Regulatory Noncoding RNA Mediated Alterations and its Effects in Stem Cell Derivatives

---

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

Submitted by

BYAPTI GHOSH

DEPARTMENT OF

LIFE SCIENCE AND BIOTECHNOLOGY

JADAVPUR UNIVERSITY

2024

In loving memory of my father,

whose unwavering belief and

encouragement inspired this journey

to fulfill his cherished dream

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
Also: ghosh.jhumur@gmail.com

TO WHOM IT MAY CONCERN

This is to certify that the thesis entitled **"Regulatory Noncoding RNA mediated alterations and its effects in stem cell derivatives"** submitted by **Smt. Byapti Ghosh** who got her name registered on 03.09.2019, 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 15/3/2024

(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, Bidhannagar
कोलकाता / Kolkata-700 091 (भारत/India)

Acknowledgements

"Let us be grateful to the people who make us happy; they are the charming gardeners who make our souls blossom." - Marcel Proust

Research is a delightful opening to knowledge, marked by gentle whispers of curiosity and wonder. On this enchanting path, we are blessed with the presence of dear companions whose warmth and encouragement light up even the darkest corners of our exploration. As I reflect on the journey of my research endeavors, I am filled with profound gratitude for the unwavering support and kindness of these extraordinary souls. Their beliefs in my dreams and their gentle guidance through both sunny days and stormy nights have been the guiding stars that brighten my path.

Firstly, my acknowledgement goes to my family. Expressing gratitude to them feels like an understatement, as their support and love have been immeasurable. Throughout this difficult time, you have been there to bless me and guide me in the right direction, ‘Baba’, I believe. My mother, ‘Ma’, have been the pillar of my education, constantly encouraging me in my endeavors and showing genuine interest in my progress. Her support and inquiries about my experiments have meant the world to me. I also want to extend my heartfelt thanks to my beloved sister, Tinna, whose affection and care have been a constant source of comfort, despite the physical distance between us. Their love and support has been a lifeline, reminding me that, even in solitude, I am never truly alone. My relatives, too, have played an invaluable role in my journey, offering unwavering support and encouragement during times of adversity. Their presence has been a beacon of hope, guiding me through the challenges of life.

It has been an honor and a privilege to be a part of Dr. Zhumur Ghosh's esteemed research laboratory. Her genuine attention, continuous support, insightful suggestions and invaluable guidance have been pivotal in shaping the successful completion of this thesis. Beyond being a mentor, she has been a beacon of inspiration, enriching my perspective not only on research but also on the profound purpose of life itself. I am deeply grateful for the countless uplifting moments, enlightening discussions and invaluable lessons learned under her tutelage. I will carry with me the invaluable wisdom and blessings bestowed upon me by her, and I remain hopeful for her continued guidance and support in the journey ahead.

I consider myself fortunate to have received experimental assistance and valuable guidance for my research from Dr. Angshuman Bagchi at Kalyani University. I extend my heartfelt gratitude and appreciation to him for his support and mentorship.

I am grateful to DST and Bose Institute for providing me with a research fellowship, which has enabled me to pursue my work.

When enveloped by a community of bright minds and compassionate labmates, the research expedition naturally flows with ease and grace. Firstly, I extend my heartfelt gratitude to my senior, Arijita Di whose wisdom and guidance have illuminated my path and enriched my journey. Her mentorship has been a guiding light, smoothing the path of my research journey and nurturing my aspirations. I am deeply thankful to my friend and colleague in the lab, Troyee, whose contagious energy and imaginative contributions have turned our research venture into a truly gratifying and inspiring one. I have found immense pleasure in receiving assistance and valuable insights from esteemed lab members Gourab Da, Arpana Di, and Aritra Da, Namrata, Sohini di and Sudip. Besides, Pritha Di and Satakshi, the new members of our lab have been great supportive. I will acknowledge my friends Abhirupa and Debadrita for their constant support.

I express my heartfelt gratitude to my beloved partner, Tanmoy, whose steadfast love and support have been a constant source of strength and inspiration in my journey.

I am deeply grateful to all the professors and members of my department for their invaluable guidance and support throughout my academic journey. Additionally, I extend my sincere appreciation to the Director of Bose Institute for providing me with the necessary institutional resources. My heartfelt thanks also go out to all members of the CIF and non-academic staffs at Bose Institute for their assistance.

Lastly, I wish to express my profound gratitude to all who have aided me directly or indirectly during my research journey, contributing to the completion of my doctoral thesis.

`

Contents

CONTENT

`

Preface

In the past decade, there has been a significant shift in the field of molecular biology concerning RNA. RNA, traditionally viewed as a mere intermediary between DNA and proteins, is now recognized as a key player in regulating genome organization and gene expression. For a long time, RNA has been considered the fundamental molecule of life and the central focus of molecular biology, serving both informational and catalytic functions. Recent evidences challenge the conventional understanding of gene regulation in higher organisms, which has persisted for the last 50 years. The traditional assumption that proteins were the primary vehicles for transmitting genetic information in complex organisms has been upended with the discovery of RNA interference (RNAi) and the advent of high-throughput sequencing. This has led research groups to redirect their attention towards exploring the world of long and small regulatory noncoding RNAs (ncRNAs), which were previously neglected. This revolution in our understanding of regulatory ncRNAs, formerly considered "junk," has unveiled a new layer of gene regulation.

Regulatory ncRNAs encompass a variety of subclasses, each distinguished by its specific size, sequence, and mechanism of action, while collectively serving the common functional purpose of regulating gene expression. Referred to as the "guardians of the genome," they play pivotal roles at multiple levels, including influencing transcriptional regulation and epigenetic processes that govern differentiation and development. Moreover, regulatory ncRNAs play a significant role in the maintenance and differentiation of stem cells. In this context it is noteworthy to mention that human pluripotent stem cells (both human embryonic and human induced pluripotent stem cells) have the property to self-renew and differentiate into cells corresponding to the three germ layers. This event makes them an invaluable tool for regenerative medicine. However, the same properties make them oncogenic as well as transmit similar possibility within their derivatives. Hence, it is extremely important to check whether there is any role of regulatory ncRNAs in such oncogenic perturbations within these induced pluripotent stem cell derivatives. Unravelling the information embedded within these regulatory ncRNAs it is imperative to unveil their functional significance. Furthermore, delving into the regulatory networks orchestrated by these ncRNAs within stem cells derivatives will provide fresh insights towards enhancing their regenerative potential. This focus of this thesis is dedicated towards analyzing the characteristics of small regulatory ncRNAs such as microRNAs (miRNAs) and PIWI interacting RNAs (piRNAs), and their role in maintaining important cellular processes. An optimized blend of in-silico and experimental approach adopted in this work will aid towards devising a roadmap for safer regenerative therapy using these stem cell derivatives.

Chapter One '*A Brief Review*' presents a concise overview of prior research concerning the characterization and analysis of miRNAs and piRNAs within the scope of this thesis. This chapter highlights the significance of these small regulatory ncRNAs in the maintenance and regulation of stem cell biology.

CONTENT

Chapter Two '*Elucidating the role of miRNAs in stem cell derivatives corresponding to the three germ layers'* reveals an overlap in gene expression patterns between cancer cells and pluripotent stem cell derivatives, providing evidence of a remnant oncogenic signature within the stem cell derivatives for all the three germ layers. miRNAs play a pivotal role in influencing such carcinogenic processes by controlling the expression of their target genes. The insights presented in this chapter will contribute to a better understanding of the dynamics of miRNA and their target genes within stem cell derivatives.

Chapter Three 'Influence of reprogramming methods towards imparting oncogenicity to stem cell derivatives' represents the influence of the reprogramming methods (that has been used to generate the iPSCs) towards inducing oncogenicity to iPSC derivatives. Subsequently small RNA-seq data was generated to identify the differentially expressed miRNAs.

Chapter Four *'Development of a prediction model to predict the oncogenic status of an iPSC derived cell'* provides the entire blueprint for the development of a prediction model where significant set of miRNA target genes serve as the feature set and subsequently various machine learning models have been implemented to check the purity of the iPSC derivatives.

Chapter Five '*piRNAQuest V.2: updating the piRNAome for silencer'* deals with the systematic analysis and organization of the biological and molecular characteristics inherent in the comprehensive piRNA dataset spanning 28 different species, as well as its association with various disease systems. The 'piRNAQuest V.2' database aims to enhance our understanding of the origins, abundance, and functions of piRNAs, which will, in turn, help us unlock the full potential of these small ncRNAs while investigating their relevance to various systems.

Chapter Six '*Investigating the role of piRNAs in stem cell derivatives*' elucidates the pivotal role of piRNAs in steering the cellular fate of stem cell derivatives and in gaining a better mechanistic insight into the piRNA mediated regulatory networks within this population that might have a role in stem cell differentiation process.

Chapter Seven '*General Conclusions and Future Perspectives*' is the concluding segment which includes a comprehensive summary of the entire thesis work and the potential prospects for future investigations in the field.

The **Appendix** contains a complete list of my publications along with my additional publications and curriculum vitae.

CHAPTER 1| A Brief Review

1.1 Introduction

For a very long time, RNA had been considered as an intermediary element for passing the genetic information to functional protein only [1]. But with advanced studies, the new RNA world has opened avenues focusing on the noncoding transcripts which were previously regarded as junks [2]. It has been found that more than 90% of mature RNAs do not code for protein, and are known as non coding RNAs (ncRNAs). Since the discovery of ribosomal RNA and transfer RNAs in the late 1950s, ncRNAs have been acknowledged for their biological importance till date **[Figure 1]**. NcRNAs can be categorized as regulatory and housekeeping ncRNAs for normal cell functional maintenance [3]. These molecules are involved in dosage compensation, X–inactivation, imprinting, transcriptional and epigenetic regulation, RNA interference (RNAi) or post-Transcriptional Gene Silencing (PTGS) [4-9].

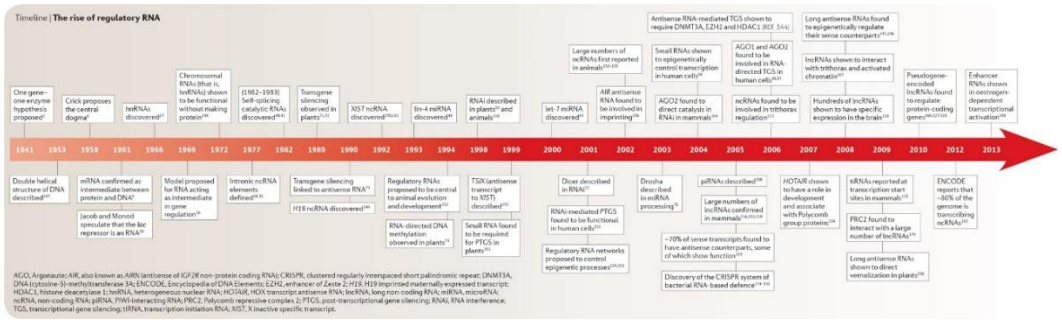


Figure 1: The Rise of Regulatory RNAs : The Timeline [10]

Small ncRNAs constitute one of the important classes of ncRNAs and their size is <200bp. These can be further sub-classified into different classes as provided in **(Table 1)** below.

Class (symbol)	Size	Features	Biological involvement	Ref
MicroRNA (miRNA)	19–25 nucleotides	• Found in viruses, plants, animals, and humans • Single stranded molecules	• Involved in cell growth, differentiation and apoptosis • Regulate gene post transcriptionally • Biomarker for many diseases	[11, 12]

Piwi interacting RNA (piRNA)	24-32 nucleotides	• Found in insects, amphibians, mammals; mainly in germ cells	• Silence RNA via the formation of RISC • Involved in embryonic development, epigenetic regulation, maintenance of germline DNA integrity	[13]
Small Interfering RNA (siRNA)	19-23 nucleotides	• Found in insects and mammals • Double-stranded RNA molecules	• Involved in posttranscriptional gene silencing • Provide defense against pathogenic nucleic acids	[14]
Transfer derived RNAs (tRFs)	14-32 nucleotides	• Found in prokaryotic and eukaryotic transcriptomes • Originated from mature tRNAs or precursor tRNAs	• Regulate gene silencing, splicing and translation • Some tRFs are involved in posttranscriptional regulation	[15, 16]
Small nucleolar RNAs (snorRNAs)	60-300 nucleotides	• Present in nucleoli of eukaryotic cell • Encoded by intronic regions	• Can maintain post-transcriptional modification and ribosome biogenesis	[17, 18]
Small nuclear RNAs (snRNAs)	Average of 150 nucleotides	• Found in nucleus of eukaryotic cells	• Process primary transcription products	[19]
Promoter associated small RNAs (PASRs)	20-200 nucleotides	• Found in leaves of Arabidopsis	• Associated with site specific DNA methylation	[20]
Transcription Initiation RNAs (tiRNA)	Approx 18 nucleotides	• Found in sects to mammals, not found in plants	• Regulate translation mechanism and maintain stress granule assembly	[21, 22]
Centromere Repeat associated small Interacting RNAs (crasiRNAs)	34-42 nucleotides	• Generated from transcription of repeats	• Recruit centromeric proteins and heterchromatin	[23]
Telomere specific Small RNAs (tel-sRNAs)	Approx 24 nucleotides	• Found in plants and mammals	• Assemble telomeric heterochromatin	[24]

Pyknons	Minimum 16 nucleotides	• Primate specific motifs • Found in Repeats in intergenic and intronic regions	• Associated with cancer biology, mainly in epithelial-to-mesenchymal transition	[25]

Table 1: The diverse subclasses of small ncRNAs

Among all the small ncRNAs (sncRNAs), the discovery of microRNAs (miRNAs) in the 1990s, started the new episode of gene regulation by ncRNAs and instigated others to reveal more types of sncRNAs like piRNAs, snoRNAs and their functional involvement towards biological regulation with the help of various technical advances **[Figure 2]**.

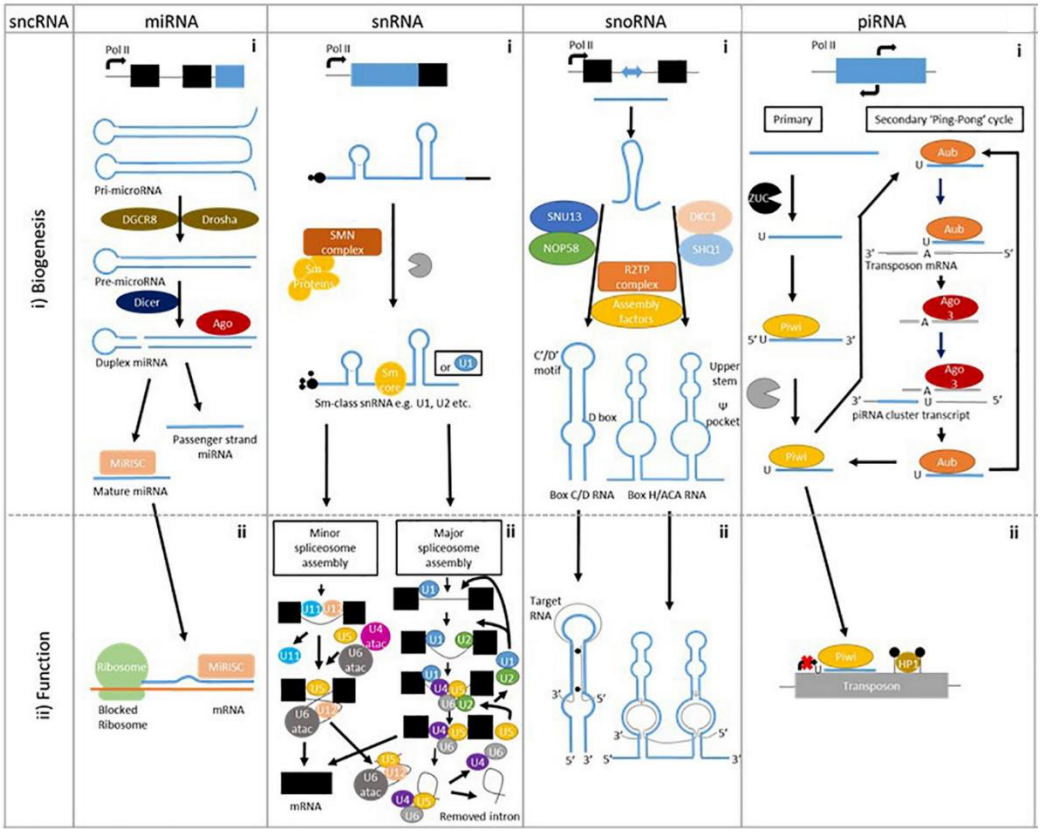


Figure 2: Different types of sncRNAs and their biological association[26]

1.2. miRNAs : The RNA interference triggering molecules

miRNAs are a group of endogenous sncRNAs which are evolutionarily conserved and single stranded molecules, size ranges from 19-25 nts. They are able to regulate gene expression both transcriptionally and translationally. miRNAs interact with partial complementary target region identifying the 'seed sequence' in the 3' untranslated region (UTR) of the messenger RNA (mRNA) [27] and the targeted

gene either got degraded or translationally inhibited by the AGO protein [28]. Biologically, these small molecules have been found to regulate different developmental and disease condition like epithelial-mesenchymal transition and metastasis [29], pluripotency [30], stem cell differentiation [31], diabetes [32], testis differentiation [33], neural plasticity and memory [34].

