of lncRNA expression across different tissue types will provide information regarding the tissue-specific functionality of this molecule. We have analyzed fourteen whole tissue RNA-seq data for *C.elegans*. For Fruitfly, tissue data for all three developmental stages. i.e. Prepupae, Third Instar Larvae and Adult have been analyzed. For Zebrafish, we have focussed on RNA-seq data of head, microglia, swim bladder, head kidney and heart. Besides, tissues like the brain, colon, kidney, spleen, lungs, skeletal muscle, testes and others are considered for the four mammalian species and chicken. Our analysis revealed that lncRNA expression is predominant in the testis of Human, Mouse and Chicken whereas its expression in the head is predominant in *Drosophila* and Zebrafish. In Cow and Rat, the predominance is observed in the kidney and colon respectively. Tissue-specific lncRNA expression profiles for the eight species are shown in Fig. 2.

### Bifunctional lncRNAs

lncRNAs have been reported to perform dual functions-both as protein-coding as well as regulatory ncRNAs. The first reported role came from the study of *Drosophila* where four micro peptides being involved in embryonic development and alternating the fate of TF Svb (shavenbaby) are coded by lncRNA tal (tarsal less) [29]. Subsequent evidence of lncRNA encoded micro peptide has also been reported in Human, Mouse, Chicken, Zebrafish, Nematode and other species as well [30–33]. This raises the question of the true non-coding identity of the lncRNAs or whether they are performing the dual role as bifunctional lncRNAs [34,35]. Small Open Reading Frames (sORFs) which were initially regarded as 'background noise' of proteomics experiments have now been proved to have the potential for coding clinically significant micro peptides [33]. sORFs are also known for sequence conservation [36]. Ribosome footprinting techniques (which distinguish between translated and untranslated ORFs), coupled with stringent computational cut-off



**Figure 2.** Distribution of lncRNAs across different tissues corresponding to the eight species.

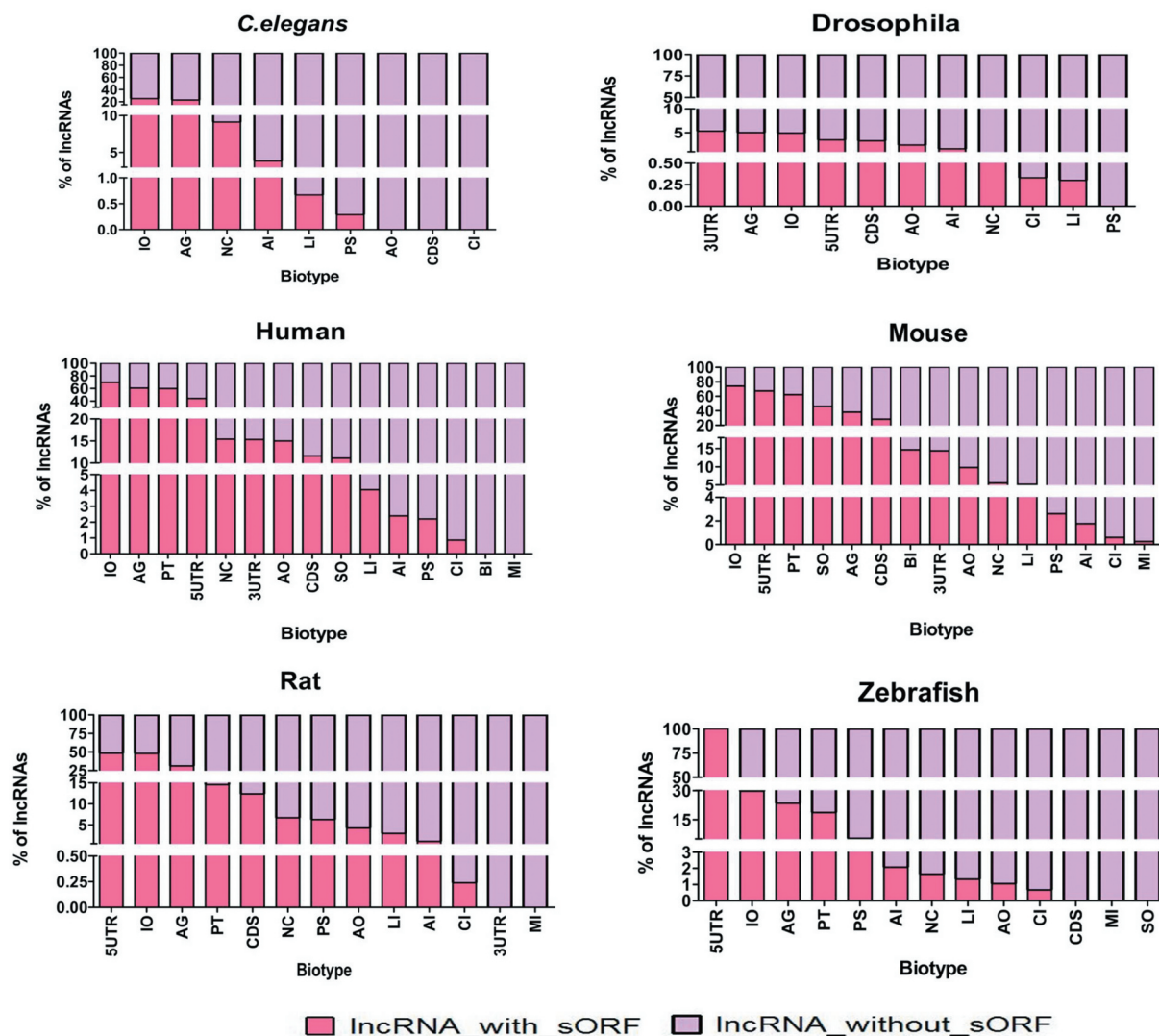
criterion [16] made it possible to identify several thousand sORFs, translated from non-coding molecules of the genome. Many of their coding capacity has already been validated [16].

In addition to categorizing lncRNAs into noncoding and putatively coding type by Coding Potential Score prediction tools [37–39], we tried to identify bi-functional lncRNAs by mapping sORFs within lncRNA loci. sORF information for six species (Human, Mouse, Rat, Fruitfly, Zebrafish, and *C. elegans*) have been retrieved from [www.sorfs.org](http://www.sorfs.org) [17]. The biotype wise distribution of sORF containing lncRNAs has been represented in Fig. 3. sORFs have been mapped mostly with IO for Human and Mouse. In *C. elegans* predominance of sORFs has been seen in IO and AG biotypes. In Rat, both IO and 5UTR biotypes harbor maximum sORFs whereas 5UTR biotypes are highly enriched with sORFs in Zebrafish. In *Drosophila*, all the three biotypes IO, 5UTR and AG biotypes are highly enriched with sORFs. Surprisingly, a minor portion of sORFs got mapped within intergenic lncRNAs(LI) (highest for Human and Mouse, 4.05% and 5.17% respectively), which lies between two coding regions. Since much of the annotated lncRNAs have not yet been experimentally validated, *insilico* prediction of their coding potential clubbed with footprinting information is a convincing method to identify bi-functional lncRNAs.

### ***lncRNAs harbouring embedded miRNAs***

A fair amount of miRNAs originate from the exonic region of lncRNAs. A noted example is the case of lncRNA H19 derived miRNA mir-675 with a co-dependent role in gastric cancer [40]. Some of their biogenesis is found to be different than that of miRNAs originating from the exonic region of protein-coding genes [41]. Unlike host protein-coding genes, where mRNA and miRNA are co-transcribed, some cellular system adopts the microprocessor-mediated mechanism of transcriptional termination to stop the transcription of the lncRNA gene whereby prioritizing the transcription of miRNA [42]. These are embedded miRNAs termed as lnc-pri-miRNAs, and they deviate from the canonical pathway of Pol II-associated cleavage and polyadenylation (CPA). Biogenesis of miRNA mir-122, involved in the metabolism of cholesterol and Hepatitis C virus replication results in terminated transcription of its host lncRNA. [43]. It is yet to be determined what causes such altered fate where either the host lncRNA or its embedded miRNA can exist in the system at a time.

In our previous version of lncRBase, we focused only on miRNAs that are coming from the same loci as that of the lncRNAs which has been named as co-localized miRNAs. This time, in addition to such information as represented in Supplementary Fig. S4, we have identified several lnc-pri-

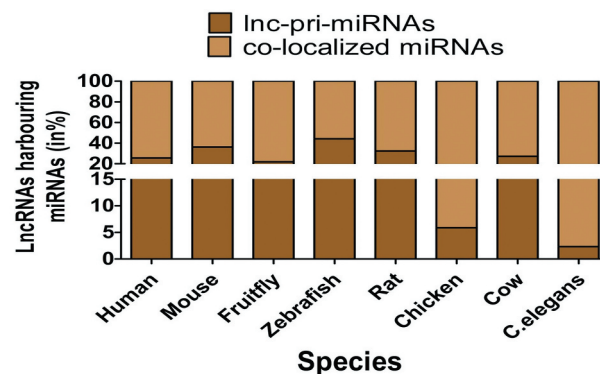


**Figure 3.** Biotype wise distribution of lncRNAs harbouring small Open Reading Frames (sORFs).

miRNAs i.e. miRNAs embedded within lncRNAs corresponding to all species as in Fig. 4. The criterion for determining lnc-pri-miRNAs has been elaborated in the materials and methods section. Zebrafish has the highest percentage (44.2%) and *C. elegans* has the lowest percentage (2.32%) of lnc-pri-miRNAs. Investigating their canonical or noncanonical mode of transcription could provide insight into their regulatory fate.

### Sub-cellular localization of lncRNAs

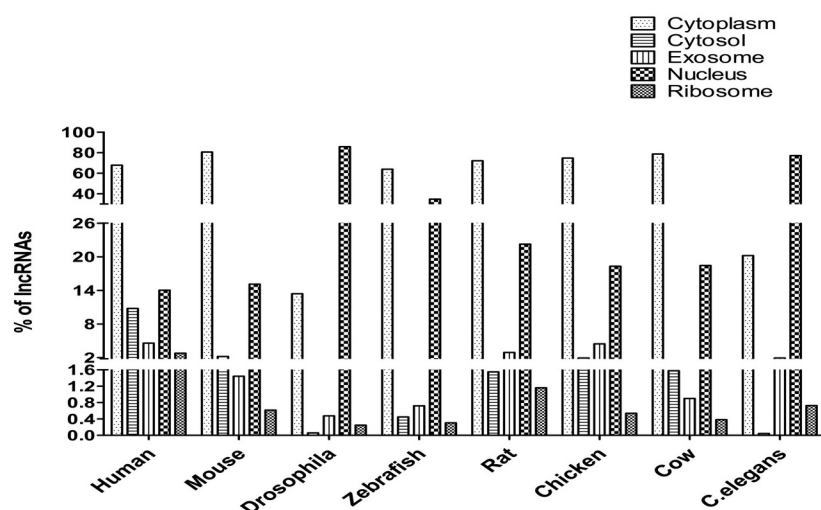
lncRNAs are known to exhibit a variable pattern of sub-cellular localization (either within the nucleus or cytoplasm or non-specifically in both) as revealed by RNA fluorescence in situ hybridization analysis [44]. The localization is intricately interlinked with its function [45,46]. However, the localization is still unknown for a vast majority of lncRNAs. We have provided the information regarding the possible localization of the lncRNA transcripts as represented in Fig. 5. Cytoplasm enrichment has been predicted for most species followed by nucleus and cytosol, the exception being



**Figure 4.** Distribution of lncRNAs harbouring lnc-pri-miRNAs and co-localized miRNAs across eight species.

Co-localized miRNAs are those which come from lncRNA loci lnc-pri-miRNAs are those which come from lncRNA loci as well as it has exact sequence match with that of the mapped region within lncRNA loci. So, the percentage of lnc-pri-miRNAs is calculated based on the total number of miRNAs coming from lncRNA loci.





**Figure 5.** Sub-cellular localization of lncRNAs for the eight species

*Drosophila* and *C.elegans* where more than 70% nucleus enrichment pattern is observed. Very recently, lncRNAs are also identified in exosomes, the smallest type of Extracellular Vesicles (EV) [47]. Exosomes itself are known for intricate cell-to-cell communication and exosomal lncRNA LINC00152 is linked to gastric cancer [48]. We have observed near about 5% exosome coverage for Human, Rat and Chicken lncRNAs. The ribosome has been observed to harbour the least no of lncRNAs in all the species. Prior knowledge of possible sub-cellular preference of a lncRNA transcript could serve as a starting point to identify its possible interacting partners.

### Transcription Factor binding site in lncRNA promoter regions

We have scanned the lncRNA promoter regions to identify the upstream regulators of the lncRNA genes. We have used the analyzed Chip-Seq data which revealed thousands of binding peaks throughout the genome in multiple tissues and under various disease conditions [49]. Subsequently, we have mapped Transcription factors (TFs) spanning 1000kb up and downstream of more than 71% lncRNA Transcription Start Site (TSS). Information on TF binding to the lncRNA promoter region could help us to understand the upstream transcriptional regulators that control the tissue-specific expression of lncRNAs, their mode of regulation, and even the disease association of many poorly annotated lncRNAs.

### lncRNA target partners and disease association

Extensive literature curation has been done to provide information on interacting partners of 408 Human and 41 Mouse lncRNA genes. Information includes the name of the targeting molecule, mode of regulation and corresponding PubMed reference. Information of 1129 lncRNA genes associated with diseases, their mode of regulation, nature of the experiment along with their PubMed IDs has also been provided. This will serve as a ready reference to the users to look into the published literature

resources while searching for an experimentally validated lncRNA target information and associated diseases.

### ClinicLSNP

ClinicLSNP hosts lncRNA-SNP information for the prevalent female cancer systems, viz. breast, ovarian and cervical cancer. We have analyzed 561 RNA seq data, which includes 280 tissue and 281 cell line samples of Human breast, ovarian, and cervical cancer to perform the transcriptome-wide variant mapping within lncRNA loci in these three female cancer systems. The results are then depleted from 88, 32 and 5 normal breast, ovary and cervix tissue-derived variants respectively to keep only these 3 cancer-specific exclusive variant information. Subsequently, these variants are matched with existing dbSNP entries [50] and have been categorized as novel and annotated (those which got mapped with the existing dbSNP entries) variants. Overall, ClinicLSNP hosts information about lncRNA-SNPs present within the 3 prevalent female cancer systems. We have come up with a total of 5,71,886 annotated and 2,43,254 novel unique variants mapping to 1,72,404 lncRNA transcripts for the three cancer types.

Further, it also hosts information regarding the overlap of these lncRNA-SNPs with (i) Repeat elements, (ii) CGI, (iii) TFBS within lncRNA loci (iv) SNP localization in trait-associated LD (Linkage Disequilibrium) region, (v) predicted potentially pathogenic variants and (vi) SNP effect on lncRNA secondary structure.

**(A) Mapping of detected SNPs within lncRNA:** The distribution of SNP-associated lncRNAs across the different Human lncRNA biotypes for both cell line and tissue datasets of the breast, cervical and ovarian cancer has been represented in Fig. 5. Despite the fact that only 0.3% and 6% of Human lncRNAs belongs to 3'UTR and AG biotype respectively (as shown in Fig. 1), they harbour the major number of SNPs (more than 60%) in both cell line and tissue datasets of breast, cervical and ovarian cancer. On the contrary, lincRNAs being the most abundant

Human lncRNA biotype (~40%), relatively few of them harbour SNPs (~30%). Such biotype specific pattern of SNP distribution across lncRNA loci tempts us to speculate its association with biotype specific functional role of lncRNAs.

**(B) Clinical relevance of the lncRNA-SNPs in the three cancer systems:** We found clinical relevance of the lncRNA-SNPs present in the three cancer systems, viz. breast, cervical and ovarian cancer system. For the dbSNP annotated lncRNA-SNPs, we have mapped their rs IDs with those present in ClinVar database [51] and SNPsnap's SNP Annotation Database [52]. In addition, the CADD tool [53] has been used to predict the clinical relevance of the novel as well as annotated lncRNA variants. Results have been represented in Supplementary Fig. S5.