Historical perspective: In the year 1993, Lee and colleagues discovered that a small 22 nucleotide untranslated RNA mediated down regulation of LIN-14 protein is necessary for developing the larval stage of *Caenorhabditis elegans* [12]. Further, different genetic studies on this nematode along with drosophila and human revealed the existence of small noncoding RNA molecules which are derived from a longer precursor with a stem-loop structure [35]. The identification and confirmation of these small RNAs which are now termed as miRNAs led researchers take both biochemical and computational approach to identify new members of this family across different species, from plants to animals.

Biogenesis: During the early studies on the biogenesis of miRNAs, different approaches were found. After the discovery of the RNase III enzyme termed Drosha in 2003 [36] along with other different studies, it was ascertained that miRNA biogenesis is a sequential and compartmentalized mechanism: (i) In nucleus, RNA polymerase II first produces primary miRNAs (pri-miRNAs) with a 5′ guanosine cap and a 3′ polyadenylated tail (ii) then long precursor RNAs (pre-miRNAs) are generated from pri-miRNA by Drosha and (iii) in cytoplasm pre-miRNAs are processed into mature miRNAs by Dicer [12] **[Figure 3]**. One strand of the mature miRNA which have most unstable base pairing at the 5′ end usually works as the guide strand. Another strand with stable base pairing at the 5′ end is usually degraded [37]. The guiding strand along with cytoplasmic Ago2 proteins forms complex with the target mRNA molecule through sequence complementarity and results in degradation or translational repression.

Functions: miRNAs play a fundamental role in the regulation of gene expression within biological systems. Acting as post-transcriptional regulators, miRNAs bind to the 3' UTR of target mRNAs, leading to translational repression or degradation of the target mRNA. This precise modulation of gene expression by miRNAs is integral to a myriad of biological processes. miRNAs regulate stem cell renewal as well as its proliferation and differentiation which is well illustrated in the study of let-7 processing by Lin-28 during stem cell differentiation [38]. Besides stem cell monitoring, miRNAs are also know to regulate the development of various organ systems with tissue specific miRNAs, e.g., miR-273 and miR-127 are required for neuronal [39] and lung [40] development respectively.

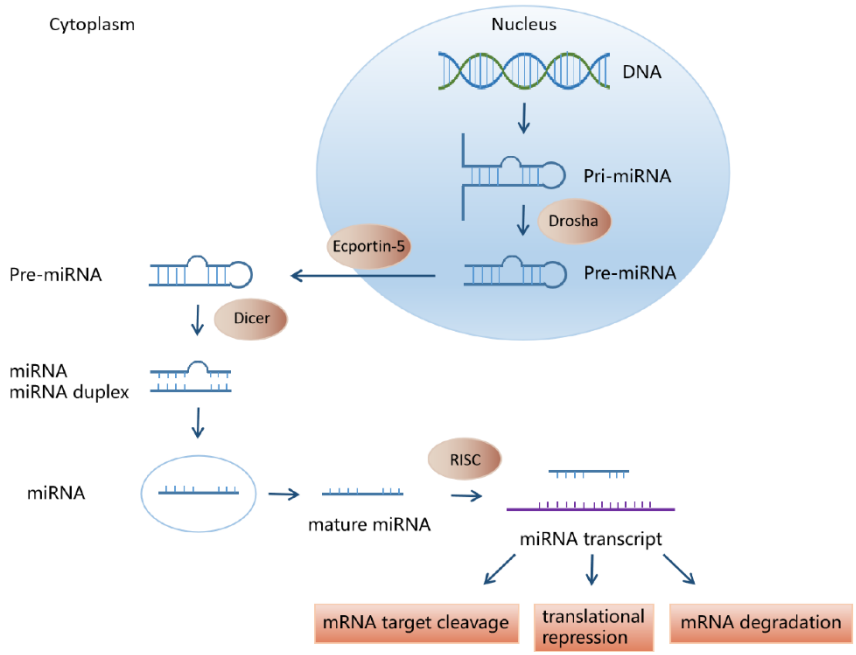


Figure 3: The canonical pathway of miRNA biogenesis [41]

In addition to their involvement in normal cellular functions, dysregulation of miRNA expression has been associated with numerous pathological conditions, including cancer [42], neurodegenerative disorders [43] and cardiovascular diseases [44]. High or low expression of specific miRNAs has been reported to correlate with wide range of diseases [12] which make these micromanagers as biomarkers [45]. The versatility of miRNAs lies in their ability to target multiple genes simultaneously, forming intricate regulatory networks that contribute to the maintenance of cellular homeostasis. The dynamic interplay between miRNAs and their target genes underscores the complexity of biological systems, highlighting the importance of miRNA function in fine-tuning gene expression and shaping the molecular landscape of living organisms.

Computational tools and Databases: The rapid progress in technology, particularly high throughput sequencing, has generated a wealth of miRNA-related data, sparking intense interest and the creation of various bioinformatics tools. These tools cater to diverse biological inquiries, encompassing miRNA identification, target prediction, expression analysis, functional involvement, pathway exploration, and disease associations. miRBase [46,47] serves as a vital repository for miRNA sequences, while algorithms like RNAfold [48] and Mfold [49] predict miRNA secondary structures. To gauge miRNA-target interactions, prediction tools consider factors like free energy, target site accessibility, evolutionary conservation, and compatibility of miRNA seed sequences with target mRNAs [50].

A schematic overview of currently available databases and tools for miRNA is presented in **Figure 4**.

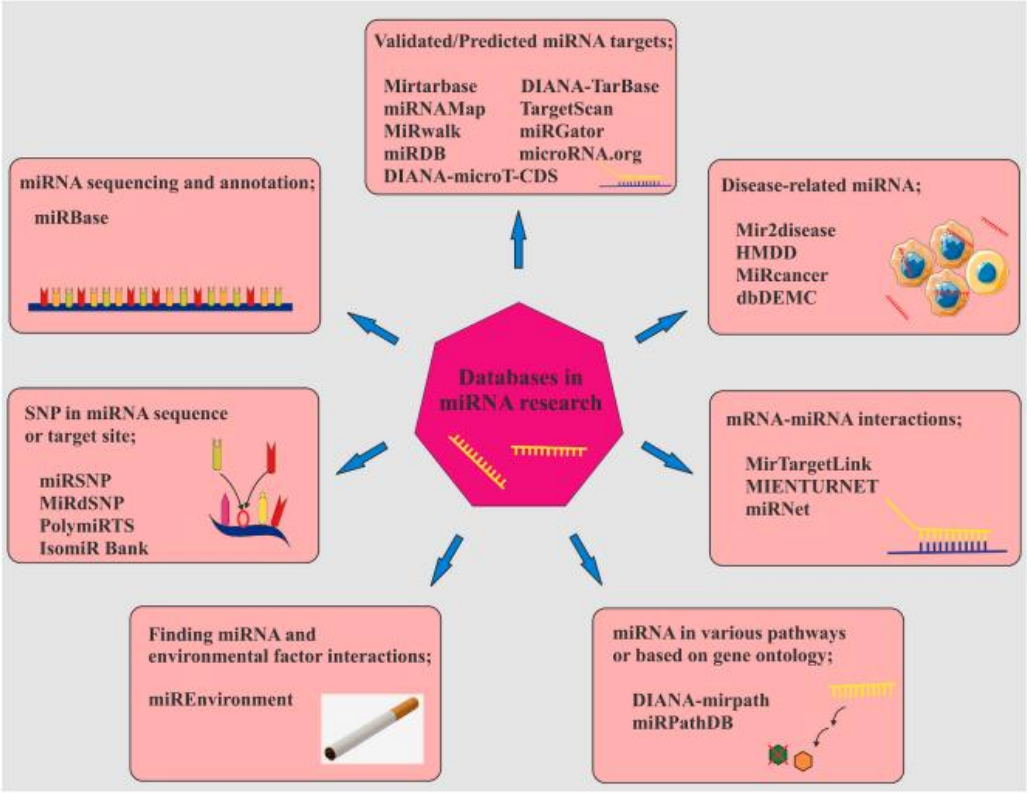


Figure 4: Databases and tools available for miRNA research [51]

Research over the past thirty years has established the importance of miRNAs indicating their potential as disease therapeutics. A better understanding of miRNA molecular pathways would provide valuable insight into the onset and progression of cancer and other diseases and would be significant for progressing therapeutic treatments.

1.3. piRNAs : The genomic players

Piwi-interacting RNAs (piRNAs) which are identified recently as a class of sncRNAs, are abundantly produced in the germline cells of eukaryotes with size ranging between 24–32 nucleotides (nt) [52]. It has been observed that each species contain about hundreds of thousands of distinct piRNA sequences, but mature piRNA sequences are not conserved even among closely related species [53]. Although initially piRNAs have been observed to be expressed exclusively in germline cells, they have been found to be expressed in somatic cells and the somatic piRNA pathway has been shown to regulate the propagation of germline cells [54]. piRNAs bind to piwi proteins to form the piwi-interacting RNA complex (piRC), which influences spermiogenesis, transposon silencing, epigenetic regulation,

genome rearrangement, protein regulation, and germ stem-cell maintenance, all of which are now recognized as major functions of piRNAs [55].

Historical perspective: As small silencing RNAs with characteristics different from those of known miRNAs or siRNAs, piRNAs were first discovered in Drosophila in 2001 [56]. Initially, these RNAs were only observed in tissues related to male and female reproduction. Later studies conducted in flies, fish, and mammals revealed a conserved association of these small RNAs with PIWI-clade Argonaute proteins, which led to the designation of these regulating players [56-59].

Biogenesis: In a number of studies, the previously unknown mechanisms of piRNA biogenesis have begun to be uncovered. These studies suggest that piRNAs are produced from both the primary processing pathway and the amplifying ping-pong mechanism **[Figure 5].** The primary processing pathway for piRNAs, or piRNA clusters, is also referred to as the hotspots for piRNA biogenesis [60]. piRNAs originate from clusters, which take up a large portion of our genome and encode thousands of piRNAs [61]. PIWI proteins are guided by mature piRNAs to target active transposable elements in the genome while also deriving from these transposons (ping-pong cycle) [62]. While active piRNA-induced silencing complex (piRISC) mainly targets transposable elements to maintain the integrity of the genome [63], it is also known to target non-transposable elements, such as protein-coding genes [64]. Similar to miRNAs, these are primarily known to repress or degrade target mRNAs that are specific to the germline [65]. According to a number of studies, piRNAs are an essential regulator for the post-transcriptional and epigenetic silencing of transposons [66].

Functions: In animals, piRNAs have emerged as genomic regulators of biological processes. Initial research has suggested that the PIWI-piRNA complex plays a role in the maintenance and development of the germline, particularly spermatogenesis [67]. Multiple organisms including nematodes, insects, fish, and mammals, have been studied to determine the epigenetic function of PIWI proteins in the control of germ and stem cells and found that piRNAs play important roles in controlling gene expression in addition to acting as transposon silencers in both germline cells and somatic cells [68].

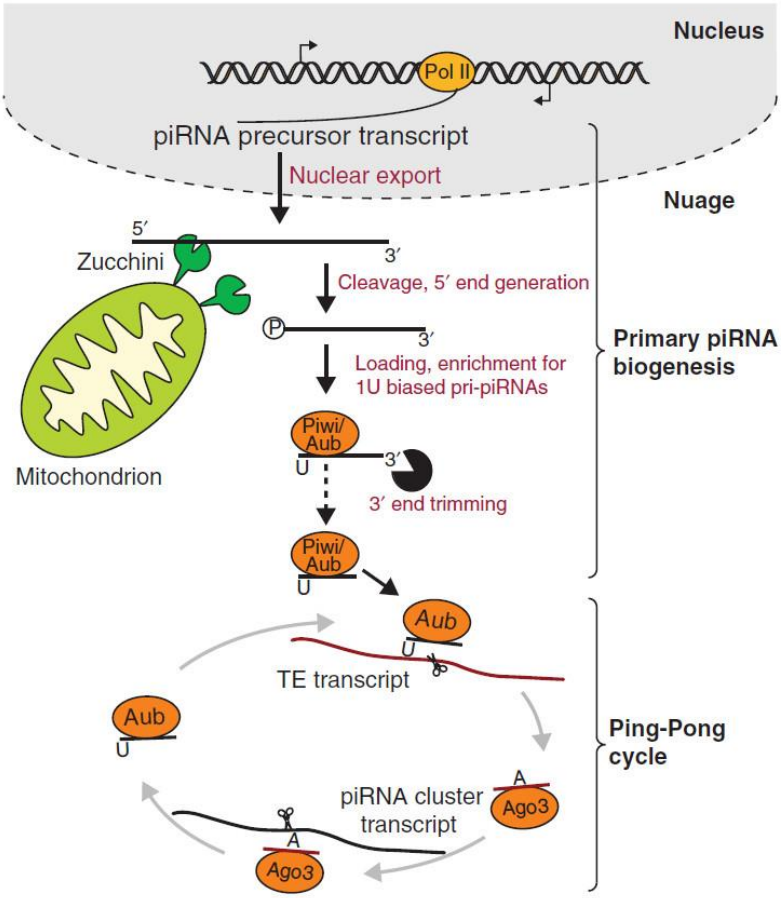


Figure 5: piRNA biogenesis pathways [69]

Understanding the function of piRNAs in human diseases has gained importance as a result of their role in gene regulation. Numerous studies have demonstrated that dysregulation of piRNAs have the potential to either promote or suppress the occurrence and progression of a number of diseases, particularly cancers [70], e.g., piRNA-823 expression is reported to be positively correlated with tumor metastasis [71] whereas piR-36712 has been discovered to suppress breast cancer cell migration, invasion, and proliferation [72].

Computational tools and Databases: In the realm of piRNA research, databases serve as invaluable repositories, housing curated piRNA sequences, annotations, and experimentally-derived information. Some widely used computational databases like piRNABank [73], piRBase [74] and piRNAdb [75] offer a comprehensive approach, providing detailed piRNA annotations, expression profiles, and functional information. Additionally, piRNAQuest [60,76] caters specifically to piRNA researchers, offering a range of piRNA-related data for both normal and disease systems. Databases like piRTarBase [77] and pirScan [78] provide information on

piRNA sequences and their associated targets. These resources empower scientists to explore the intricate world of piRNAs and unravel their significance in epigenetic regulation and genome stability.