(i) Mapping with ClinVar IDs: ClinVar database [51] holds information on clinically significant variants where the variants are classified into 13 types based on 'clinical significance'. In order to look for the clinical significance of the dbSNP annotated lncRNA variants obtained in our analysis, we have mapped them with ClinVar IDs. For the breast, ovarian and cervical cancer cell lines, 2.5%, 2.3% and 3 % of lncRNA-SNPs respectively got mapped with ClinVar IDs. For tissue datasets, the mapping pattern of breast, ovarian and cervical cancer lncRNA-SNPs with ClinVar IDs are 2.1%, 2.6% and 5.4% respectively (Supplementary Fig. S5).

(ii) Mapping TAG SNPs: A group of SNPs or haplotypes inherited together in a genomic region of high linkage disequilibrium are often associated with a disease or phenotypic traits [54]. One such representative SNP from the cluster is helpful for risk association studies without the need to genotype every single variant. 57.9%, 58.2% and 76.6% of our dbSNP annotated lncRNA variants have been found to match with as TAG SNPs for breast, ovary, and cervical cancer cell lines and tissues respectively.

As observed from the graph in Supplementary Fig. S5, more than 60% of dbSNP annotated variants are found to be trait-associated in all the three disease systems, of which lncRNA variants for cervical cancer (corresponding to both cell line and tissue dataset) shows the highest enrichment.

(iii) Identification of potentially pathogenic variants: The deleterious score for all the variants has been predicted with CADD v1.5 [53], a tool that incorporates more than 60 genomic annotations to score SNPs and indels anywhere in the genome assembly. The suggested cut off as referred by Rentzsch et al., 2019 [53] for identifying potential pathogenic variants is set between 10 and 15 with scaled 'C score' >10 indicating top 10% deleterious variants, 'C score' >20 showing the top 1% deleterious variants and so on. Overall, 13.9%, 9.2% and 11.7% of our analyzed breast, cervical and ovarian cancer variants respectively fall within the risk range of 15 to 60. Detailed statistics of cell line and tissue-specific enrichment of annotated and novel SNPs have been provided in Supplementary Fig. S5.

**(C) SNP associated with Repeat Elements:** Initially considered as 'junk DNA', Repeat elements comprising 50% of the genome consists of interspersed repeat sequences mainly. These are known for their regulatory and structural roles [55]. Further, a study on Alu repeat elements by Payer et al., observed 44 such Alu insertion polymorphism to be in strong

Linkage Disequilibrium ( $r^2 > 0.7$ ) with trait-associated SNPs [56]. Moreover, lncRNAs are known to regulate Staufen mediated mRNA decay by forming an imperfect base pairing with 3'UTR Alu repeat elements [57].

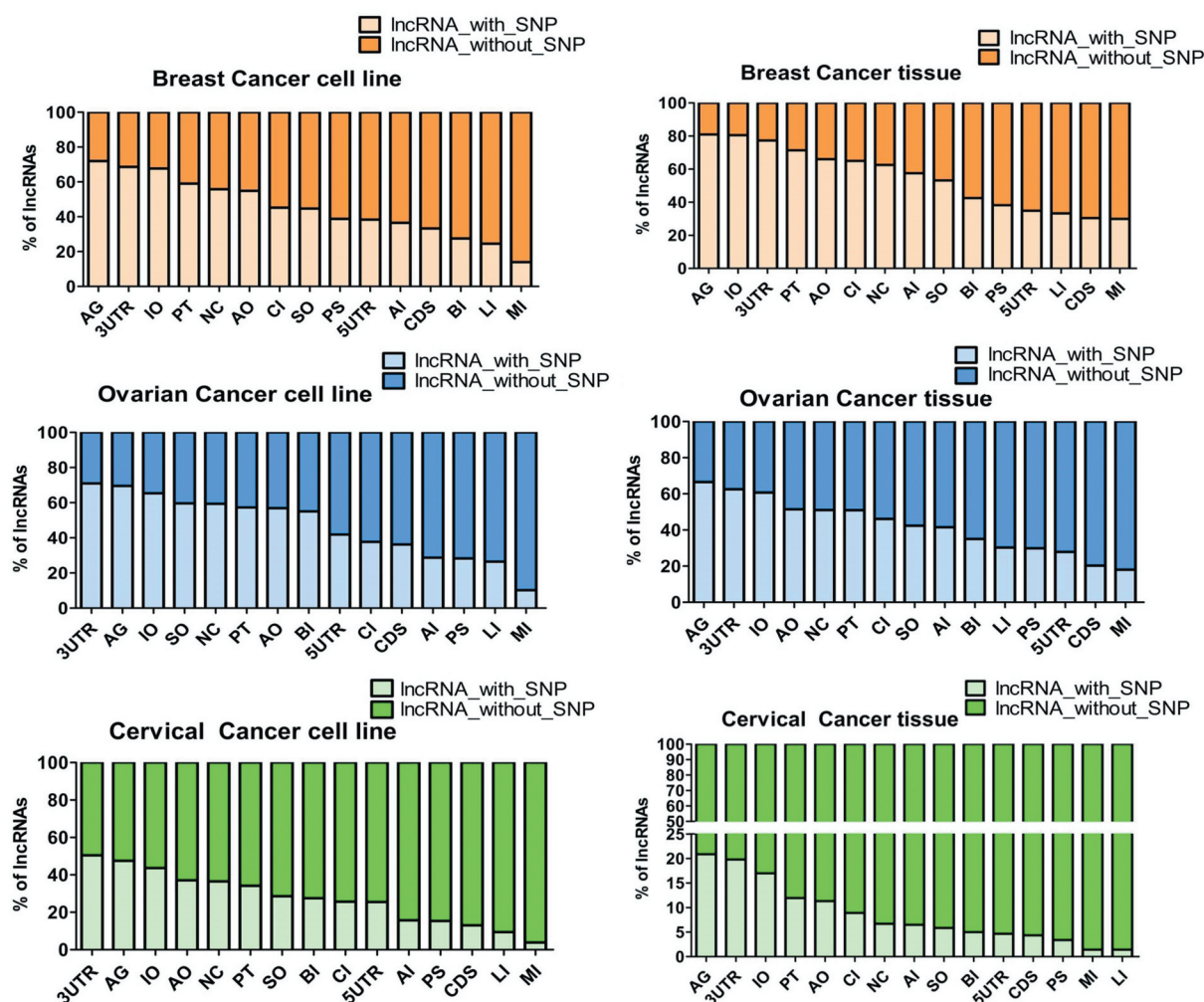
Our analysis shows the percentage of SNPs associated with repeat elements within lncRNA loci (Supplementary Fig. S6) as well as the distribution of repeat class (in percentage) within SNPs present in the lncRNA loci in the three cancer systems (Fig. 6). lncRNA-SNPs have been observed to be mostly enriched with SVA Retroposon and Satellite class repeats for these three cancers. SINE repeat elements come next followed by others. Thus the enrichment of a lncRNA-SNP with a repeat element has the potential to alter the function of that lncRNA and needs to be checked further.

**(D) Regulatory SNPs within lncRNA loci:** lncRNAs are well known for their role in transcriptional regulation of a nearby or distant gene by hiring chromatin modifiers or other modulating complexes [58]. Loss or gain in the binding sites due to the presence of regulatory SNPs (rSNPs) result in perturbed interaction and hinder downstream effect and hence such information regarding rSNPs within lncRNA loci could provide us more information regarding their altered function in transcriptional regulation. Literature evidence of such altered function in transcriptional regulation due to the presence of variants are as follows:

Buroker reviewed a few disease developments associated with variant induced de-novo or loss of transcription factor binding site [59]. Liu et al. identified breast cancer-associated genetic variants affecting TF binding [60].

In this work, we found the presence of rSNPs (regulatory regions being TFBS) within lncRNA loci. TFBS information has been retrieved from Unibind analyzed Chip-seq datasets [49]. 39798, 10633 and 28407 such rSNPs present within lncRNA loci were found throughout breast, ovarian and cervical cancer respectively. In other words, among the dbSNP annotated lncRNA variants, the maximum number of rSNPs are found in cervical cancer, whereas those in novel variants are observed in breast cancer (Supplementary Fig. S6). Human Papilloma Virus which causes cervical cancer has been reported to bring about mutation within immune and DNA repair-related genes during integration in the Human genome [61]. Hence polymorphism within the regulatory region, as obtained (within lncRNA loci) in our analysis can be a mechanism adopted by the virus to modulate the host transcriptome environment.

**(E) CpG-SNPs within lncRNA-loci:** A short stretch of DNA in which a high frequency of the dinucleotide CG is found, compared to other regions is known as the CpG Island. The cytosines in CpG dinucleotides can be methylated to form 5-methyl-cytosine, leading to the formation of transcriptional repression complex, resulting in shutting down of the island and abolished transcription [62]. CpG-SNPs are point mutation at CpG sites which may lead to hyper/hypomethylation if an SNP allele has the potential to generate/destroy CpG dinucleotide. CpG-SNP rs7766585 has been found to have a strong association with breast cancer risk [63]. The impact of SNPs within the CpG Islands in oncogenes and tumor suppressor genes has also been documented [64]. Our analysis revealed *chr15:100554315–100559288*, *chr21:8431968–8441142* and



**Figure 6.** Distribution of SNP associated lncRNAs across different biotypes.

Abbreviations used: NC: Non Coding; MI: lncRNA; SO: Sense Overlapping; PS: Pseudogenes; CI: Completely Intronic; 3UTR: 3/UTR overlapping; PT: Processed Transcript; 5UTR: 5/UTR overlapping; CDS: CDS overlapping; IO: Intron Overlapping; LI: Long Intergenic; AO: Antisense Overlapping; AI: Antisense Intronic; AG: Ambiguous; BI: Bidirectional lncRNAs.

*chr14:19300376–19302642, within lncRNA loci to be SNP enriched CGI regions. Further, 4.2%, 4.5% and 6.8% of lncRNAs harbouring SNPs have been found to be present within the CpG regions in breast, ovarian and cervical cancer respectively.*

CGI name, length, GC percent, CpG percent, and CpG density have been noted for SNP overlapped CGI regions falling within the lncRNA loci. Tissue and cell line specific distribution of novel and annotated lncRNA variants within CGI have been represented in Supplementary Fig. S6 for all three cancers.

(:

(F) *Effect of SNP on lncRNA secondary structure:* The recruitment of LSD1 and PCR2 complex by HOTAIR is directly associated with its conserved secondary structure [65]. A combinatorial point mutation in the stem-loop structure of Rox1 lncRNA in drosophila inhibits MSL complex binding with subsequent loss of dosage compensation and male lethality [66]. Chemical and enzymatic probing approach to study the conserved structure of lncRNAs

revealed secondary and tertiary structures to be highly conserved and directly related to their biological functions [67]. Thus, a single base pair alteration could lead to structural instability and functional alteration. Using RNAsnp [68] 15.4%, 15.73% and 15.5% of the SNP containing lncRNAs in breast, ovarian and cervical cancer respectively have been predicted to have structural disruption as they harbour SNP(s). This prediction is based on a base pair probability cut off score(p-val). Based on this, around 15% of the detected lncRNAs, harbouring SNPs for each disease system, have

**Table 3.** Case studies with the lncRNA-SNPs hosted in ‘ClinicLSNP’ corroborating with experimentally validated data.

Disease System	lncRNA	SNP	Ref allele	Alt allele
Epithelial ovarian Carcinoma	HOXA11-AS	<b>rs17427875</b>	A	T
Breast Cancer	CCAT2	<b>rs6983267</b>	G	T
Cervical Cancer	HOTAIR	<b>rs2366152</b>	A	G



significantly higher chances of having structural perturbation (Supplementary Fig. S7).

(G) *A case study with ClinicLSNP results:* In order to check whether our results corroborate with the experimentally validated data, we have done some case studies with the lncRNA-SNPs hosted in 'ClinicLSNP' (given in Table 3).

(i) The homeobox A (HOXA) region of protein-coding genes is involved in regulating embryogenesis and ovarian carcinogenesis of the female reproductive system. Richard et al showed that **rs17427875** (A > T), a variant within the exon of downregulated HOXA11-AS in ovarian tumors can be linked to reduced cancer risk implying the tumor repressor function of HOXA11-AS [69].

(ii) Redis et al analyzed the expression, function and clinical correlation of CCAT2, a lncRNA overlapping SNP **rs6983267** (G > T), in breast cancer patients [70].

(iii) Concerning tumorigenesis in the case of virus-induced cervical cancer, Saha et al have identified the role SNP **rs2366152** (A > G) by affecting the secondary structure of HOTAIR, a lncRNA which actively participates in cellular chromatin reprogramming [71].

Our findings are on par with these validated results. We have detected the above variants in similar systems as that seen in these experiments. All these three variants are also found to be annotated as trait-associated in our analysis.

## Materials and methods

### Improved content and new features

#### Data procurement

lncRNA entries have been expanded to six new species besides Human and Mouse. The genome build and transcript sources for all eight species are mentioned in Supplementary Table ST1. For Human, the build is updated from hg19 to hg38. For the NONCODE [6] data for *C.elegans* and Chicken, the genome build has been lifted over from ce10 to ce11 and from gal-gal 4 to gal-gal5 respectively using UCSC lift over utility [72]. Data are retrieved in various formats such as fasta, bed, and gtf. Repeat elements, CGI fasta and coordinates, as well as Refseq annotated intron, 5'UTR/3'UTR/CDS exon information are downloaded from UCSC [72]. The miRNA and piRNA related information corresponding to the same build is downloaded from miRBase [73] and National Centre for Biotechnology Information (NCBI) respectively [74]. For probe remapping, microarray probes are downloaded from the Affymetrix website (<http://www.affymetrix.com>). Ribosome profiled data of six species have been downloaded from sorfs.org [75]. Normal tissue-specific total and long RNA sequence data has been downloaded from NCBI Gene Expression Omnibus (GEO) [76].

### Data processing and refinement

#### A. Redundancy check and assigning alias ID

The procedure of redundancy removal and ID assignments is similar to that followed in LncRBase [13] which is as follows: [abbreviation of species name] LB\_ [subtype] \_ [number] or LB\_ [subtype] \_ [number. numerical index] (for multiple

transcripts within the same chromosomal loci). However, for a transcript falling under multiple biotype categories, the subtype has been referred to as ambiguous (AG), instead of assigning multiple IDs for a single transcript. For eg. hsaLB\_AG\_208424 has been assigned for the transcript which is both 5'UTR exon overlapping as well as intron overlapping.

#### B. Association with CpG island, Repeat elements and small non-coding RNAs

For all species, mapping of repeat elements, CGI, identifying piRNA associated lncRNAs (Supplementary Fig. S2), miRNA associated lncRNAs (Supplementary Fig. S4) and remapping microarray probes with lncRNA transcripts follow the same method as that followed in the first version of LncRBase [13].

#### C. Determining putative lnc-pri-miRNAs

The coordinates of species-specific miRNAs have been mapped to the respective coordinates of lncRNA transcripts. lncRNA and miRNA transcripts coming from the same loci were further aligned using BLAST-like alignment tool (BLAT) [77] with block count = 1 and gap count = 0. These transcripts fulfilling both the criteria of coordinate matching as well as exact sequence matching has been annotated to be lnc-pri-miRNAs.

#### D. Coding capacity of lncRNAs:

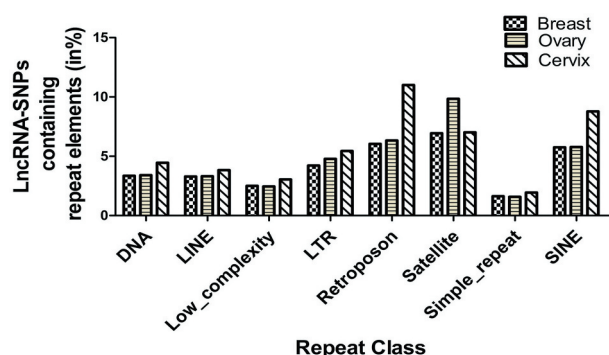
In the first version of LncRBase, standalone Coding-Potential Assessment Tool (CPAT) [38], was used to check the coding probability of the lncRNA transcripts. However, CPAT limits the assessment of coding potential to only four species (Human, Mouse, Fruitfly and Zebrafish). Hence, this time we have used two other renowned tools viz. CPC2 [37] and PLEK [39] in addition to CPAT. CPC2 is species neutral and PLEK is an alignment-free tool.