1.4. SncRNAs and stem cells

Stem cells are specialized cells within the body characterized by two key attributes: (1) Their capacity for sustained self-renewal through numerous cell division cycle while retaining their undifferentiated state, and (2) their potential to differentiate into progenitor cells representing diverse cell lineages [79,80]. Totipotent cells, such as the zygote, possess the inherent ability to generate all embryonic and extraembryonic cells, enabling them to give rise to an entire organism. Pluripotent stem cells, exemplified by mouse embryonic stem cells (mESCs) or human embryonic stem cells (hESCs), induced pluripotent stem cells, can produce cells representing all three germ layers (ectoderm, mesoderm, and endoderm) but lack the capacity to generate extraembryonic cells [81]. The versatility of pluripotent stem cells to self-renew and differentiate into cells of the three germ layers make them an invaluable tool for regenerative medicine. Adult stem cells, which are multipotent stem cells found in specific tissues like bone marrow, adipose tissue, and blood (including umbilical cord blood) [82], maintain specialization primarily within a particular tissue or lineage, such as hematopoietic stem cells (HSCs) that can only differentiate into cells of the hematopoietic lineage. Unipotent stem cells are further constrained in their developmental potential, giving rise to only a single cell type [83]; for example, erythrocyte progenitor cells exclusively generate erythrocytes. Hence, the canonical developmental pathway **[Figure 6]** follows the progression from totipotent stem cells to pluripotent stem cells, then to multipotent stem cells, and ultimately to unipotent stem cells and mature cells. Notably, both the capacity for self-renewal and the potential for differentiation diminish as cells progress from a totipotent state to a mature cell state.

Human pluripotent stem cells, including both human embryonic and human induced pluripotent stem cells (ESCs and iPSCs), possess the remarkable ability to self-renew and differentiate into cells representing the three germ layers (ectoderm, endoderm and mesoderm). This exceptional property renders them invaluable tools in the field of regenerative medicine. However, these very characteristics also imbue them with oncogenic potential and the capacity to pass on similar traits to their derivative cells. Notably, literature reports have provided evidence of a substantial overlap in gene expression patterns between cancer cells and pluripotent stem cell derivatives [84,85], encompassing iPSC derivatives and ESC derivatives, further showing their oncogenic propensity.

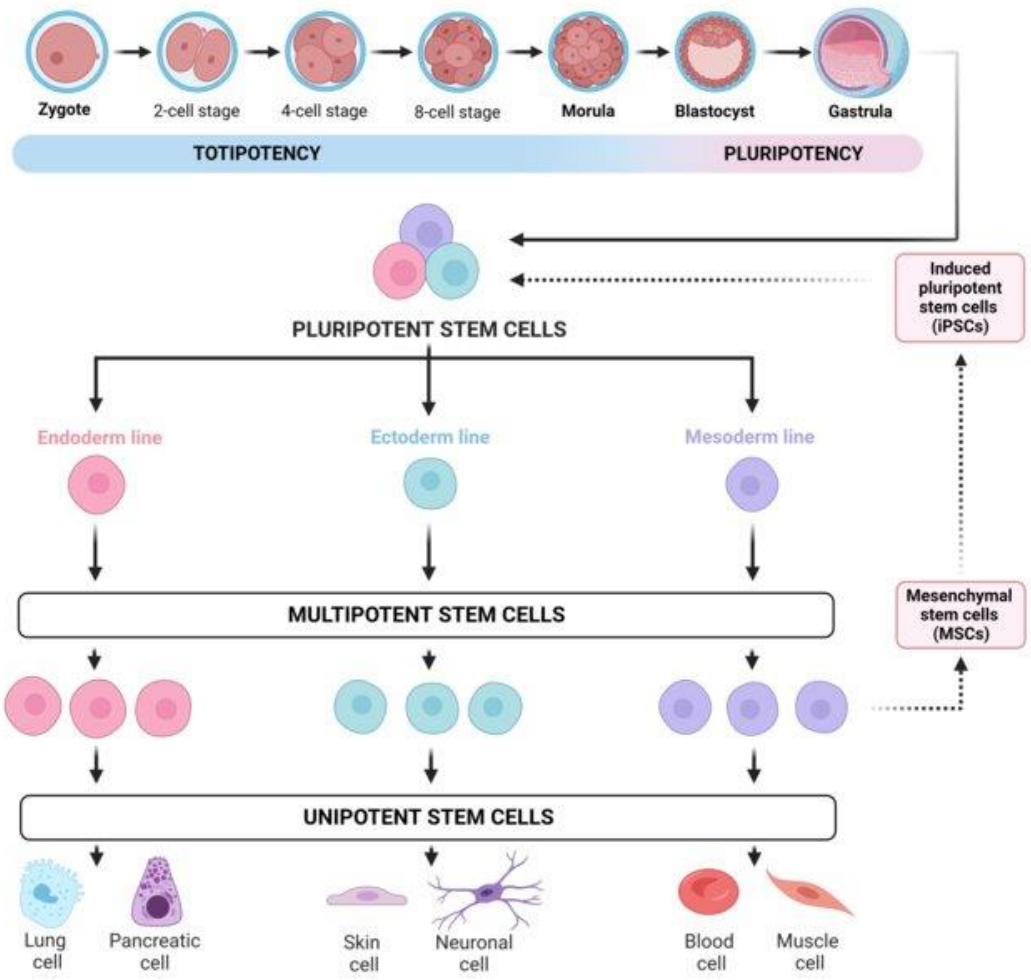


Figure 6 : Cellular potency and Stages of differentiation [86]

Numerous investigations aimed at disrupting the functionality of RNA silencing components provide strong evidence that small RNA pathways play a significant role in regulating cell division, sustaining cellular integrity, and guiding the differentiation processes of stem cells. Hence, sncRNAs are the center of attraction in developmental studies from the last 20 years for their biological and molecular association in cancer development and progression, and also in drug resistance. There are numerous tumor types that have 'stem cell' populations that can support the growth and metastasis of the tumor. Despite making up a very small portion of all tumor cells, these are the ones responsible for drug resistance and tumor recurrence [87]. Beside cancer systems, it has been found that pluripotent stem cells which can be used in regenerative therapy can transmit oncogenic properties within their derivatives which hinder their use in clinical field. The two profound classes of sncRNAs viz. miRNAs and piRNAs, play promising roles in eukaryotic stem cells and stem cell malignancies [88,89].

miRNAs have a significant impact on stem cell maintenance [**Figure 7**] due to their crucial role in regulating gene expression. They are central players in determining whether a stem cell remains undifferentiated or gets differentiated into specialized cell types. Certain miRNAs inhibit the expression of genes associated with differentiation, thereby preserving the stem cell's capacity for self-renewal. miR-302 is a well-known miRNA that plays a important role in maintaining the pluripotent state of ESCs. It target genes are associated with differentiation, aiding in the maintenance of undifferentiated state of ESCs [90]. Conversely, other miRNAs are involved in promoting differentiation by suppressing genes that keep cells in an undifferentiated state [91]. This drives stem cell towards cell lineages. In tissues with regenerative capacities, such as the skin or the gastrointestinal tract, miRNAs are instrumental in orchestrating stem cell-driven repair processes after injury or damage. By modulating miR-126, the regenerative response of endothelial progenitor cells in damaged blood vessels could be achieved [92]. Dysregulation of miRNAs in stem cells has been linked to various diseases and cancers. Aberrant miRNA expression can disrupt normal stem cell function and contribute to the development of malignancies [93].

miRNAs also play a pivotal role in cellular reprogramming, a process that involves converting one type of cell into another, typically to generate iPSCs. The role of miRNAs in cellular reprogramming encompasses several key aspects:

Enhancement of Reprogramming Efficiency: Certain miRNAs can significantly enhance the efficiency of cellular reprogramming [94]. They can act as facilitators by promoting the transition of somatic cells to a pluripotent state, reducing the time required for the reprogramming process.

Regulation of Pluripotency Factors: miRNAs are involved in the fine-tuning of key pluripotency factors such as OCT4, SOX2, and NANOG. These factors are essential for maintaining the pluripotent state and miRNAs help regulate their expression levels during cellular reprogramming [95].

Suppression of Differentiation Pathways: During reprogramming, miRNAs play a crucial role in suppressing the expression of genes associated with the original cell type's identity. This inhibition is necessary to erase the cell's previous characteristics and enable the acquisition of a pluripotent state.

Epigenetic Modulation: miRNAs contribute to the epigenetic changes required for cellular reprogramming. They can influence DNA methylation patterns and histone modifications, promoting the chromatin remodeling necessary for the transition to a pluripotent state [94].

Stabilization of Reprogrammed Cells: miRNAs play a role in stabilizing the reprogrammed state by regulating the gene expression involved in cell cycle control, apoptosis, and differentiation. This helps ensure the long-term stability and viability of iPSCs [96].

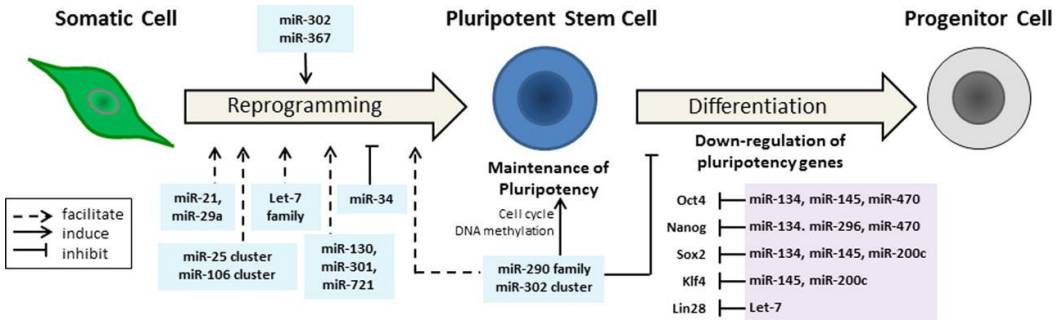


Figure 7 : miRNA in stemness maintenance [97]

In the context of stem cells, piRNAs have also been found to regulate the self-renewal and differentiation processes, ensuring the maintenance of stemness and the fine-tuning of lineage commitment [**Figure 8**] [98,99]. For instance, specific piRNAs have been implicated in the control of pluripotency genes in embryonic stem cells, influencing their ability to differentiate into various cell types. Studies have identified piRNAs that target and regulate key pluripotency genes like Oct4 and Nanog in embryonic stem cells [100]. By modulating the expression of these genes, piRNAs help maintain the undifferentiated state of stem cells, ensuring their pluripotency and self-renewal capacity. In regenerative medicine, understanding and harnessing tissue-specific piRNAs can be instrumental in guiding stem cells towards the desired lineage for tissue repair and regeneration. For example, piRNAs have been investigated for their role in promoting the differentiation of cardiac progenitor cells into functional cardiomyocytes [101]. piRNAs have also been implicated in cancer stem cells, which share some characteristics with normal stem cells. Dysregulation of piRNAs in cancer stem cells can influence tumorigenesis and tumor progression [102]. Exploring piRNA-based interventions in cancer stem cells may have implications for cancer therapies and regenerative approaches aimed at restoring normal tissue function post-cancer treatment.

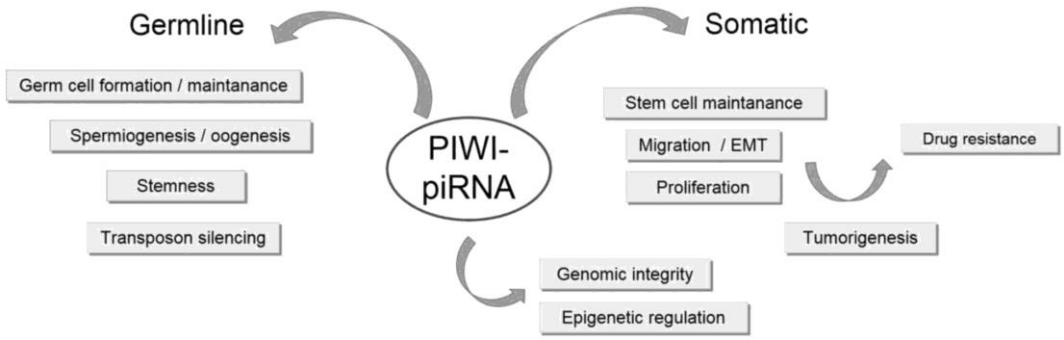


Figure 8 : PIWI-piRNA pathways in cellular functions [103]

Understanding the role of miRNAs and piRNAs in stem cell biology has been gained importance as it might lead towards novel avenues for future therapeutic

interventions. In summary, these sncRNAs play a crucial role in both the maintenance of stem cells and their potential application in regenerative therapies. Manipulating miRNAs and piRNAs hold promise for enhancing the stem cells' therapeutic potential and addressing various diseases and tissue injuries. However, it's essential to continue research into the specific sncRNAs and their targets to develop safe and effective regenerative therapies. Researchers are exploring miRNA/piRNA based strategies to enhance tissue regeneration, combat degenerative diseases, and even target cancer stem cells. Researchers are exploring the use of miRNAs and piRNAs to improve the efficacy of stem cell-based regenerative therapies. By manipulating their expression, it is possible to enhance the survival, differentiation, and integration of transplanted stem cells into damaged tissues.

The work presented in this thesis represents a comprehensive effort to enhance our understanding of the functions of miRNAs and piRNAs. The initial section of the thesis delves into the miRNA and mRNA signatures, which hold the key to unraveling the oncogenic contamination within induced pluripotent stem cell derivatives, irrespective of the three germ layers. In the second part of our work, we scrutinize whether an oncogenic signature is present within iPSC derivatives, irrespective of the specific reprogramming method employed for iPSC generation. Our in-silico analysis reveals a robust oncogenic signature in iPSC-derivatives, causing them to cluster with their cancer cell counterparts rather than their primary cell counterparts, regardless of the reprogramming method used. Combining all the information, we have developed a prediction algorithm that can assist researchers in assessing the purity of their generated iPSC derivatives. The final section of the thesis centers on the characterization of piRNAs emphasizing their role in maintaining stemness as well as differentiation properties. Collectively, this research offers novel insights into these pivotal micromanagers, which modulate critical attributes essential towards maintenance and regulation of stem cell biology.

References:

1. Ling H, Girnita L, Buda O, et al. Non-coding RNAs: the cancer genome dark matter that matters! Clinical chemistry and laboratory medicine. 2017 May 1;55(5):705-714.
2. Nowak R. Mining treasures from 'junk DNA'. Science. 1994 Feb 4;263(5147):608-10.
3. Morey C, Avner P. Employment opportunities for non-coding RNAs. FEBS letters. 2004 Jun 1;567(1):27-34.
4. Brown CJ, Ballabio A, Rupert JL, et al. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature. 1991 Jan 3;349(6304):38-44.
5. Latos PA, Pauler FM, Koerner MV, et al. Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science. 2012 Dec 14;338(6113):1469-72.