We tabulated scores from all the three tools and the corresponding predictions along with it. The updated version of CPAT (CPAT1.1.3) has been used for Human, Mouse, Fruitfly and Zebrafish. CPC2 and PLEK have been used for all the eight species. The potential coding threshold for CPAT is 0.364, 0.44, 0.39 and 0.38 for Human, Mouse, Fruitfly and Zebrafish respectively and that for CPC2 and PLEK is 0.5, 0 respectively for all species.

For putative micro peptide coding evidence, coordinate and other information of sORF sequences (the approximate cut off for sORF length are around 300 nts), have been retrieved from sorfs.org [75] for six species (Human, Mouse, Fruitfly, Zebrafish, Rat and *C.elegans*). Coordinate mapping of sORFs with lncRNA loci and corresponding sequence alignment (using BLAT) has been done to screen out the lncRNAs bearing putative sORFs.

#### E. Sub-cellular localization of lncRNAs

A web-based tool named LncLocator by Cio et al [78], has been used to predict sub-cellular localization of lncRNA transcripts for all the species. This tool utilizes Neural Network based stacked ensemble approach, fusing four base classifiers to a single model. The performance of this tool is better than the other contemporary tools as it integrates the power of individual predictors into one model through stacked ensemble method, which is known for having better performance than their individual counterparts [79].



**Figure 7.** Distribution of lncRNA-SNPs associated with repeat elements across different repeat classes

This tool predicts a putative score for five sub-cellular locations viz: Cytoplasm, Nucleus, Ribosome, Cytosol, and Exosome with an overall accuracy of 0.59. The most probable sub-cellular location for a particular lncRNA transcript has been selected to be the one having the highest score. We have provided the information regarding the possible localization of the lncRNA transcripts as represented in Fig. 7.

#### F. TFBS in lncRNA promoter region

High confidence TF-DNA interaction data in the form of bed files have been downloaded from Unibind [49], a database which holds binding site predictions of 232 unique TFs processed from 1983 Chip-seq peak datasets. Custom scripts have been used to sort and map the data within lncRNA promoter regions restricted to 1000 bp upstream and downstream of TSS.

#### G. lncRNA profile across multiple tissues

Normal tissue-specific raw long RNA-Seq data (provided in **Supplementary File SF1**) for all the species obtained from GEO [80] were aligned to their respective reference genome using Hisat2 [81] after quality checking (using FastQC [82]) and adapter trimming (using Cutadapt [83]). Stringtie was used for assembling transcripts, and transcript FPKM files were generated thereafter [81].

### For ClinicLSNP

#### Raw data corresponding to the three cancer systems:

Raw RNA sequence cell line and tissue data of ovarian, breast and cervical cancers and their normal counterparts have been downloaded from GEO [76] and ArrayExpress [84]. Annotated SNP data for Human (GRCh37) has been downloaded from NCBI dbSNP(v151)[50]. A total of 280 raw RNA-Seq tissue data, 281 cell line data including 147 different cell lines corresponding to three cancers and 125 normal tissue RNA-seq data (Ovarian = 32, Breast = 88, Cervix = 5) have been analyzed for SNP detection. For better precision, the cancer data has been categorized based on its subtype (ovarian n = 10, breast n = 3 and cervix n = 2). Input data and results are provided in **Supplementary File SF2** and **Supplementary Table ST2** respectively.

#### 1. SNP detection:

For each sample, the quality of reads has been checked with FastQC [82]. All reads with quality >35 have been retained. Adapters have been trimmed with Cutadapt [83]. Paired-end raw sequence reads have been mapped with Hisat2 to the Human reference genome (hg38). Sorting and indexing of BAM files are done using SAMtools 0.1.19 [85]. Prior to variant calling, Bam files have been pre-processed with Opossum 0.2, a tool that performs quality control measures such as discarding duplicate reads, poorly mapped reads, secondary alignments and also merges overlapping reads [86]. Opossum is most compatible to work with Variant Caller Platypus which has been then called using default parameters for SNP and indel detection [87]. Platypus is a much faster variant detection tool with similar sensitivity to that of leading variant callers available. Those variants with QUALITY = PASSED have been filtered out from the output. To retain high-quality variants, those with the number of variants containing reads (TR) >5 are only considered. If TR<5 in a particular sample (corresponding to a particular subtype), the variant is selected only if it is present in more than 60% of the samples under that particular subtype. Custom scripts are used to merge all the variants containing output files under each subtype followed by redundancy removal. Variants from normal counterparts are then depleted from the cancerous subtype with custom scripts. Variants are annotated with dbSNP(v151) using 'bcftools annotate'[85]. Finally, variants are mapped with the lncRNA transcripts present in this new version of lncRBase using bedtools Intersect [88].

#### 2. Mapping SNP association with Repeat and CGI

SNPs have been mapped to repeat elements belonging to different repeat class and families as well as CpG islands downloaded from UCSC [72]

#### 3. Mapping TFBS overlap with SNP

TFBS information from analyzed CHP-seq data has been downloaded from Unibind.uio.no [49]. Custom scripts have been used to sort the data and map SNP within lncRNA loci to the TFBS. The output records the TF name, position, sequence, Data accession No. and cell line/tissue information corresponding to the mapped SNP.

#### 4. Mapping TAG SNPs

Information regarding trait-associated variants have been downloaded from SNPsnap [52], an SNP based enrichment analysis web server which retains common variants (>1% maximum allele frequency) from 1000 Genomes Project Phase 3 encompassing three super populations: European, East Asian and West African. 1,99,16,464 total variants are included with identifiers consisting of both chromosomal coordinates and rs IDs. For variants with chromosomal coordinate identifiers, the genomic coordinates were converted from genome assembly hg19 to hg38 using the LiftOver utility in UCSC (<http://genome.ucsc.edu>). Matched SNPs are marked as 'yes' under the category 'TagSNP' in SNP search result page.

#### 5. Determining structure perturbation score due to the presence of SNP

A web-based/standalone tool, RNAsnp [68] is being popularly used for structure disruptive SNP prediction. The

A

**Browse by lncRNA Biotype and Coding Potential**

Organism  Chromosome  Position  To   
[eg: From 1 To 1000000000]

LncRNA Biotype  Coding potential tool  Coding potential   
(# CPAT1.2.3 tools for Human, Mouse Zebrafish and Drosophila)

**Search for sORF overlap**

Organism  Chromosome  LncRNA Biotype  Search for sORF :

B

Search results for lncRNA: **hsaLB\_LI\_31323**

General Information			
LncRNA ID:	hsaLB_LI_31323		
Biotype(s):	linc		
Position:	chr11: 1998176-1997814 (-)		
Length:	2281		
GC Percent:	63.4371		
Gene Symbol:	H19		
Assembly:	GRCh38 hg38		
Source ID:	ENST00000412788		
Source Alias ID:	NONHSAT017462.2, uc001ha.5, H19-9		

Coding potential Score			
Tool	Score	Prediction	Foot Printing Evidence
CPAT1.2.3	0.964342958	putatively_coding	yes
CPC2	0.822735	putatively_coding	sORF Count(s)
PLEK	-1.18566	noncoding	4

Association with small ncRNA	
miRNA count :	1 <input type="button" value="View"/>
piRNA count :	0

Association with other genomic elements	
Associated Repeat Elements:	0
Promoter associated CpG Island:	0
Mapped Microarray probes:	5 <input type="button" value="View"/>

Localization Score				
Cytoplasm	Nucleus	Ribosome	Cytosol	Exosome
0.0807937	0.0585251	0.0270052	0.0301655	0.80351 <input checked="" type="checkbox"/>

C

Expression value (FPKM)					
Tissue	Adrenal (GSM1244381)	Adrenal (GSM12453456)	Adrenal (GSM12444170)	Testis (GSM12453457)	Spleen (GSM12453444)
FPKM	65.613899	32.084637	21.727032	12.032747	7.169962

[View More](#)

Regulatory region & TFs				
Gene	Chromosome	Promoter Start	Promoter End	TF Details
H19	chr11	1994176	1996176	<a href="#">view</a>

D

**LncRBase V2**

Completely Intronic  
Intergenic  
Intronic Overlapping  
Antisense Intronic  
Antisense Overlapping  
Pseudogene  
Sense Overlapping  
Processed Transcript  
3'UTR Overlapping  
5'UTR Overlapping  
Non coding  
Miscellaneous  
Ambiguous  
CDS Overlapping  
Mitochondrial  
Bidirectional  
Macro

Select  
Prepupe  
Central Nervous System  
Multidendritic sensory neuron  
Head  
Brain  
Ovary  
Testis  
Seminal vesicles  
Antennae

Expression ClinicL SNP Download

Search lncRNA expression in Drosophila

Tissue  GSM ID's   
GSM1032806  
GSM1032807  
LncRNA Type

\* antennae, brain, head, seminal vesicles, testis and ovary in ADULT  
\* Central Nervous System(CNS), Multidendritic Sensory Neuron in THIRD INSTAR LARVAE

**Figure 8.** Multiple web interfaces for easy access of LncRBase V.2.

A. General search page options. B. Detailed information of individual LncRBase IDs including general information, coding potential Score, association with small ncRNAs, association with genomic elements and cellular localization score. C. Detailed information of individual LncRBase IDs including tissue wise expression values and Transcription Factor overlap within the regulatory region. D. Tissue-specific search for individual species across multiple Data Accession IDs.

standard empirical p-value cut off is  $\leq 0.2$ , a lower score indicates possible structural effects on RNA secondary structure in the presence of an SNP.

## Database implementation

In LncRBase V.2, user query is mainly processed through simple search options, and information is displayed on the web interface after retrieving from relational databases (Figs. 8 and 9). The General Output page (Fig. 8) displays necessary information about the lncRNA transcript and provides multiple options for probing into further details. The Detailed Output page shows complete information about the lncRNA transcript.

(a) The following search options are under the 'Transcripts' menu:

1. **Search by lncRNA Accession ID:** User can select the name of the organism and then input specific lncRNA Accession ID from (a) LncRBase V.2 like hsaLB\_LI\_10017 for Human, dreLB\_AO\_675 for Zebrafish, etc. or from (b) any of the source databases (Ensembl Gene90, UCSC ID, NONCODE v5.0, H-InvDB 8.0, LNCipedia) to get detailed

information about that transcript. On submitting the query, a result page comes up with query feature specific information.

2. **Search by lncRNA Gene Symbol:** User can select the name of the organism and provide known lncRNA Gene Symbol as input to search for the gene-specific transcript entries listed in LncRBase V.2; the results page displays the list of information about all the transcripts (with LncRBase V.2 IDs) corresponding to that gene symbol as well as provide literature evidence about that particular lncRNA gene. Detailed transcript information will open by clicking on any of the displayed LncRBase IDs.

3. **Browse by lncRNA subtype, coding potential, and sORF overlap:** For the first part, the user needs to select organism name, chromosome number, lncRNA biotype, coding potential tool name and coding potential type to view the corresponding result. Users can also avail of the extra option of specifying genome locus. For the second part, organism name, chromosome input and lncRNA biotype would retrieve the sORF count of the retrieved lncRNAs.



**A**

Search by SNP ID or Position

ID / Position :

[e.g. rs61350339 or chr17:82671305]

**B**

Search by Cell line Name

Cell Line :

[Breast, Cervical & Ovarian cancer]

**C**

Search SNP for Breast Cancer

Subtype

**D**

SNP information of rs2232677

SNP ID:	rs2232677
Position:	chr13:27621874
Ref.:	C
Alt:	A
Clinvar ID:	311643
Clinvar Type:	Likely benign
CADD Score:	14.32
Tag SNP:	Yes
Repeat Overlap:	1
TF Overlap:	Yes
CpG Overlap:	1

SNP within lncRNA

lncRBase ID	Structure disruptive score
hsaLB AG 54589	0.3476 ✖
hsaLB PT 54588	0.3287 ✖

Analyzed Cancer Dataset for rs2232677

Cancer Type	Sample Count	Details
Cervix Squamous cell carcinoma tissue	1	<input type="button" value="View"/>
Ovarian High grade Serous cell line	2	<input type="button" value="View"/>
Ovarian High grade Serous tissue	9	<input type="button" value="View"/>

**Figure 9.** Multiple web interfaces for easy accessing of ClinicLSNP.

A. Search by SNP ID or position. B. Search by Breast, Cervical and Ovarian cancer Cell line name. C. Individual Cancer subtype-specific search page D. Detailed information of individual SNP IDs.

**4. Search for associated genomic elements:** Users can search for associated ‘Repeat elements’, ‘CGI’ and remapped microarray ‘probe’ by selecting the desired option with organism name and chromosome number.

**5. Search for association with small ncRNA:** Organism and chromosome-specific search of lncRNAs associated with small non-coding RNAs i.e. ‘miRNAs’ and ‘piRNAs’ can also be avoided.

**(b) Species-specific lncRNA expression profile:** Clicking an organism name under the ‘Expression’ menu will open the organism-specific lncRNA expression query page. Selecting a specific ‘tissue’ will show all the corresponding data ‘accession IDs’. An additional option of limiting the number of outputs is to submit the query corresponding to a particular ‘lncRNA biotype’.

**(c) Search options under Section ‘ClinicLSNP’:**

**1. Search by SNP ID/position:** Within this search option, Single rsID or genomic location is provided as input to obtain detailed variant information such as overlap with Repeat, TF, CGI and, trait association. Hyperlink to each attribute (if present), would lead to detailed information regarding it. Hyperlink to dbSNP and related Clinvar ID(if present) is also provided. Further, a hyperlink is provided to the corresponding lncRNA IDs leading to the transcript details page. Information is also provided regarding the cancer datasets where this SNP has been found.

**2. Search by Cell line:** Here variant information can be searched within specific cell lines. The search result will provide information regarding all the variants obtained within that specific cell line as well as their location, allele information and corresponding lncRNA IDs. A ‘Download total data’ option has been provided on this page to download the entire data.

**3. Search by Disease:** This will facilitate users to search ClinicLSNP by individual Cancer type. Output results will be obtained based on user input information on Cancer subtype, data type (cell line or tissue data), variant type (novel or annotated) and genomic location.

The home page provides a quick tour regarding the contents of the entire database. Details of all this information have been provided on our ‘Information’ page under the Help menu. Further, the ID conversion page under the Help menu facilitates searching the first version of lncRBase for Human and Mouse lncRNAs corresponding to their IDs in this new version of lncRBase.

lncRBase V.2 has been developed using MySQL, which is an open-source relational database management system. This web server runs in a Linux environment using Apache HTTP Server. It is also free and open-source cross-platform web server software. The interface layer has been designed using HTML, CSS, and JavaScript. The database is connected to the web interface using PHP module. Since PHP is a server-side scripting language it creates dynamic pages and access data from MySQL to produce output (Fig. 10)

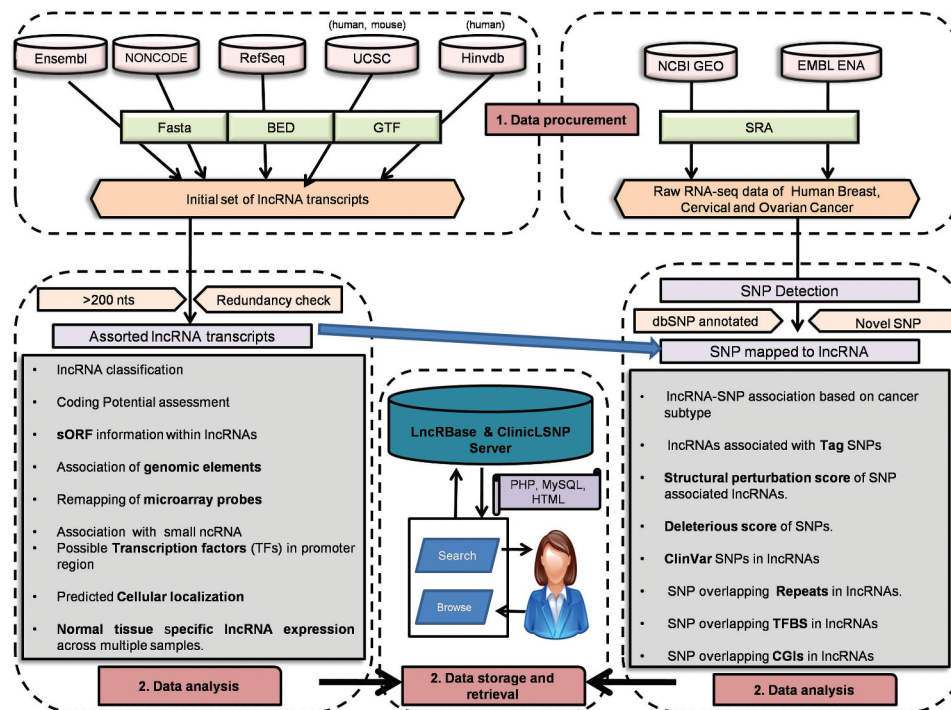


Figure 10.