6. Rassoulzadegan M, Grandjean V, Gounon P, et al. RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. Nature. 2006 May 25;441(7092):469-74.
7. Wagner KD, Wagner N, Ghanbarian H, et al. RNA induction and inheritance of epigenetic cardiac hypertrophy in the mouse. Developmental cell. 2008 Jun;14(6):962-9.
8. Mattick JS. Non-coding RNAs: the architects of eukaryotic complexity. EMBO reports. 2001 Nov;2(11):986-91.
9. Castel SE, Martienssen RA. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nature reviews Genetics. 2013 Feb;14(2):100-12.
10. Morris KV, Mattick JS. The rise of regulatory RNA. Nature reviews Genetics. 2014 Jun;15(6):423-37.
11. Jang JH, Lee TJ. The role of microRNAs in cell death pathways. Yeungnam University journal of medicine. 2021 Apr;38(2):107-117.
12. Bhaskaran M, Mohan M. MicroRNAs: history, biogenesis, and their evolving role in animal development and disease. Veterinary pathology. 2014 Jul;51(4):759-74.
13. Huang S, Yoshitake K, Asakawa S. A Review of Discovery Profiling of PIWI-Interacting RNAs and Their Diverse Functions in Metazoans. International journal of molecular sciences. 2021 Oct 16;22(20).
14. Dana H, Chalbatani GM, Mahmoodzadeh H, et al. Molecular Mechanisms and Biological Functions of siRNA. International journal of biomedical science : IJBS. 2017 Jun;13(2):48-57.
15. Kazimierczyk M, Wojnicka M, Biala E, et al. Characteristics of Transfer RNA-Derived Fragments Expressed during Human Renal Cell Development: The Role of Dicer in tRF Biogenesis. International journal of molecular sciences. 2022 Mar 26;23(7).
16. Yu X, Xie Y, Zhang S, et al. tRNA-derived fragments: Mechanisms underlying their regulation of gene expression and potential applications as therapeutic targets in cancers and virus infections. Theranostics. 2021;11(1):461-469.
17. Huang ZH, Du YP, Wen JT, et al. snoRNAs: functions and mechanisms in biological processes, and roles in tumor pathophysiology. Cell death discovery. 2022 May 12;8(1):259.
18. Liu X, Xie W, Meng S, et al. Small Nucleolar RNAs and Their Comprehensive Biological Functions in Hepatocellular Carcinoma. Cells-Basel. 2022 Aug 26;11(17).
19. Hull R. Chapter 8 - Origins and Evolution of Plant Viruses. In: Hull R, editor. Plant Virology (Fifth Edition). Boston: Academic Press; 2014. p. 423-476.
20. Ma X, Han N, Shao C, et al. Transcriptome-Wide Discovery of PASRs (Promoter-Associated Small RNAs) and TASRs (Terminus-Associated Small RNAs) in Arabidopsis thaliana. PloS one. 2017;12(1):e0169212.
21. Taft RJ, Kaplan CD, Simons C, et al. Evolution, biogenesis and function of promoter-associated RNAs. Cell cycle. 2009 Aug;8(15):2332-8.
22. Ivanov P, Emara MM, Villen J, et al. Angiogenin-induced tRNA fragments inhibit translation initiation. Molecular cell. 2011 Aug 19;43(4):613-23.
23. Lindsay J, Carone DM, Brown J, et al. Unique small RNA signatures uncovered in the tammar wallaby genome. BMC genomics. 2012 Oct 17;13:559.
24. Radion E, Morgunova V, Ryazansky S, et al. Key role of piRNAs in telomeric chromatin maintenance and telomere nuclear positioning in Drosophila germline. Epigenetics & chromatin. 2018 Jul 12;11(1):40.
25. Evangelista AF, de Menezes WP, Berardinelli GN, et al. Pyknon-Containing Transcripts Are Downregulated in Colorectal Cancer Tumors, and Loss of PYK44 Is Associated With Worse Patient Outcome. Frontiers in genetics. 2020;11:581454.
26. Watson CN, Belli A, Di Pietro V. Small Non-coding RNAs: New Class of Biomarkers and Potential Therapeutic Targets in Neurodegenerative Disease. Frontiers in genetics. 2019;10:364.

27. Gulyaeva LF, Kushlinskiy NE. Regulatory mechanisms of microRNA expression. Journal of translational medicine. 2016 May 20;14(1):143.
28. Eulalio A, Huntzinger E, Izaurralde E. Getting to the root of miRNA-mediated gene silencing. Cell. 2008 Jan 11;132(1):9-14.
29. Bracken CP, Gregory PA, Khew-Goodall Y, et al. The role of microRNAs in metastasis and epithelial-mesenchymal transition. Cellular and molecular life sciences : CMLS. 2009 May;66(10):1682-99.
30. Leonardo TR, Schultheisz HL, Loring JF, et al. The functions of microRNAs in pluripotency and reprogramming. Nature cell biology. 2012 Nov;14(11):1114-21.
31. Korpal M, Kang Y. The emerging role of miR-200 family of microRNAs in epithelial-mesenchymal transition and cancer metastasis. RNA biology. 2008 Jul-Sep;5(3):115-9.
32. Fernandez-Valverde SL, Taft RJ, Mattick JS. MicroRNAs in beta-cell biology, insulin resistance, diabetes and its complications. Diabetes. 2011 Jul;60(7):1825-31.
33. Rakoczy J, Fernandez-Valverde SL, Glazov EA, et al. MicroRNAs-140-5p/140-3p modulate Leydig cell numbers in the developing mouse testis. Biology of reproduction. 2013 Jun;88(6):143.
34. Bredy TW, Lin Q, Wei W, et al. MicroRNA regulation of neural plasticity and memory. Neurobiology of learning and memory. 2011 Jul;96(1):89-94.
35. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004 Jan 23;116(2):281-97.
36. Lee Y, Ahn C, Han J, et al. The nuclear RNase III Drosha initiates microRNA processing. Nature. 2003 Sep 25;425(6956):415-9.
37. Okamura K, Liu N, Lai EC. Distinct mechanisms for microRNA strand selection by Drosophila Argonautes. Molecular cell. 2009 Nov 13;36(3):431-44.
38. Viswanathan SR, Daley GQ, Gregory RI. Selective blockade of microRNA processing by Lin28. Science. 2008 Apr 4;320(5872):97-100.
39. Hobert O. Architecture of a microRNA-controlled gene regulatory network that diversifies neuronal cell fates. Cold Spring Harbor symposia on quantitative biology. 2006;71:181-8.
40. Bhaskaran M, Wang Y, Zhang H, et al. MicroRNA-127 modulates fetal lung development. Physiological genomics. 2009 May 13;37(3):268-78.
41. Liu J, Zhou F, Guan Y, et al. The Biogenesis of miRNAs and Their Role in the Development of Amyotrophic Lateral Sclerosis. Cells-Basel. 2022 Feb 7;11(3).
42. Zhang B, Pan X, Wang Q, et al. Computational identification of microRNAs and their targets. Computational biology and chemistry. 2006 Dec;30(6):395-407.
43. Li S, Lei Z, Sun T. The role of microRNAs in neurodegenerative diseases: a review. Cell biology and toxicology. 2023 Feb;39(1):53-83.
44. Zhou SS, Jin JP, Wang JQ, et al. miRNAS in cardiovascular diseases: potential biomarkers, therapeutic targets and challenges. Acta pharmacologica Sinica. 2018 Jul;39(7):1073-1084.
45. Condrat CE, Thompson DC, Barbu MG, et al. miRNAs as Biomarkers in Disease: Latest Findings Regarding Their Role in Diagnosis and Prognosis. Cells-Basel. 2020 Jan 23;9(2).
46. Griffiths-Jones S, Saini HK, van Dongen S, et al. miRBase: tools for microRNA genomics. Nucleic acids research. 2008 Jan;36(Database issue):D154-8.
47. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research. 2014 Jan;42(Database issue):D68-73.
48. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, et al. ViennaRNA Package 2.0. Algorithms for molecular biology : AMB. 2011 Nov 24;6:26.
49. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research. 2003 Jul 1;31(13):3406-15.

50. Peterson SM, Thompson JA, Ufkin ML, et al. Common features of microRNA target prediction tools. Frontiers in genetics. 2014;5:23.
51. Mortazavi SS, Bahmanpour Z, Daneshmandpour Y, et al. An updated overview and classification of bioinformatics tools for MicroRNA analysis, which one to choose? Computers in biology and medicine. 2021 2021/07/01/;134:104544.
52. Zuo L, Wang Z, Tan Y, et al. piRNAs and Their Functions in the Brain. International journal of human genetics. 2016 Mar-Jun;16(1-2):53-60.
53. Mani SR, Juliano CE. Untangling the web: the diverse functions of the PIWI/piRNA pathway. Molecular reproduction and development. 2013 Aug;80(8):632-64.
54. Barckmann B, El-Barouk M, Pelisson A, et al. The somatic piRNA pathway controls germline transposition over generations. Nucleic acids research. 2018 Oct 12;46(18):9524-9536.
55. Han YN, Li Y, Xia SQ, et al. PIWI Proteins and PIWI-Interacting RNA: Emerging Roles in Cancer. Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology. 2017;44(1):1-20.
56. Aravin A, Gaidatzis D, Pfeffer S, et al. A novel class of small RNAs bind to MILI protein in mouse testes. Nature. 2006 Jul 13;442(7099):203-7.
57. Houwing S, Kamminga LM, Berezikov E, et al. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell. 2007 Apr 6;129(1):69-82.
58. Girard A, Sachidanandam R, Hannon GJ, et al. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature. 2006 Jul 13;442(7099):199-202.
59. Lau NC, Seto AG, Kim J, et al. Characterization of the piRNA complex from rat testes. Science. 2006 Jul 21;313(5785):363-7.
60. Ghosh B, Sarkar A, Mondal S, et al. piRNAQuest V.2: an updated resource for searching through the piRNAome of multiple species. RNA biology. 2022;19(1):12-25.
61. Siomi MC, Sato K, Pezic D, et al. PIWI-interacting small RNAs: the vanguard of genome defence. Nature reviews Molecular cell biology. 2011 Apr;12(4):246-58.
62. Czech B, Hannon GJ. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. Trends in biochemical sciences. 2016 Apr;41(4):324-337.
63. Moyano M, Stefani G. piRNA involvement in genome stability and human cancer. Journal of hematology & oncology. 2015 Apr 21;8:38.
64. Zhang P, Kang JY, Gou LT, et al. MIWI and piRNA-mediated cleavage of messenger RNAs in mouse testes. Cell research. 2015 Feb;25(2):193-207.
65. Post C, Clark JP, Sytnikova YA, et al. The capacity of target silencing by Drosophila PIWI and piRNAs. Rna. 2014 Dec;20(12):1977-86.
66. Watanabe T, Lin H. Posttranscriptional regulation of gene expression by Piwi proteins and piRNAs. Molecular cell. 2014 Oct 2;56(1):18-27.
67. Thomson T, Lin H. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. Annual review of cell and developmental biology. 2009;25:355-76.
68. Wu X, Pan Y, Fang Y, et al. The Biogenesis and Functions of piRNAs in Human Diseases. Molecular therapy Nucleic acids. 2020 Sep 4;21:108-120.
69. Le Thomas A, Toth KF, Aravin AA. To be or not to be a piRNA: genomic origin and processing of piRNAs. Genome biology. 2014 Jan 27;15(1):204.
70. Liu Y, Zhang J, Li A, et al. Identification of PIWI-interacting RNA modules by weighted correlation network analysis. Cluster Computing. 2019 2019/01/01;22(1):707-717.
71. Huang G, Hu H, Xue X, et al. Altered expression of piRNAs and their relation with clinicopathologic features of breast cancer. Clinical & translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico. 2013 Jul;15(7):563-8.
72. Tan L, Mai D, Zhang B, et al. PIWI-interacting RNA-36712 restrains breast cancer progression and chemoresistance by interaction with SEPW1 pseudogene SEPW1P RNA. Molecular cancer. 2019 Jan 12;18(1):9.

73. Sai Lakshmi S, Agrawal S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic acids research. 2008 Jan;36(Database issue):D173-7.
74. Wang J, Shi Y, Zhou H, et al. piRBase: integrating piRNA annotation in all aspects. Nucleic acids research. 2022 Jan 7;50(D1):D265-D272.
75. Ricardo P, Pedro AFG. piRNAdb: A piwi-interacting RNA database. bioRxiv. 2021:2021.09.21.461238.
76. Sarkar A, Maji RK, Saha S, et al. piRNAQuest: searching the piRNAome for silencers. BMC genomics. 2014 Jul 4;15:555.
77. Wu WS, Brown JS, Chen TT, et al. piRTarBase: a database of piRNA targeting sites and their roles in gene regulation. Nucleic acids research. 2019 Jan 8;47(D1):D181-D187.
78. Wu WS, Huang WC, Brown JS, et al. pirScan: a webserver to predict piRNA targeting sites and to avoid transgene silencing in C. elegans. Nucleic acids research. 2018 Jul 2;46(W1):W43-W48.
79. International Stem Cell I, Adewumi O, Aflatoonian B, et al. Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. Nature biotechnology. 2007 Jul;25(7):803-16.
80. Mitalipov S, Wolf D. Totipotency, pluripotency and nuclear reprogramming. Advances in biochemical engineering/biotechnology. 2009;114:185-99.
81. Jiang Y, Jahagirdar BN, Reinhardt RL, et al. Pluripotency of mesenchymal stem cells derived from adult marrow. Nature. 2002 Jul 4;418(6893):41-9.
82. Gimble JM, Katz AJ, Bunnell BA. Adipose-derived stem cells for regenerative medicine. Circulation research. 2007 May 11;100(9):1249-60.
83. Singh VK, Saini A, Kalsan M, et al. Describing the Stem Cell Potency: The Various Methods of Functional Assessment and In silico Diagnostics. Frontiers in cell and developmental biology. 2016;4:134.
84. Ghosh Z, Huang M, Hu S, et al. Dissecting the oncogenic and tumorigenic potential of differentiated human induced pluripotent stem cells and human embryonic stem cells. Cancer research. 2011 Jul 15;71(14):5030-9.
85. Riggs JW, Barrilleaux BL, Varlakhanova N, et al. Induced pluripotency and oncogenic transformation are related processes. Stem cells and development. 2013 Jan 1;22(1):37-50.
86. Erceg Ivkosic I, Fures R, Cosic V, et al. Unlocking the Potential of Mesenchymal Stem Cells in Gynecology: Where Are We Now? Journal of personalized medicine. 2023 Aug 13;13(8).
87. Visvader JE, Lindeman GJ. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. Nature reviews Cancer. 2008 Oct;8(10):755-68.
88. Garg M. MicroRNAs, stem cells and cancer stem cells. World journal of stem cells. 2012 Jul 26;4(7):62-70.
89. Garcia-Borja E, Siegl F, Mateu R, et al. Critical appraisal of the piRNA-PIWI axis in cancer and cancer stem cells. Biomarker research. 2024 Feb 1;12(1):15.
90. Balzano F, Cruciani S, Basoli V, et al. MiR200 and miR302: Two Big Families Influencing Stem Cell Behavior. Molecules. 2018 Jan 30;23(2).
91. Greve TS, Judson RL, Blelloch R. microRNA control of mouse and human pluripotent stem cell behavior. Annual review of cell and developmental biology. 2013;29:213-239.
92. Zhang Y, Xu Y, Zhou K, et al. MicroRNA-126 and VEGF enhance the function of endothelial progenitor cells in acute myocardial infarction. Experimental and therapeutic medicine. 2022 Feb;23(2):142.
93. Yoshida K, Yamamoto Y, Ochiya T. miRNA signaling networks in cancer stem cells. Regenerative therapy. 2021 Jun;17:1-7.
94. Pascale E, Caiazza C, Paladino M, et al. MicroRNA Roles in Cell Reprogramming Mechanisms. Cells-Basel. 2022 Mar 10;11(6).

95. Mathieu J, Ruohola-Baker H. Regulation of stem cell populations by microRNAs. Advances in experimental medicine and biology. 2013;786:329-51.
96. Li N, Long B, Han W, et al. microRNAs: important regulators of stem cells. Stem cell research & therapy. 2017 May 11;8(1):110.
97. Heinrich EM, Dimmeler S. MicroRNAs and stem cells: control of pluripotency, reprogramming, and lineage commitment. Circulation research. 2012 Mar 30;110(7):1014-22.
98. Rojas-Rios P, Chartier A, Pierson S, et al. Aubergine and piRNAs promote germline stem cell self-renewal by repressing the proto-oncogene Cbl. The EMBO journal. 2017 Nov 2;36(21):3194-3211.
99. Rojas-Rios P, Simonelig M. piRNAs and PIWI proteins: regulators of gene expression in development and stem cells. Development. 2018 Sep 7;145(17).
100. Ding X, Li Y, Lu J, et al. piRNA-823 Is Involved in Cancer Stem Cell Regulation Through Altering DNA Methylation in Association With Luminal Breast Cancer. Frontiers in cell and developmental biology. 2021;9:641052.
101. La Greca A, Scarafia MA, Hernandez Canas MC, et al. PIWI-interacting RNAs are differentially expressed during cardiac differentiation of human pluripotent stem cells. PloS one. 2020;15(5):e0232715.
102. Cheng Y, Wang Q, Jiang W, et al. Emerging roles of piRNAs in cancer: challenges and prospects. Aging. 2019 Nov 13;11(21):9932-9946.
103. Litwin M, Szczepanska-Buda A, Piotrowska A, et al. The meaning of PIWI proteins in cancer development. Oncology letters. 2017 May;13(5):3354-3362.

CHAPTER 2| Elucidating the role of miRNAs in stem cell derivatives corresponding to the three germ layers

Abstract:

Induced pluripotent stem cell (iPSC) derived cells represent specific cell fate whose regenerative potential gets compromised due to oncogenic contamination within them. Epigenetic disturbances are believed to underlie this dysregulation in certain cell fates. Reported role of microRNAs (miRNAs) towards imparting oncogenicity within iPSCs instigated us to look into its role towards inducing oncogenicity within the iPSC derived cells corresponding to the three germ layers viz. the ectoderm, endoderm, and mesoderm. A comprehensive analysis of gene and miRNA expression profiles corresponding to these derivatives, their cancerous counterparts, and primary cells, lead us to unveil an "oncogenic signature" present within the iPSC derivatives corresponding to all the three germ layers. Furthermore, we elucidated the key miRNA-mRNA interactions responsible for such oncogenic contamination. Overall, our study revealed the presence of oncogenic contamination within the iPSC-derived cells corresponding to all the 3 germ layers and deciphered the role of a set of miRNA-mRNA interactions for it. This provides a roadmap towards designing certain stringent functional assays to ensure the status of these cells before applying them for regenerative therapy.

Presented this work at 4th International Conference on Translational Research: Recent Development and Innovations in Human Health and Agriculture Research, 2018: "Revealing the oncogenic signatures induced by miRNAs in iPSC-derivatives"

2.1. Introduction

Induced Pluripotent Stem Cells (iPSCs) represent a revolutionary breakthrough in the field of regenerative medicine. iPSCs are artificially reprogrammed cells that possess a remarkable ability to differentiate into various cell types within the human body [1] and that too without any ethical concerns and minimizing the risk of immune rejection. Originally derived from adult cells, such as blood or skin cells, iPSCs are genetically manipulated to revert to a pluripotent state, akin to embryonic stem cells. This process allows for the generation of patient-specific cells, holding significant promise for personalized medicine and the treatment of a wide array of diseases. iPSCs not only sidestep the ethical concerns connected with embryonic stem cells (ESCs) but also open new avenues for understanding disease mechanisms and lead to the development of innovative therapeutic approaches **[Figure 1]**. The discovery of iPSCs has spurred groundbreaking advancements in biomedical research, paving the way for the potential regeneration of damaged tissues and organs. In 2006, Takahashi and Yamanaka identified a fundamental set of four transcription factor (TF) genes—SOX2, KLF4, OCT4 and C-MYC, (commonly referred to as OSKM or the Yamanaka factors) for generating iPSCs

[2]. When subcutaneously injected into immunocompromised mice, these iPSCs formed teratomas and contributed to various tissues in developing embryos upon blastocyst injection [2]. However, these initial "first-generation" iPSCs seemed partially reprogrammed and lacked crucial deterministic patterns exhibited by fully pluripotent cells. Subsequent research, including work by Yamanaka's team [3], soon refined these initial findings, producing a next generation of modified iPSCs that more closely resembled ESCs both molecularly and functionally [4,5].

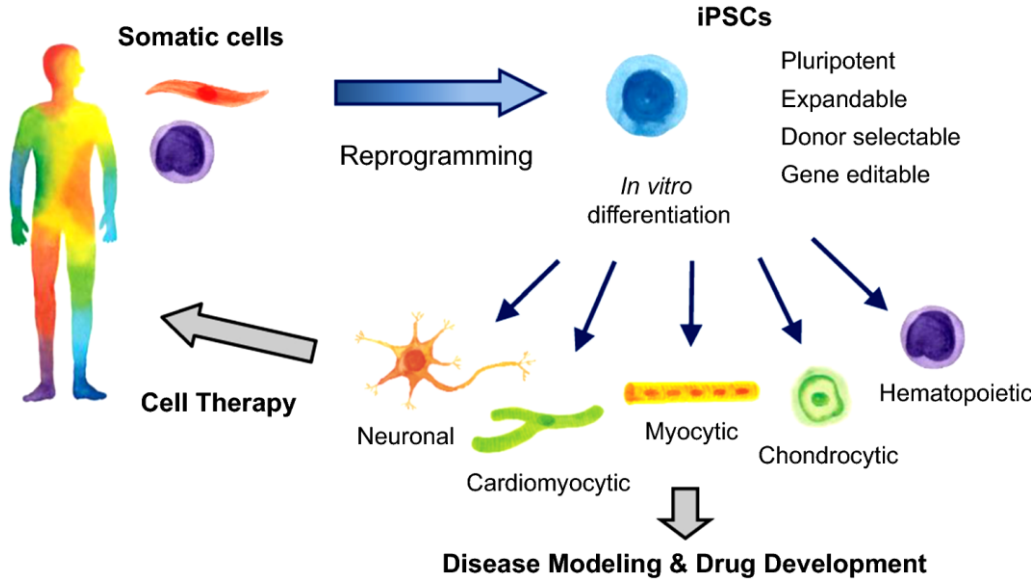


Figure 1: Applications of induced pluripotent stem cells [6]

In the current landscape of stem cell medicine, the journey toward secure and effective regenerative therapies employing human induced pluripotent stem cells (hiPSCs) remains intricate and poses challenges, particularly concerning safety aspects as these cells differentiate and may form teratomas post-transplantation. The overlapping factors for pluripotency and tumorigenicity introduce intricacies into this domain. The activation of functional programs during cellular reprogramming, inducing pluripotency, shares commonalities with analogous processes in cancer cells [7]. Instances arise where inadequately differentiated tumour cells exploit normal developmental programs, including factors linked to ESCs and pertinent pathways. Recent investigations emphasize the presence of shared master regulatory genes in both stem cells and tumor cells [8-11]. As a strategic response, the proposition involves inducing iPSCs to differentiate into the intended cell type corresponding to a particular lineage before transplantation. This tactic aims to mitigate the challenge of tumorigenicity inherent in pluripotent cells.

While the tumorigenic capability of pluripotent cells appears significantly diminished in vivo when subjected to in vitro predifferentiation, reports indicate that the resulting differentiated stem cell derivatives may still possess the potential to form tumors **[Figure 2]**. Various studies offer evidence of oncogenic contamination within the iPSC derived

cells [12] as well as tumor development following the transplantation of iPSC derivatives [13-15]. These findings underscore the potential risk associated with using re-differentiated cells due to the presence of oncogenic contaminants. Therefore, it is imperative to exercise utmost caution in eliminating such tumorigenic cells from the transplantable population before deeming them safe for clinical application. The potential cause of this oncogenicity might be attributed to transcriptional misregulation. miRNA plays a crucial role in the regulation of diverse transcriptional factors. Given the pluripotent nature of iPSCs, capable of differentiating into cells corresponding to the three germ layers, it is essential to examine oncogenic contaminations within pluripotent stem cell derivatives across ectoderm, endoderm, and mesoderm layers.

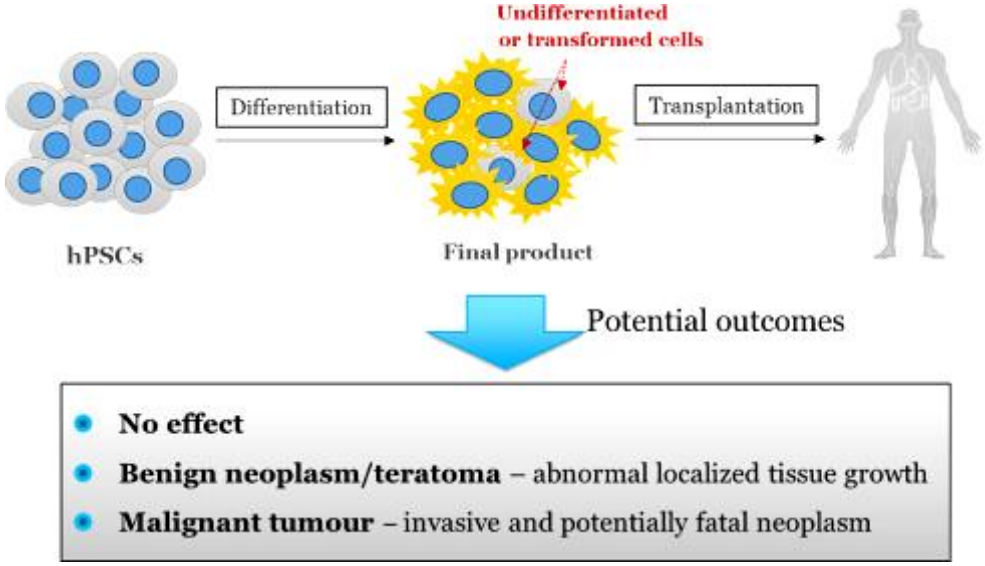


Figure 2: Potential Outcome of PSC-Derived Products after Transplantation [16]

Utilizing this knowledge, we have sought to unveil whether miRNA plays a role towards imparting such oncogenic contamination and if so what are the miRNA-mRNA interactions that play a dominant role in such event irrespective of the cells belonging to the three germ layers. Overall, we have been able to tease apart the significant set of miRNA-mRNA interactions which are playing a key role behind the presence of such residual oncogenicity within the iPSC derived cells irrespective of the 3 germ layers.

2.2. Methods

2.2.1. Dataset collection:

We took miRNA and mRNA expression datasets (microarray) corresponding to iPSC-derivatives, primary cells and its cancer counterpart for all three germ layers (i.e. ectoderm, endoderm and mesoderm), as provided in **Table 1(A, B, C)** from GEO database (https://www.ncbi.nlm.nih.gov/geo/). Based on the availability of datasets, Neuron, Hepatocyte and Cardiomyocyte were chosen for ectodermal, endodermal and mesodermal lineage respectively.

Sample Type	Data Type	iPSC-derived neuron		Neuroblastoma		Primary neuron	
		Accession No.	No. of samples	Accession No.	No. of samples	Accession No.	No. of samples
mRNA	Microarray Data	GSE75701	3	GSE51978	3	GSE12679	6
				GSE13273	2	GSE19332	7
		GSE74358	8	GSE4600	3	GSE40438	8
miRNA		GSE62721	2	GSE94483	6	GSE34016	3

Table 1A: Input dataset for Ectodermal lineage

Sample Type	Data Type	iPSC-derived Hepatocyte		Hepatocarcinoma		Primary Hepatocyte	
		Accession No.	No. of samples	Accession No.	No. of samples	Accession No.	No. of samples
mRNA	Microarray Data	GSE62962	2	GSE18269	3	GSE62962	2
				GSE29084	2		
		GSE14897	3	GSE23031	3	GSE18269	6
miRNA		GSE66075	3	GSE90146	2	GSE40113	4

Table 1B: Input dataset for Endodermal lineage

Sample Type	Data Type	iPSC-derived Cardiomyocytes		Sarcoma		Primary Cardiomyocyte	
		Accession No.	No. of samples	Accession No.	No. of samples	Accession No.	No. of samples
mRNA	Microarray Data	GSE60291	3	GSE59704	2	GSE17800	8
				GSE32911	1	GSE14975	5
miRNA		GSE35672	36	GSE36982	15	GSE36946	20

Table 1C: Input dataset for Mesodermal lineage

2.2.2. Experimental Grouping:

Global gene expression pattern and miRNA profile analysis was done for cells selected corresponding to the three germ layers. Primary cells served as the control, iPSC derivatives served as the test sample 1 and the cancer counterparts for the specific lineage served as the test sample 2 (**Table 2**).

Germ Layer	Tissue	Control	Test Sample 1	Test Sample 2
Ectoderm	Neuron	Primary Neuron (PN)	iPSC-derived Neuron (iPSC_N)	Neuroblastoma (NB)
Endoderm	Hepatocyte	Primary Hepatocyte (PH)	iPSC-derived Hepatocyte (iPSC_H)	Hepatocarcinoma (HCC)
Mesoderm	Cardiomyocyte	Primary Cardiomyocyte (PC)	iPSC-derived Cardiomyocyte (iPSC_C)	Sarcoma (SC)

Table 2: Experimental Grouping of the analysis on the samples obtained for iPSC-derivatives, cancer counterpart and the primary cell corresponding to the germ layer

2.2.3. Microarray data analysis:

To mitigate study-specific batch effects, the raw miRNA and mRNA datasets underwent normalization using the frozen RMA algorithm (fRMA) [17], which offers greater precision compared to standard RMA methods. This approach involves background correction, normalization, and summarization steps similar to other algorithms, while also addressing between-probe and between-batch variability. The fRMA normalization was executed within the R-Bioconductor framework followed by analysis in MEV (www.tm4.org/mev.html). Subsequently, analysis of variance (ANOVA) was performed, incorporating the Benjamini–Hochberg false discovery rate (FDR) multiple testing correction, with a significance level set at a p-value of ≤ 0.05. Additionally, a fold-change cutoff of ≥ 2.0 was applied to identify the differentially expressed (DE) genes and miRNAs between different groups. For mRNA expression data, the average probe intensity was calculated and then considered as the gene expression level for genes with more than one probe. Further, hierarchical cluster analysis was performed with the DE genes and DE miRNAs.

2.2.4. Target Prediction and Functional analysis:

In order to check the effect of miRNA-mRNA interactions, target prediction was performed using TargetScan [18] for negatively correlated sets (DE mRNAs and DE miRNAs corresponding to the three germ layers) using default parameters, following filtration using ParasoR [19], which refines the targeted genes based on accessible regions. Targeted genes were further functionally annotated using Ingenuity Pathway Analysis (IPA, Qiagen).

2.2.5. Screening epigenetically associated genes and miRNA:

Genes known to function as epigenetic modifiers and chromatin remodelers were sourced from EpiFactors [20]. miRNAs with established roles in exerting epigenetic functions were obtained from EpimiR [21].

2.3. Results and Discussion

2.3.1. Oncogenic signature within the ectodermal lineage specific cells derived from iPSCs:

The DE genes and DE miRNAs across the test and control samples have been shown in **Table 3**. Venn diagram in **Figure 3** shows the DE genes and DE miRNAs which are commonly differentially expressed in iPSC_N and its cancerous counterpart NB with the primary cells i.e PN as control. Further hierarchical clustering **[Figure 4]** with this set of genes and miRNAs depicted the clustering of the iPSC_N with its cancerous counterpart NB rather than clustering with its primary counterpart i.e PN.

Control	Test samples	DE mRNAs		DE miRNAs	
		Up	Down	Up	Down
PN	iPSC_N	5527	2225	160	102
PN	NB	4943	2874	132	83

Table 3: Differentially expressed mRNAs and miRNAs across the different samples

Hence, these set of miRNAs and mRNAs bear the clue towards the remanance of such oncogenic contamination within the iPSC derived neurons corresponding to ectodermal lineage.