## Conclusion

Since our last release in 2014, there has been quite an increase in the number of lncRNAs identified and catalogued. With the advent of novel technologies, it has been possible to look into the structure, function and interacting partners of some of the lncRNA molecules to a particular extent. Still, for the majority of them, their nature and mode of operation are vague, though their association with disease risk is coming up from time to time. Initially, we developed LncRBase, with the aim to build a non-redundant comprehensive catalogue of these versatile transcripts in order to provide a better understanding of these molecules in the context of their genomic localization, tissue-specific expression, remapped microarray probes, overlaps with genomic elements and their association with other small ncRNAs. In this new version of LncRBase, we have extended our findings to six more species covering mammals, bird, insect, nematode, and fish. The newly discovered Human and Mouse lncRNA transcripts have also been incorporated, and obsolete ones have been removed. Along with the existing features of the former version of LncRBase, this database has come up with 6 new key updates: (i) User-friendly nomenclature of lncRNA transcripts with newly added biotypes (ii) Coding potential assessment by multiple tools along with possible micro peptide coding evidence from footprinting studies (iii) TFBS in the lncRNA promoter region (in Human) (iv) Tissue-specific lncRNA expression analysis across multiple samples per tissue for all species (v) Elucidating possible lnc-pri-miRNA transcripts and (vi) predicting cellular localization of lncRNAs. Literature curated information about the role of lncRNAs in various disease systems along with their interacting partners have also been incorporated. To the best of our knowledge,

such extensive characterization of lncRNAs corresponding to multiple species besides Human and Mouse has not been reported previously. LncRBase V.2 also hosts ClinicLSNP, a repository of lncRNA variants (SNP/insertion/deletion) in three female cancer systems. We have come up with a set of novel and annotated lncRNA-SNPs for each cancer subtype. Annotation with Clinvar (for the annotated IDs) has been done to understand their clinical relevance. CADD score indicates the deleteriousness of variants, which is meant for screening functional SNPs. 'Tag' indicates Trait associated variants in the region of Linkage Disequilibrium. The presence of lncRNA variants within CGI, TFBS and STR might have a significant downstream effect. The structure disruptive score of a lncRNA-SNP pair is indicative of the structural perturbation efficiency of a particular variant for that lncRNA transcript position. However, it cannot be used to calculate structural perturbation by in-dels and hence p-values for those lncRNA-variant associations have been tagged as 'NA' here. ClinicLSNP would serve as a readily available source of clinically relevant lncRNA variants and would contribute towards functional studies of disease-associated lncRNAs.

The updated features, over the existing ones, would make LncRBase V.2 a comprehensive, data-intensive repertoire of lncRNA. Our future aim is to update LncRBase V.2 with newly annotated lncRNAs adding new features to it, to keep pace with the rapid growth rate of lncRNA research worldwide.

## Disclosure of potential conflicts of interest

No potential conflict of interest was reported by the authors.



## Acknowledgments

We are grateful to Council of Scientific and Industrial Research (CSIR) and Science and Engineering Research Board (SERB) for financial support. We thank European Genome-phenome Archive (EGA) and Frederic de-Sauvage of Genentech for providing us access to the EGAD00001000725 dataset. We also thank and acknowledge Dr. Sohini Chakraborty of New York University School of Medicine for critical comments on the database.

## Funding

This work was supported by the Council of Scientific and Industrial Research; Science and Engineering Research Board.

## Availability

LncRBase V.2 is freely available at <http://dibresources.jcbose.ac.in/zhu/mur/lncrbase2/>. Files can be freely downloaded and used in accordance with the GNU Public License.

## References

- [1] Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* **2009**;23:1494–1504.
- [2] Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. *Annu Rev Biochem.* **2012**;81:145–166.
- [3] Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. *Nat Rev Genet.* **2009**;10:155–159.
- [4] Cerik S, Schwarzenbacher D, Adiprasito JB, et al. Current status of long non-coding rnas in human breast cancer. *Int J Mol Sci.* **2016**;17:1485.
- [5] Flicek P, Amode MR, Barrell D, et al. Ensembl 2012. *Nucleic Acids Res.* **2012**;40:D84–90.
- [6] Fang S, Zhang L, Guo J, et al. NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* **2018**;46:D308–D14.
- [7] O’Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**;44:D733–45.
- [8] Volders P-J, Anckaert J, Verheggen K, et al. LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res.* **2018**;47:D135–D9.
- [9] Zhou KR, Liu S, Sun WJ, et al. ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Res.* **2017**;45:D43–D50.
- [10] Gao Y, Wang P, Wang Y, et al. Lnc2Cancer v2.0: updated database of experimentally supported long non-coding RNAs in human cancers. *Nucleic Acids Res.* **2019**;47:D1028–D33.
- [11] Bao Z, Yang Z, Huang Z, et al. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* **2018**;47:D1034–D7.
- [12] Ren C, An G, Zhao C, et al. Lnc2Catlas: an atlas of long noncoding RNAs associated with risk of cancers. *Sci Rep.* **2018**;8:1909.
- [13] Chakraborty S, Deb A, Maji RK, et al. LncRBase: an enriched resource for lncRNA Information. *PloS One.* **2014**;9:e108010.
- [14] van Heesch S, van Itersson M, Jacobi J, et al. Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol.* **2014**;15:R6.
- [15] Chen LL. Linking long noncoding RNA localization and function. *Trends Biochem Sci.* **2016**;41:761–772.
- [16] Yeasmin F, Yada T, Akimitsu N. Micropeptides encoded in transcripts previously identified as long noncoding RNAs: a new chapter in transcriptomics and proteomics. *Front Genet.* **2018**;9:144.
- [17] Olexiouk V, Crappe J, Verbruggen S, et al. sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* **2016**;44:D324–9.
- [18] Reich DE, Gabriel SB, Altshuler D. Quality and completeness of SNP databases. *Nat Genet.* **2003**;33:457–458.
- [19] Hajjari M, Rahnama S. Association between SNPs of long non-coding RNA HOTAIR and Risk of Different Cancers. *Front Genet.* **2019**;10:113.
- [20] Lin Y, Guo W, Li N, et al. Polymorphisms of long non-coding RNA HOTAIR with breast cancer susceptibility and clinical outcomes for a southeast Chinese Han population. *Oncotarget.* **2018**;9:3677–3689.
- [21] Dong J, Su M, Chang W, et al. Long non-coding RNAs on the stage of cervical cancer (Review). *Oncol Rep.* **2017**;38:1923–1931.
- [22] Worku T, Bhattarai D, Ayers D, et al. Long non-coding RNAs: the new horizon of gene regulation in ovarian cancer. *Cell Physiol Biochem.* **2017**;44:948–966.
- [23] Ning S, Yue M, Wang P, et al. LincSNP 2.0: an updated database for linking disease-associated SNPs to human long non-coding RNAs and their TFBSSs. *Nucleic Acids Res.* **2017**;45:D74–D8.
- [24] Miao YR, Liu W, Zhang Q, et al. lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. *Nucleic Acids Res.* **2018**;46:D276–D80.
- [25] Wang M, Tao X, Jacob MD, et al. Stress-induced low complexity RNA activates physiological amyloidogenesis. *Cell Rep.* **2018**;24:1713–21 e4.
- [26] Cabili MN, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **2011**;25:1915–1927.
- [27] Kaushik K, Leonard VE, Kv S, et al. Dynamic expression of long non-coding RNAs (lncRNAs) in adult zebrafish. *PloS One.* **2013**;8:e83616.
- [28] Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A.* **2002**;99:3740–3745.
- [29] Pueyo JI, Couso JP. Tarsal-less peptides control Notch signalling through the Shavenbaby transcription factor. *Dev Biol.* **2011**;355:183–193.
- [30] Matsumoto A, Pasut A, Matsumoto M, et al. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature.* **2017**;541:228–232.
- [31] Cai B, Li Z, Ma M, et al. LncRNA-Six1 encodes a micropeptide to activate Six1 in Cis and is involved in cell proliferation and muscle growth. *Front Physiol.* **2017**;8:230.
- [32] Mackowiak SD, Zauber H, Bielow C, et al. Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.* **2015**;16:179.
- [33] Kondo T, Plaza S, Zanet J, et al. Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science.* **2010**;329:336–339.
- [34] Nam JW, Choi SW, You BH. Incredible RNA: dual Functions of Coding and Noncoding. *Mol Cells.* **2016**;39:367–374.
- [35] Ulveling D, Francastel C, Hube F. When one is better than two: RNA with dual functions. *Biochimie.* **2011**;93:633–644.
- [36] Choi SW, Kim HW, Nam JW. The small peptide world in long noncoding RNAs. *Brief Bioinform.* **2019**; 20:1853–1864.
- [37] Kang YJ, Yang DC, Kong L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **2017**;45:W12–W6.
- [38] Wang L, Park HJ, Dasari S, et al. CPAT: coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **2013**;41:e74.
- [39] Li A, Zhang J, Zhou Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics.* **2014**;15:311.
- [40] Yan J, Zhang Y, She Q, et al. Long noncoding RNA H19/miR-675 axis promotes gastric cancer via FADD/Caspase 8/Caspase 3 signaling pathway. *Cell Physiol Biochem.* **2017**;42:2364–2376.
- [41] Zhang T, Nie K, Tam W. BIC is processed efficiently to microRNA-155 in Burkitt lymphoma cells. *Leukemia.* **2008**;22:1795–1797.