Figure 3: Venn diagram showing the overlap of the differentially expressed A) mRNAs and B) miRNAs between PN versus iPSC_N & PN versus NB

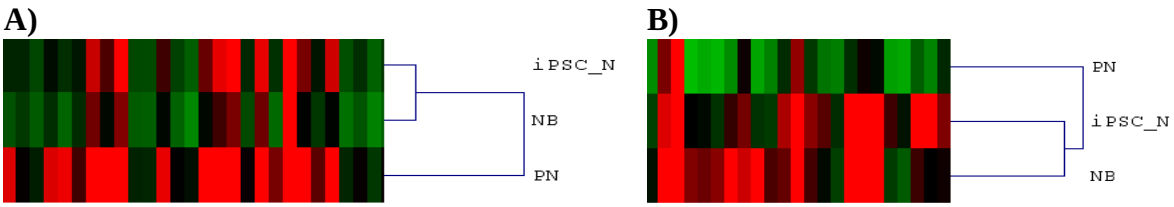


Figure 4: Heatmap showing the expression profile for the commonly expressed A) 5545 mRNAs and B) 128 miRNAs in neuronal lineage corresponding to ectoderm layer. Clustering shows iPSCN to be closest to the cancer counterpart NB

2.3.2. Oncogenic signature within the endodermal lineage specific cells derived from iPSCs:

The DE genes and DE miRNAs across the test and control samples have been depicted in **Table 4**. Venn diagram in **Figure 5** illustrates the DE genes and DE miRNAs which are commonly differentially expressed in iPSC_H and its cancerous counterpart HCC with the primary cells i.e PH as control.

Control	Test samples	DE mRNAs		DE miRNAs	
		Up	Down	Up	Down
PH	iPSC_H	2335	1987	301	62
PH	HCC	2572	1926	290	50

Table 4: Differentially expressed mRNAs and miRNAs across the different samples

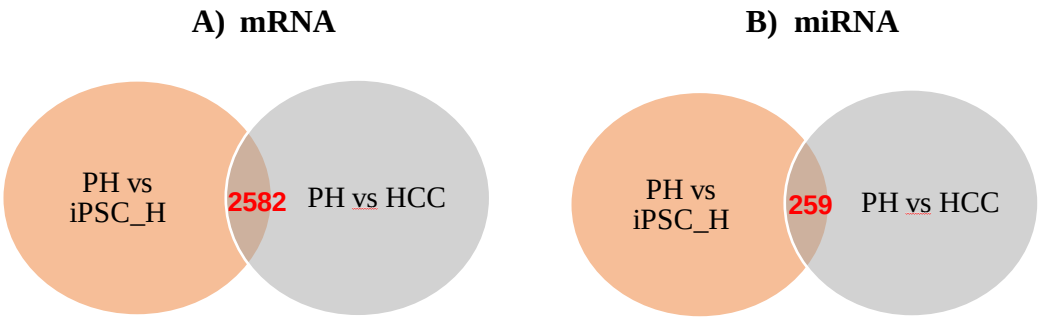


Figure 5: Venn diagram showing the overlap of the differentially expressed A) mRNAs and B) miRNAs between PH versus iPSC_H & PH versus HCC

Hierarchical clustering of iPSC_H, PH and HCC has been done using the expression profile of these common set of mRNAs and miRNAs [**Figure 6A and Figure 6B**]. **Figure 6** showed that iPSC_H clusters with its cancerous counterpart HCC rather than PH (which is the primary hepatocyte) suggesting the presence of oncogenic contamination within iPSC-derivative of endodermal lineage.

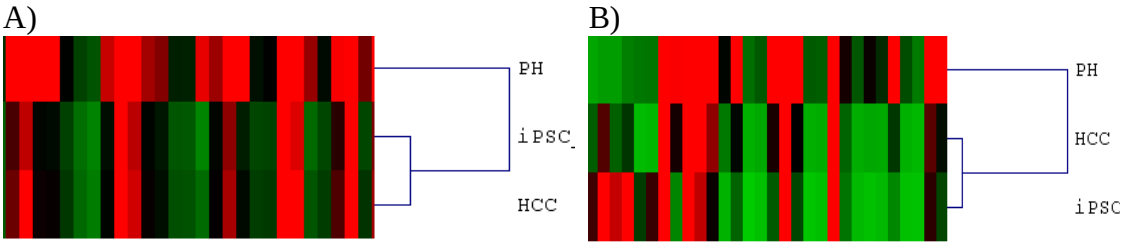


Figure 6: Heatmap showing the expression profile for the commonly expressed A) 2582 mRNAs and B) 259 miRNAs in hepatocyte lineage corresponding to endoderm layer. Clustering shows iPSC_H to be closest to the cancer counterpart HCC

2.3.3. Oncogenic signature within the mesodermal lineage specific cells derived from iPSCs:

The DE genes and DE miRNAs across the test and control samples have been prsented in **Table 5**. Venn diagram in **Figure 7** displays the DE genes and DE miRNAs which are commonly differentially expressed in iPSC_C and its cancerous counterpart SC with the primary cells i.e PC as control. Further hierarchical clustering [**Figure 8**] with this set of genes and miRNAs depicted the clustering of the iPSC_C with its cancerous counterpart SC rather than clustering with its primary counterpart, PC. Hence, these set of miRNAs and mRNAs bear the clue towards the remanance of such oncogenic contamination within the iPSC derived cardiomyocytes corresponding to mesodermal lineage.

Control		DE mRNAs		DE miRNAs	
	Test samples	Up	Down	Up	Down
PC	iPSC_C	3561	2167	360	360
PC	SC	4899	3427	404	320

Table 5: Differentially expressed mRNAs and miRNAs across the different samples

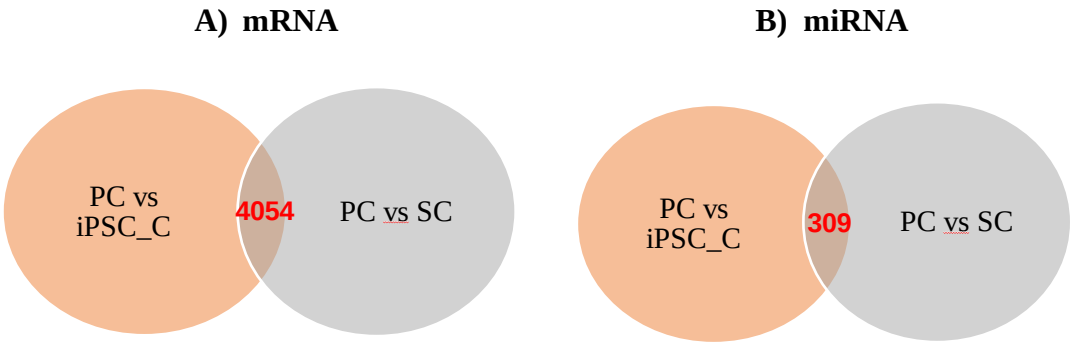


Figure 7: Venn diagram showing the overlap of the differentially expressed A) mRNAs and B) miRNAs between PC versus iPSC_C & PC versus SC

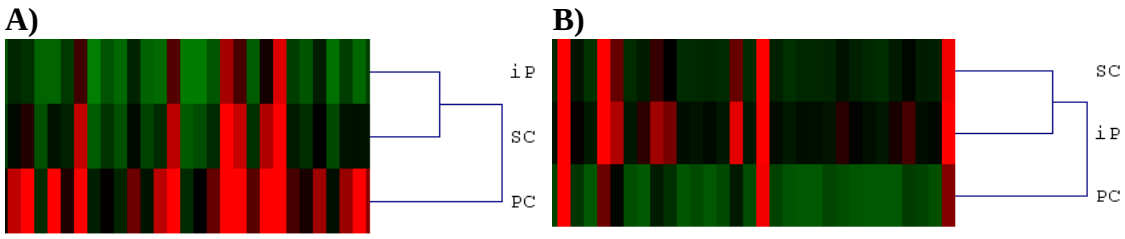


Figure 8: Heatmap showing the expression profile for the commonly expressed A) 4054 mRNAs and B) 309 miRNAs in cardiomyocyte lineage corresponding to mesoderm layer. Clustering shows iPSC_C to be closest to the cancer counterpart SC

2.3.4. Target prediction and pathway analysis with the signature set of miRNAs and genes:

Next, we looked for the significant set of miRNA-mRNA interactions pertaining to such oncogenic contaminations within the iPSC derivatives corresponding to all the 3 germ layers. For this, target prediction (using targetscan) was conducted separately for the negatively correlated miRNA and mRNA gene expression datasets corresponding to each germ layers. Subsequently, ParasoR [19] was utilized to filter the targeted genes based on the presence of accessible regions within the target gene. **Table 6** represents the predicted miRNA-mRNA target pairs comprising of '*oncogenic signature*' specific miRNAs and mRNAs.

Further, the targeted genes of each germ layer underwent rigorous functional and pathway analysis using IPA to find out their involvement in oncogenicity. We observed that a majority of the top significant pathways were associated with carcinogenesis and identified several shared pathways across the three germ layers **[Figure 9]**.

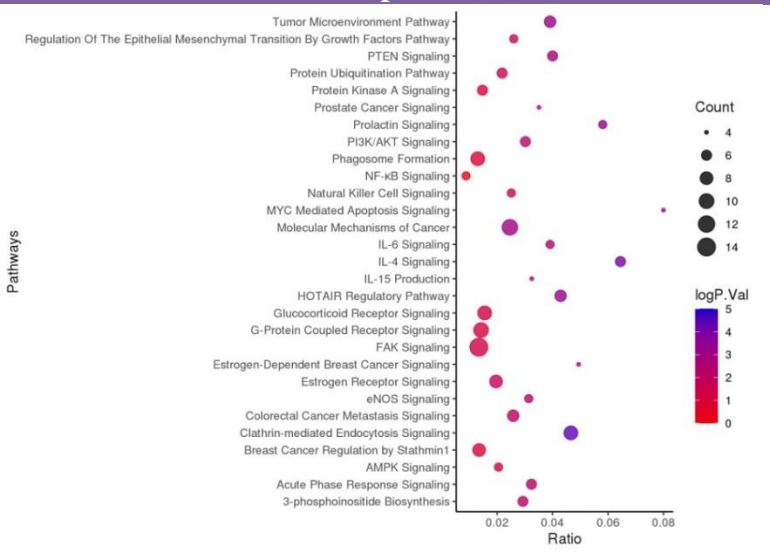
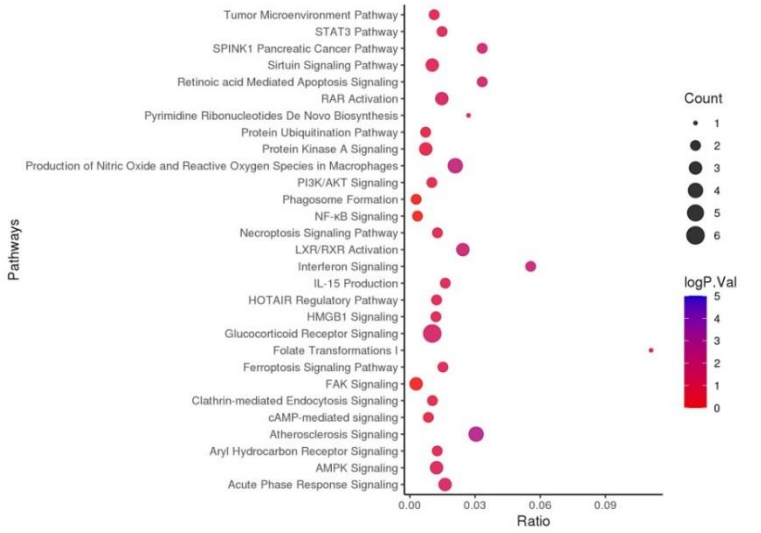
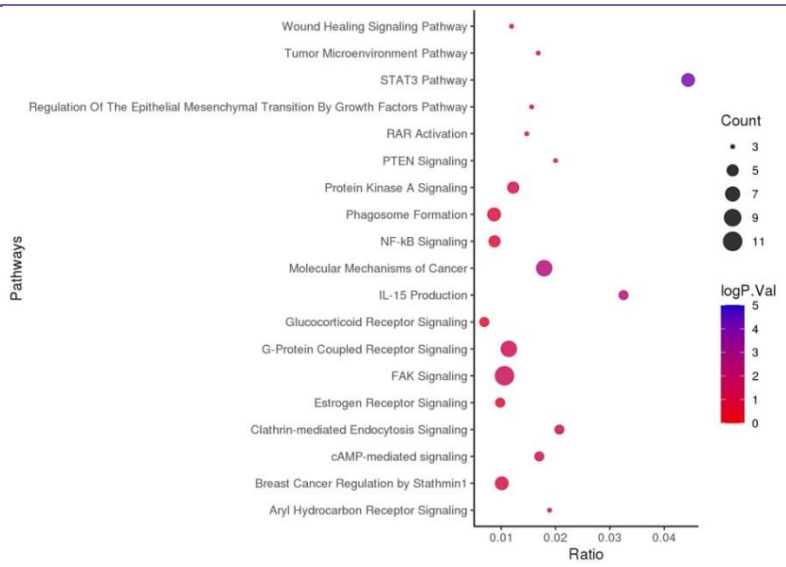
Germ Layer	Dotplot
Ectoderm	
Endoderm	
Mesoderm	

Figure 9: Dot plot Analysis of the miRNA target genes across the Ectoderm, Endoderm and Mesoderm

Subsequently, from pathway analysis, it has been observed that 116 significant pathways were common among the three germ layers [**Figure 10**].

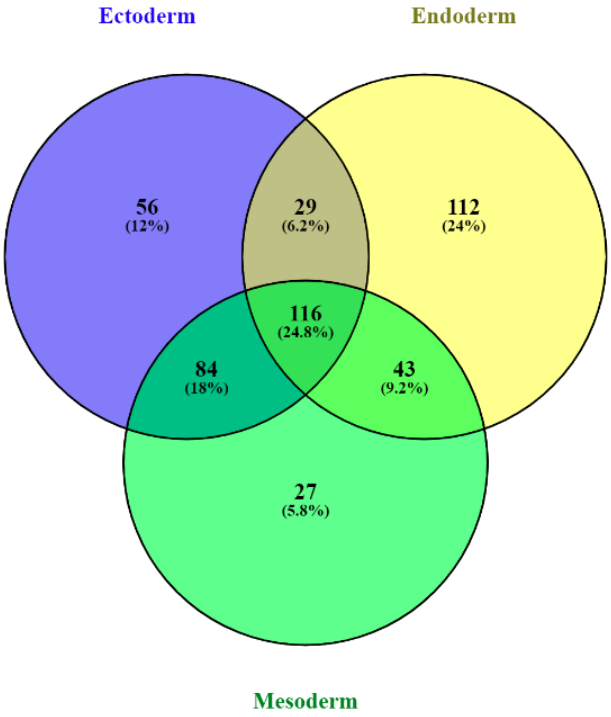


Figure 10: Common pathways among the three germ layers

These 116 pathways were thoroughly investigated to ascertain their association with oncogenicity. We specifically selected those pathways which are directly implicated in general carcinogenesis, excluding those associated with tissue-specific cancers resulting in total 47 pathways. We extracted gene information corresponding to these 47 pathways along with their targeting miRNAs for each germ layer separately, and investigated their relationships with epigenetic modifications and functions. Epigenetic modifications, such as DNA methylation and histone acetylation, can alter gene expression patterns without changing the underlying DNA sequence. Dysregulation of these modifications can lead to aberrant gene expression, disrupting cellular processes and promoting carcinogenesis [22-25]. **Table 6** represents the miRNA-gene target pairs for ectoderm, endoderm and mesoderm which are (both miRNA and the target gene) associated with either histone modification or chromatin remodelling or other type of epigenetic modification.

Germ Layer	Targeting miRNA	Targeted Gene	Epigenetic Function
Ectoderm	hsa-miR-184	PRKCB	Histone modification
	hsa-miR-9-5p	MBD4, MSH6, CBX2, BRCA1, TLK1, UBE2H, USP3	DNA and Histone Modification
	hsa-miR-18a-3p	HDAC11, CBX7	Histone modification
	hsa-miR-375	SP100	Chromatin remodeling cofactor
	hsa-miR-218-5p	GADD45B	Chromatin remodeling
	hsa-miR-146b-5p	ATR	Histone modification write
	hsa-miR-330-5p	HDAC11	Histone modification erase
	hsa-miR-27a-5p	CBX7	Histone modification read
	hsa-miR-425-3p	ARRB1	Histone modification
Endoderm	hsa-miR-196a-5p, hsa-miR-16-5p, hsa-miR-28-5p, hsa-miR-483-5p	ZBTB16	Histone modification erase cofactor
	hsa-miR-126-5p	SP100, PRKAA1, KAT2B	Histone modification write and Chromatin remodelling cofactor
	hsa-miR-18a-3p, hsa-miR-19b-3p	CBX7	Histone modification read
	has-miR-15b-3p, has-miR-20a-3p, has-miR-30e-3p	PRKAA1	Histone modification write
	has-miR-1275, hsa-miR-140-5p	GADD45B	Chromatin remodeling
	hsa-miR-421	PRKAG2	Histone modification write cofactor
Mesoderm	hsa-miR-520e	DNMT3A	DNA modification
	hsa-miR-507	ACTB	Chromatin remodeling cofactor
	hsa-miR-196a-5p	ZBTB16	Histone modification erase cofactor
	hsa-miR-429	HDAC4, NCOR2	Histone modification
	hsa-miR-377-5p	BRCC3, DOT1L, PRMT2	Histone modification
	hsa-miR-638	PRKAB2	Histone modification write cofactor
	hsa-miR-516b-3p	CTBP1	Chromatin remodeling
	hsa-miR-193b-5p	PHF19, MBD1	Histone modification write cofactor
	hsa-miR-618	CHEK1	Histone modification write
	hsa-miR-31-3p	UBE2B	Histone modification write

Table 6: miRNA-mRNA target pairs bearing 'oncogenic signature' and their associated role in epigenetic events

Our analysis unveiled an "oncogenic signature" present within iPSC-derivatives across the three germ layers, suggesting a predisposition to tumorigenicity. Notably, a subset of miRNA-mRNA target pairs bearing this signature is known to influence epigenetic events

which might be the underlying cause of the presence of such remnant oncogenic contamination within these derivatives. Studying the interplay between epigenetic modifications and miRNA/mRNA regulation in iPSC derivatives is paramount for several reasons. Firstly, iPSC-based therapies hold immense promise for regenerative medicine, but concerns surrounding tumorigenicity necessitate rigorous safety assessments. Investigating how these miRNA-induced epigenetic changes influence their target genes is crucial for understanding potential mechanisms of oncogenic contamination. These findings will help to develop strategies to mitigate the risk of tumorigenesis, ensuring the safety and efficacy of iPSC-derived therapies for clinical translation. Moreover, uncovering the regulatory networks governing epigenetic modifications in iPSC derivatives enhances our fundamental understanding of cellular reprogramming and differentiation processes, facilitating the optimization of protocols for generating safer and more reliable iPSC-based therapies. Overall, this study not only addresses critical safety concerns but also contributes to advancing the field of iPSC research towards its clinical applications in regenerative medicine.

2.4. Conclusion

Navigating the path towards safe and effective iPSC-based regenerative therapies presents considerable challenges. Thorough screening to assess the tumorigenic nature of pluripotent stem cells and their derivatives is imperative to ensure their safe utilization in regenerative therapy. Therefore, in this study, we employed an extensive bioinformatic approach to explore miRNA-mRNA target interactions, which possess the potential to trigger oncogenic contamination within iPSC-derived cells corresponding to the three germ layers. We identified a set of miRNA-mRNA target pairs exhibiting an "oncogenic signature." These mRNA targets are known to influence epigenetic events, likely contributing to the occurrence of oncogenic contamination within these derivatives.

This study aimed to outline a roadmap for producing safer iPSC-derivatives. We've uncovered insights into the miRNA-induced epigenetic alterations potentially driving oncogenic transformation within these derivatives across three germ layers. Further functional assays will be crucial for bolstering this conclusion. Overall, our findings offer valuable insights into deciphering the factors governing pluripotency versus tumorigenicity, thereby paving the way for safer regenerative therapies and enhancing the success of iPSC technology in the foreseeable future.

References:

1. Singh VK, Kalsan M, Kumar N, et al. Induced pluripotent stem cells: applications in regenerative medicine, disease modeling, and drug discovery. Frontiers in cell and developmental biology. 2015;3:2.
2. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006 Aug 25;126(4):663-76.

3. Okita K, Ichisaka T, Yamanaka S. Generation of germline-competent induced pluripotent stem cells. Nature. 2007 Jul 19;448(7151):313-7.
4. Maherali N, Sridharan R, Xie W, et al. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. Cell stem cell. 2007 Jun 7;1(1):55-70.
5. Wernig M, Meissner A, Foreman R, et al. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature. 2007 Jul 19;448(7151):318-24.
6. Sugimoto N, Eto K. Generation and manipulation of human iPSC-derived platelets. Cellular and molecular life sciences : CMLS. 2021 Apr;78(7):3385-3401.
7. Iglesias JM, Gumuzio J, Martin AG. Linking Pluripotency Reprogramming and Cancer. Stem cells translational medicine. 2017 Feb;6(2):335-339.
8. Ben-Porath I, Thomson MW, Carey VJ, et al. An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nature genetics. 2008 May;40(5):499-507.
9. Chiou SH, Yu CC, Huang CY, et al. Positive correlations of Oct-4 and Nanog in oral cancer stem-like cells and high-grade oral squamous cell carcinoma. Clinical cancer research : an official journal of the American Association for Cancer Research. 2008 Jul 1;14(13):4085-95.
10. Sperger JM, Chen X, Draper JS, et al. Gene expression patterns in human embryonic stem cells and human pluripotent germ cell tumors. Proceedings of the National Academy of Sciences of the United States of America. 2003 Nov 11;100(23):13350-5.
11. Wong DJ, Liu H, Ridky TW, et al. Module map of stem cell genes guides creation of epithelial cancer stem cells. Cell stem cell. 2008 Apr 10;2(4):333-44.
12. Ghosh Z, Huang M, Hu S, et al. Dissecting the oncogenic and tumorigenic potential of differentiated human induced pluripotent stem cells and human embryonic stem cells. Cancer research. 2011 Jul 15;71(14):5030-9.
13. Roy NS, Cleren C, Singh SK, et al. Functional engraftment of human ES cell-derived dopaminergic neurons enriched by coculture with telomerase-immortalized midbrain astrocytes. Nature medicine. 2006 Nov;12(11):1259-68.
14. Wernig M, Benninger F, Schmandt T, et al. Functional integration of embryonic stem cell-derived neurons in vivo. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2004 Jun 2;24(22):5258-68.
15. Xie X, Cao F, Sheikh AY, et al. Genetic modification of embryonic stem cells with VEGF enhances cell survival and improves cardiac function. Cloning and stem cells. 2007 Winter;9(4):549-63.
16. Sato Y, Bando H, Di Piazza M, et al. Tumorigenicity assessment of cell therapy products: The need for global consensus and points to consider. Cytotherapy. 2019 Nov;21(11):1095-1111.
17. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics. 2010 Apr;11(2):242-53.
18. Agarwal V, Bell GW, Nam JW, et al. Predicting effective microRNA target sites in mammalian mRNAs. eLife. 2015 Aug 12;4.
19. Kawaguchi R, Kiryu H. Parallel computation of genome-scale RNA secondary structure to detect structural constraints on human genome. BMC bioinformatics. 2016 May 6;17(1):203.
20. Medvedeva YA, Lennartsson A, Ehsani R, et al. EpiFactors: a comprehensive database of human epigenetic factors and complexes. Database : the journal of biological databases and curation. 2015;2015:bav067.
21. Dai E, Yu X, Zhang Y, et al. EpimiR: a database of curated mutual regulation between miRNAs and epigenetic modifications. Database : the journal of biological databases and curation. 2014;2014:bau023.

22. Kyriakou G, Melachrinou M. Cancer stem cells, epigenetics, tumor microenvironment and future therapeutics in cutaneous malignant melanoma: a review. Future oncology. 2020 Jul;16(21):1549-1567.
23. Lee JE, Kim MY. Cancer epigenetics: Past, present and future. Seminars in cancer biology. 2022 Aug;83:4-14.
24. Recillas-Targa F. Cancer Epigenetics: An Overview. Archives of medical research. 2022 Dec;53(8):732-740.
25. Ravindran Menon D, Hammerlindl H, Torrano J, et al. Epigenetics and metabolism at the crossroads of stress-induced plasticity, stemness and therapeutic resistance in cancer. Theranostics. 2020;10(14):6261-6277.

CHAPTER 3

CHAPTER 3| Influence of reprogramming methods towards imparting oncogenicity to stem cell derivatives

Abstract:
The presence of oncogenic contamination in induced pluripotent stem cell (iPSC) derivatives across all the three germ layers raise concerns regarding the safety and reliability of these cell populations for therapeutic applications. The generation of iPSCs through various reprogramming methods involving different factors instigates us to check whether this can be one of the potential causes of the remnant oncogenicity within iPSC derivatives. Hereby, in this chapter, we specifically focused on iPSC-derived endothelial cells belonging to the mesodermal layer and analyzed datasets corresponding to iPSC-derived endothelial cells where different reprogramming methods have been used to generate the iPSCs, alongside control and cancerous counterparts. Our analysis unveils the presence of an oncogenic signature within iPSC derived cells irrespective of the germ layers as well as the reprogramming methods used to generate the iPSCs. It will serve as the basic framework for developing a quality control measure to predict the eligibility of these iPSC derivates to be used for regenerative therapy.

Presented this work at Cold Spring Harbor Laboratory Conference Regulatory & Non-Coding RNAs, 2022: "Investigating the non coding RNA mediated oncogenicity in induced pluripotent stem cell derivatives"

3.1. Introduction

It has been observed that induced pluripotent stem cell(iPSC) derived cells possess an oncogenic contamination which limits its usage for regenerative therapy. In our previous chapter (Chapter 2), it has been shown that this oncogenic signature persists within iPSC derivatives corresponding to all the three germ layers viz. ectoderm, endoderm and mesoderm. From here we traced back our journey to look into the generation of the iPSCs to investigate if the clue lies there.

iPSCs are generated from somatic cells that have been reprogrammed using combination of different transcription factors [1]. Reprogramming factors, crucial in this conversion, play essential roles in transforming somatic cells into a pluripotent state akin to embryonic stem cells. Following Shinya Yamanaka's groundbreaking research, retroviral vectors containing transcription factors like Oct4, Sox2, Klf4, and c-Myc (OSKM) have become extensively employed. Later lentiviral-based reprogramming became another efficient and robust method [2]. But scientists have started using different cocktail of transcription factors to generate iPSC as the use of viral vectors has raised concerns due to potential genomic integration and the associated risk of tumorigenicity [3].

To address these concerns, alternative methods have been explored. Episomal vectors, designed to minimize genomic integration, provide a safer approach [4]. These vectors can replicate autonomously in the host cell without integrating into the genome, reducing the risk of unintended genetic changes. Sendai virus-mediated reprogramming is a non-integrating method that employs an RNA virus, specifically the Sendai virus, to deliver reprogramming factors into target cells [5]. This method is advantageous for cellular reprogramming as it ensures transient expression of reprogramming factors, minimizing the risk of genomic alterations and providing a safer alternative for the generation of iPSCs. Another alternative involves mRNA transfection, where the reprogramming factors are delivered as mRNA molecules rather than being integrated into the genome [6]. This approach allows for transient expression of these factors, avoiding long-term genetic alterations and minimizing the risk of tumorigenicity. These alternatives address safety concerns and provide versatility in iPSC generation. In addition to vector-based methods, innovative approaches utilizing small molecules to modulate signalling pathways associated with pluripotency have gained traction [7]. Small molecules offer a more controlled and defined means of reprogramming, allowing for fine-tuning of cellular processes. As the field progresses, ongoing research aims to optimize these methods **[Figure 1]**, considering their respective advantages and challenges. Understanding the intricacies of these reprogramming strategies is crucial for tailoring iPSC production to specific research or clinical needs, ensuring the safe and efficient derivation of pluripotent stem cells for a variety of applications in regenerative medicine and beyond.

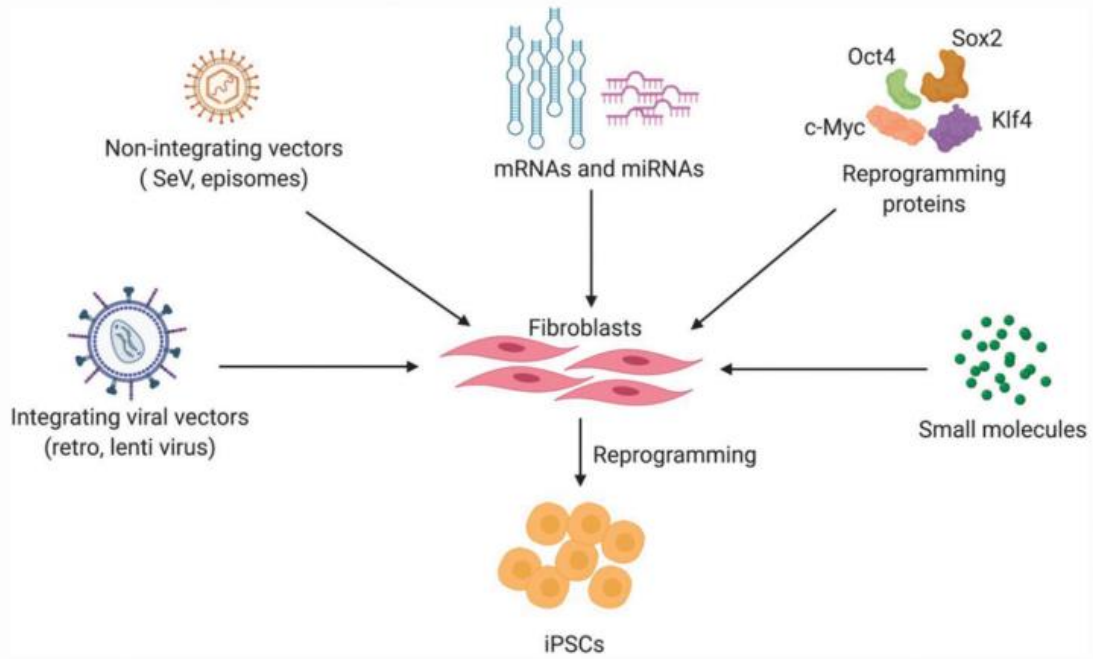


Figure 1: An overview of primary reprogramming techniques employed for the derivation of induced pluripotent stem cells (iPSCs) [8]

In this chapter, we addressed the question as to whether the reprogramming method used for iPSC generation influences the occurrence of oncogenic transformation in the iPSC-derived cells. We have chosen iPSC derived endothelial cells (iPSC_EC), derived from the mesodermal lineage, as our experimental system to carry out this task. The rationale for selecting this cell type is based on its relevance to vascular biology and regenerative medicine along with the availability of data, specifically; data pertaining to various reprogramming methods is accessible only for endothelial cells. By analyzing datasets corresponding to iPSC-derived endothelial cells, where the iPSCs have been generated using different reprogramming methods, alongside human umbilical vein endothelial cells (HUVEC) as control and appropriate cancerous counterparts, we sought to elucidate the relationship between reprogramming methods and oncogenic propensity. Our *in-silico* analysis revealed a persistent oncogenic signature within all these iPSC derived endothelial cells irrespective of the reprogramming method. This observation puts up a pressing need towards developing quality control measures in iPSC-derived cell-based therapies. This work warrants further investigation to unveil the underlying molecular mechanisms behind such remnant oncogenic signature.

As microRNAs (miRNAs) play a crucial role in modulating carcinogenic processes by regulating the expression of various mRNAs, as demonstrated in Chapter 2, we directed our focus toward these regulatory noncoding molecules. We generated small RNA sequencing data in our lab corresponding to the test and the control groups to elucidate the miRNA signatures within them. Our findings were supported by wet lab validations.

In light of these findings, it is imperative to gain a deeper understanding regarding the molecular mechanisms driving oncogenic contamination in iPSC derivatives. Such insights will not only enhance the safety and efficacy of iPSC-based therapies but also pave the way for the development of innovative strategies to mitigate the risk of oncogenic transformation in iPSC-derived cell populations.