- [42] Diederichs S. Micro-terminator: 'Hasta la vista, lncRNA!'. *Nat Struct Mol Biol.* **2015**;22:279–281.
- [43] Dhir A, Dhir S, Proudfoot NJ, et al. Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs. *Nat Struct Mol Biol.* **2015**;22:319–327.
- [44] Cabili MN, Dunagin MC, McClanahan PD, et al. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* **2015**;16:20.
- [45] Carlevaro-Fita J, Johnson R. Global positioning system: understanding long noncoding RNAs through subcellular localization. *Mol Cell.* **2019**;73:869–883.
- [46] Kopp F, Mendell JT. Functional classification and experimental dissection of long noncoding RNAs. *Cell.* **2018**;172:393–407.
- [47] Luo J, Xiong Y, Fu PF, et al. Exosomal long non-coding RNAs: biological properties and therapeutic potential in cancer treatment. *J Zhejiang Univ Sci B.* **2019**;20:488–495.
- [48] Sun Z, Yang S, Zhou Q, et al. Emerging role of exosome-derived long non-coding RNAs in tumor microenvironment. *Mol Cancer.* **2018**;17:82.
- [49] Gheorghe M, Sandve GK, Khan A, et al. A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Res.* **2019**;47:e21.
- [50] Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **2001**;29:308–311.
- [51] Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **2014**;42:D980–5.
- [52] Pers TH, Timshel P, Hirschhorn JN. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics.* **2015**;31:418–420.
- [53] Rentzsch P, Witten D, Cooper GM, et al. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **2019**;47:D886–D94.
- [54] International HapMap C. The International HapMap Project. *Nature.* **2003**;426:789–796.
- [55] Gemayel R, Vences MD, Legendre M, et al. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet.* **2010**;44:445–477.
- [56] Payer LM, Steranka JP, Yang WR, et al. Structural variants caused by Alu insertions are associated with risks for many human diseases. *Proc Natl Acad Sci U S A.* **2017**;114:E3984–E92.
- [57] Gong C, Maquat LE. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature.* **2011**;470:284–288.
- [58] Fernandes JCR, Acuna SM, Aoki JI, et al. Long non-coding RNAs in the regulation of gene expression: physiology and disease. *Noncoding RNA.* **2019**;5:17.
- [59] Buroker NE, Regulatory SN. Ps and transcriptional factor binding sites in ADRBK1, AKT3, ATF3, DIO2, TBXA2R and VEGFA. *Transcription.* **2014**;5:e964559.
- [60] Liu Y, Walavalkar NM, Dozmorov MG, et al. Identification of breast cancer associated variants that modulate transcription factor binding. *PLoS Genet.* **2017**;13:e1006761.
- [61] Joo J, Omae Y, Hitomi Y, et al. The association of integration patterns of human papilloma virus and single nucleotide polymorphisms on immune- or DNA repair-related genes in cervical cancer patients. *Sci Rep.* **2019**;9:13132.
- [62] Li E, Zhang Y. DNA methylation in mammals. *Cold Spring Harb Perspect Biol.* **2014**;6:a019133.
- [63] Harlid S, Ivarsson MI, Butt S, et al. A candidate CpG SNP approach identifies a breast cancer associated ESR1-SNP. *Int J Cancer.* **2011**;129:1689–1698.
- [64] Samy MD, Yavorski JM, Mauro JA, et al. Impact of SNPs on CpG Islands in the MYC and HRAS oncogenes and in a wide variety of tumor suppressor genes: a multi-cancer approach. *Cell Cycle.* **2016**;15:1572–1578.
- [65] Wu L, Murat P, Matak-Vinkovic D, et al. Binding interactions between long noncoding RNA HOTAIR and PRC2 proteins. *Biochemistry.* **2013**;52:9519–9527.
- [66] Ilik IA, Quinn JJ, Georgiev P, et al. Tandem stem-loops in roX RNAs act together to mediate X chromosome dosage compensation in *Drosophila*. *Mol Cell.* **2013**;51:156–173.
- [67] Li R, Zhu H, Luo Y. Understanding the functions of long non-coding RNAs through their higher-order structures. *Int J Mol Sci.* **2016**;17:702.
- [68] Sabarinathan R, Tafer H, Seemann SE, et al. RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. *Hum Mutat.* **2013**;34:546–556.
- [69] Richards EJ, Permuth-Wey J, Li Y, et al. A functional variant in HOXA11-AS, a novel long non-coding RNA, inhibits the oncogenic phenotype of epithelial ovarian cancer. *Oncotarget.* **2015**;6:34745–34757.
- [70] Redis RS, Sieuwerts AM, Look MP, et al. CCAT2, a novel long non-coding RNA in breast cancer: expression study and clinical correlations. *Oncotarget.* **2013**;4:1748–1762.
- [71] Sharma Saha S, Roy Chowdhury R, Mondal NR, et al. Identification of genetic variation in the lncRNA HOTAIR associated with HPV16-related cervical cancer pathogenesis. *Cell Oncol.* **2016**;39:559–572.
- [72] Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **2004**;32:D493–6.
- [73] Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **2011**;39:D152–7.
- [74] Geer LY, Marchler-Bauer A, Geer RC, et al. The NCBI BioSystems database. *Nucleic Acids Res.* **2010**;38:D492–6.
- [75] Oleksiuk V, Van Crielinge W, Menschaert G. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* **2018**;46:D497–D502.
- [76] Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol.* **2006**;411:352–369.
- [77] Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res.* **2002**;12:656–664.
- [78] Cao Z, Pan X, Yang Y, et al. The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. *Bioinformatics.* **2018**;34:2185–2194.
- [79] Pan XY, Tian Y, Huang Y, et al. Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach. *Genomics.* **2011**;97:257–264.
- [80] Clough E, Barrett T. The gene expression omnibus database. *Methods Mol Biol.* **2016**;1418:93–110.
- [81] Pertea M, Kim D, Pertea GM, et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc.* **2016**;11:1650–1667.
- [82] S. A. FastQC: a quality control tool for high throughput sequence data. **2010**. [cited 2017 Oct 29]. Available from: <http://wwwbioinformaticsbabrahamacuk/projects/fastqc>
- [83] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **2011**;17:3.
- [84] Kanz C, Aldebert P, Althorpe N, et al. The EMBL nucleotide sequence database. *Nucleic Acids Res.* **2005**;33:D29–33.
- [85] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* **2009**;25:2078–2079.
- [86] Oikonen L, Lise S. Making the most of RNA-seq: pre-processing sequencing data with Opossum for reliable SNP variant detection. *Wellcome Open Res.* **2017**;2:6.
- [87] Rimmer A, Phan H, Mathieson I, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet.* **2014**;46:912–918.
- [88] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* **2010**;26:841–842.