3.2. Methods

3.2.1. Different methods of iPSC generation:

The method of iPSC generation is the key event which directs a cell to be reprogrammed. Numerous techniques are available for generating iPSC lines, yet those most appropriate for investigating human diseases and developing therapies should demonstrate sufficient efficiency to produce iPSCs. These methods should also be capable of reprogramming cells derived from both skin fibroblasts and blood. Several reprogramming approaches meet these criteria and can be applied to obtain iPSCs in projects aimed at both fundamental scientific understanding and therapeutic advancements. We have considered the following different reprogramming methods for our analysis based on availability of data (**Table 1**).

Method	Trancription Factors used
Sendai Virus mediated reprogramming	OCT4, SOX2, KLF4 and cMYC
Episomal reprogramming	OCT3/4, shp53, SOX2, KLF4, LMYC, and LIN28
Lentiviral reprogramming	OCT4, SOX2, KLF4 and MYC

Table 1: Different reprogramming methods and their transcription factors

3.2.2. Dataset collection and Experimental Grouping:

We took mRNA expression datasets (sequencing) corresponding to iPSC-derived endothelial cells (iPSC_EC), HUVEC as positive contol and breast cancer and ovarian cancer as cancerous counterpart for all the different methods of iPSC generation from GEO database (**Table 2**). In the analysis, MCF7 and OVCAR3 cell lines had been considered for breast cancer and ovarian cancer, respectively.

	iPSC derivative		Primary Counterpart		Cancerous Counterpart			
Reprogramming method	iPSC-derived EC		HUVEC		Breast Cancer (MCF7)		Ovarian cancer (OVCAR3)	
	Accession No.	No. of samples	Accession No.	No. of samples	Accession No.	No. of samples	Accession No.	No. of samples
Sendai	GSE195559	3	GSE191240	3	GSE215084	3	GSE166767	3
Episomal	GSE214306	3						
Lentiviral	GSE141136	3						

Table 2: RNAseq data corresponding to different reprogramming methods along with Primary and Cancerous Counterpart

Global gene expression profile analysis was done with these samples, where HUVEC served as the control, iPSC_EC served as the test sample 1 and the cancer counterparts as test sample 2 (**Table 3**).

iPSC derivative	Cancerous Counterpart		Primary Counterpart
iPSC-derived Endothelial Cell (iPSC_EC)	Breast Cancer (BC)	Ovarian Cancer (OVC)	Human umbilical vein endothelial cell (HUVEC)

Table 3: Experimental Grouping of the samples

3.2.3. Long RNA sequencing data analysis:

The quality of the raw reads was checked using FastQC [9]. Reads were initially aligned to the reference genome using Hisat2 [10]. Samtools [11] was used to assemble the files which were required for differential gene expression analysis using Cuffdiff [12]. Differentially regulated genes were determined using p-value ≤0.05 and fold-change cut-off ≥2.0.

3.2.4. Pathway analysis:

To conduct functional annotation of the differentially expressed gene set, we utilized QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, http://www.ingenuity.com/) software. This software employs the Ingenuity Pathways Knowledge Base (Ingenuity Systems, Inc., Redwood City, CA) to allocate biological functions to genes. The research direction shifted towards investigating the pathways of the similarly regulated gene expression of the three methods, driven by the goal of identifying crucial pathways regulated regardless of the reprogramming method, with a particular focus on their implications in oncogenicity.

3.2.5. Cell culture and Preparation for sequencing:
We had procured HUVEC from Invitrogen and MCF7 (Breast Cancer cell line) and OVCAR3 (ovarian cancer cell line) from CLS (https://www.clinisciences.com/). iPSC_EC cell line was a generous gift from Dr. Joseph Wu lab, Stanford University. The iPSC_EC cell line was generated via sendai virus mediated reprogramming. HUVEC, MCF7 and OVCAR3 were cultured in Medium 200 (GIBCO), DMEM (GIBCO) and RPMI (GIBCO) respectively, supplemented with 10% fetal bovine serum (FBS) and 0.1% penicillin/streptomycin and at 37°C, in a humidified atmosphere of 5% CO2 **[Figure 2]**. Cells were passaged in a fixed duration of time and cryopreserved in a regular manner.

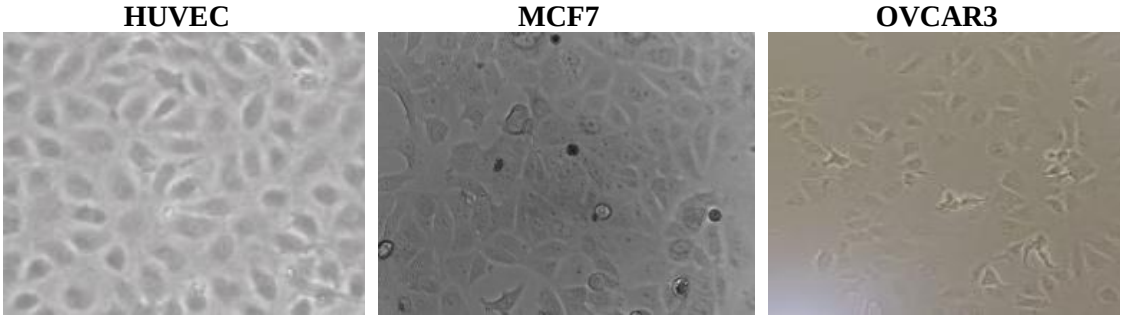


Figure 2: Culture and maintenance of the cell lines

For sequencing, all the samples (cryoprserved vials of HUVEC, OVCAR3 and iPSC_EC) were taken out from liquid nitrogen and pelleted down after thawing. Cell pellets were washed with freshly prepared PBS and labeled properly. There were two replicates for each sample. The vials were then sent to Genotypic, Bangalore for sequencing. NEXTflex™ Small RNA-Seq Kit v3 was used for library preparation. Illumina-compatible sequencing libraries were quantified by Nanodrop and checked in Agilent RNA Bioanalyzer for integrity.

3.2.6. Experimental Grouping and Small RNA sequencing data analysis:
Raw small RNA expression profile data was received in fastq format. HUVEC served as the control, iPSC_EC served as the test sample 1 and OVCAR3 as the cancerous counterpart test sample 2 (as previously mentioned in **Table 3**). For breast cancer, online available data from GEO (https://www.ncbi.nlm.nih.gov/geo/) (GSE171282 – 3replicates) for MCF7 had been utilized.

The quality of the raw reads was checked using FastQC [9]. BLAST [13] was used to align the reads with known miRNAs obtained from miRBase [14]. Differentially regulated miRNAs were determined using Deseq2 [15], with fold-change cut-off ≥2.0. The workflow for small RNA analysis is shown in **Figure 3**.

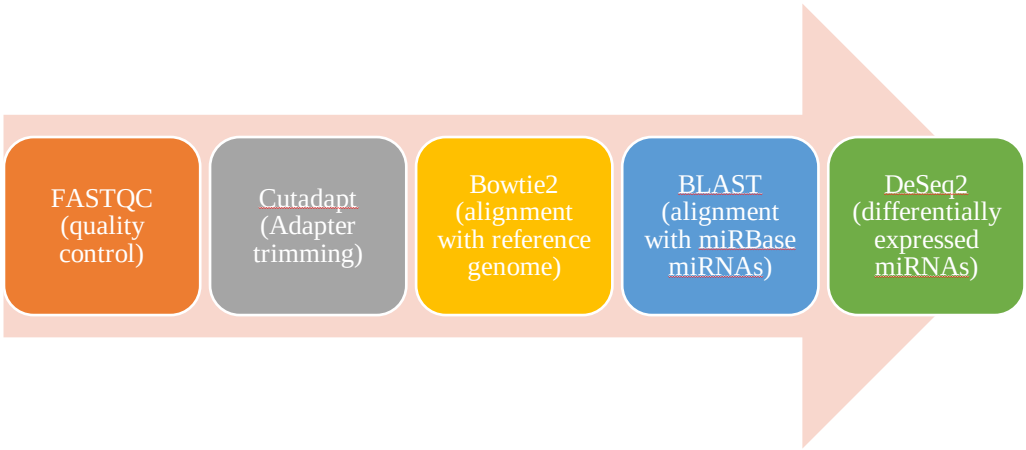


Figure 3: Workflow for small RNA sequencing

3.2.7. *miRNA Target gene prediction:*

Differentially expressed miRNAs were subsequently selected to forecast their target genes from the pool of commonly regulated gene set obtained from pathways analysis. TargetScan [16] was employed for predicting gene targets, followed by filtration using Parasor [17], which refines the targeted genes based on accessible regions.

3.2.8. Total RNA and small RNA isolation followed by cDNA synthesis:

Long and small RNA were separately extracted from the four cell lines utilizing the miRNeasy Mini Kit (Qiagen, Cat. 217004). In brief, cells at approximately 80% to 90% confluence in T-25 flasks were lysed using Qiazol. RNA extraction was carried out from the lysate following the manufacturer's protocol. The quantification of RNA was performed using a NanoDrop-LITE spectrophotometer. For cDNA synthesis, 1µg of total RNA was utilized in reverse transcription PCR using the Verso cDNA synthesis Kit (Thermo Scientific), following the manufacturer's instructions.

3.2.9. Quantitative real time PCR:

The amplification of target genes was carried out using SYBR® Select Master Mix (Applied Biosystems™) on a 7500 Fast Real-Time PCR System (Applied Biosystems). All mRNA quantification data were normalized to GAPDH (an endogenous control). The fold differences in target gene expression were calculated relative to HUVEC. For miRNA quantification, qRT-PCR was conducted using the miRCURY LNA miRNA PCR Starter Kit (Qiagen #339320) following the manufacturer's protocol, with mir-103a-3p serving as the endogenous control for miRNAs. Statistical analysis was performed utilizing a paired t-test, with a p-value of ≤ 0.05 considered statistically significant.

3.3. Results and Discussion

3.3.1. Oncogenic signature within the iPSC_EC generated via different reprogramming method:

The differentially expressed genes (DEGs) for different experimental groups have been put up in Table 4.

Control	Test samples	Differentially expressed mRNAs	
		Upregulated	Downregulated
HUVEC	iPSC_EC (Sendai Virus)	2467	2454
	iPSC_EC (Episomal Vector)	2913	2642
	iPSC_EC (Lentiviral)	2145	2970
	BC	2800	3325
	OVC	2398	3486

Table 4: DEGs across the different samples

Thereafter, we screened the common set of mRNAs [**Figure 4A, B, C**] across all the 3 sets for all types of reprogramming methods.

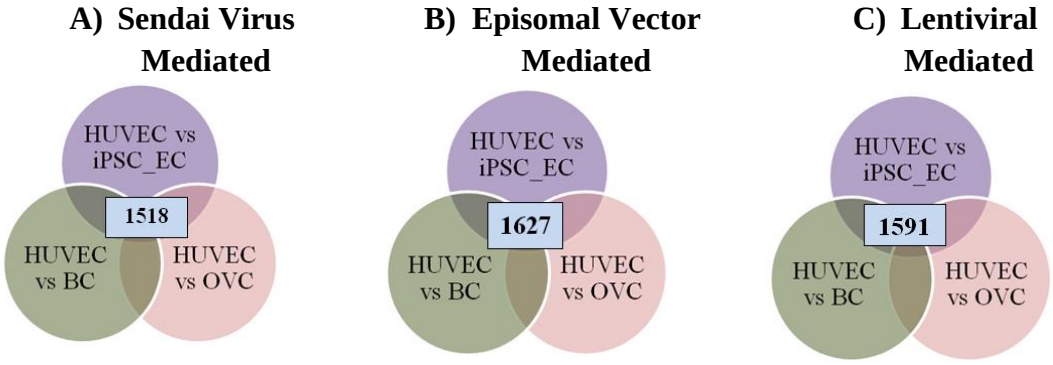


Figure 4: Venn diagram showing the overlap of the differentially expressed mRNAs among HUVEC versus iPSC_EC, HUVEC vs BC & HUVEC versus OVC of different methods

Hierarchical clustering of iPSC_EC, HUVEC and cancerous counterparts had been done using the expression profile of these common set of mRNAs (**Figure 5A, B, C**).

Reprogramming Methods **Heatmap**

A) Sendai Virus Mediated

B) Episomal Vector Mediated

C) Lentiviral Mediated

Figure 5: Heatmap showing the expression profile for the commonly expressed mRNAs in the samples. Clustering shows iPSC_EC to be closest to the cancerous counterparts in all the three methods

The heat map in Figure 5 reveals that that iPSC_EC (for which iPSC is generated via different reprogramming methods) cluster with the cancerous counterparts rather than with the primary counterpart HUVEC. This indicates the presence of oncogenic contamination within iPSC-derivative irrespective of the reprogramming method used for generating the iPSCs.

3.3.2. Common gene signature associated Pathway Analysis:

We observed that 277 genes are upregulated and 502 genes (779 in total) are downregulated in all the iPSC_ECs corresponding to different reprogramming methods along with the cancerous systems **[Figure 6]**. The pathway analysis of these commonly regulated genes was conducted to assess their functional implications using IPA. Given the same cell type, though generated from different reprogramming methods, it can be inferred that iPSC_EC should maintain similar modifications to exhibit oncogenic traits. By elucidating these consistent regulated pathways, the study aimed to uncover fundamental mechanisms underlying oncogenicity that transcend specific reprogramming techniques, enhancing our understanding as to what are the key factors behind such oncogenicity.

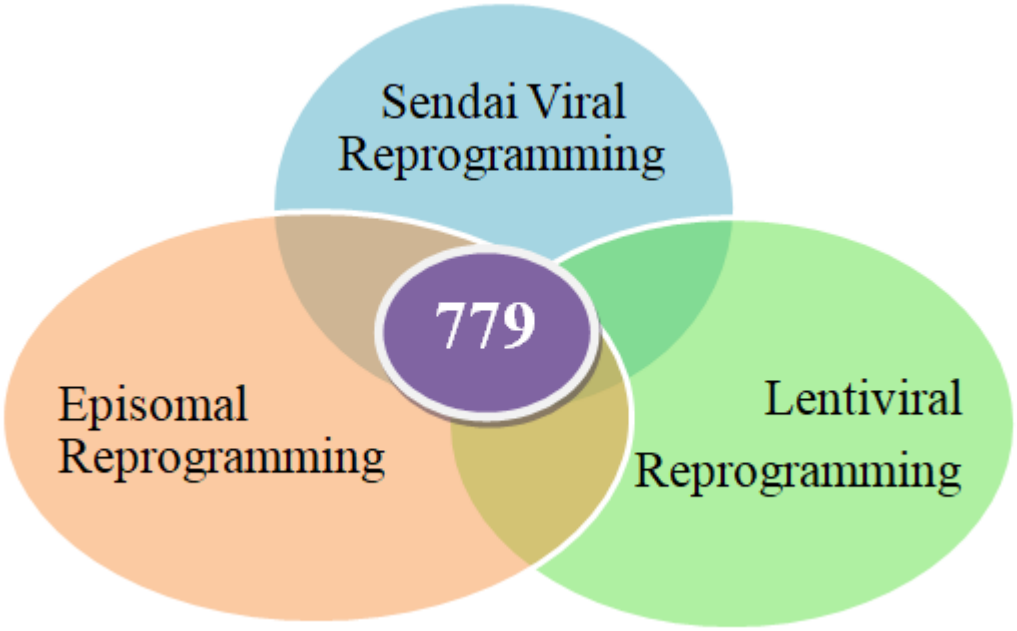


Figure 6: Commonly expressed genes with same regulation pattern among three reprogramming methods

178 significant pathways were identified from the pathway analysis. These pathways were meticulously examined to determine their association with oncogenicity. We focused on pathways directly linked to general carcinogenesis, excluding those associated with tissue-specific cancers, resulting in a total of 122 pathways **[Figure 7]** for further analysis.

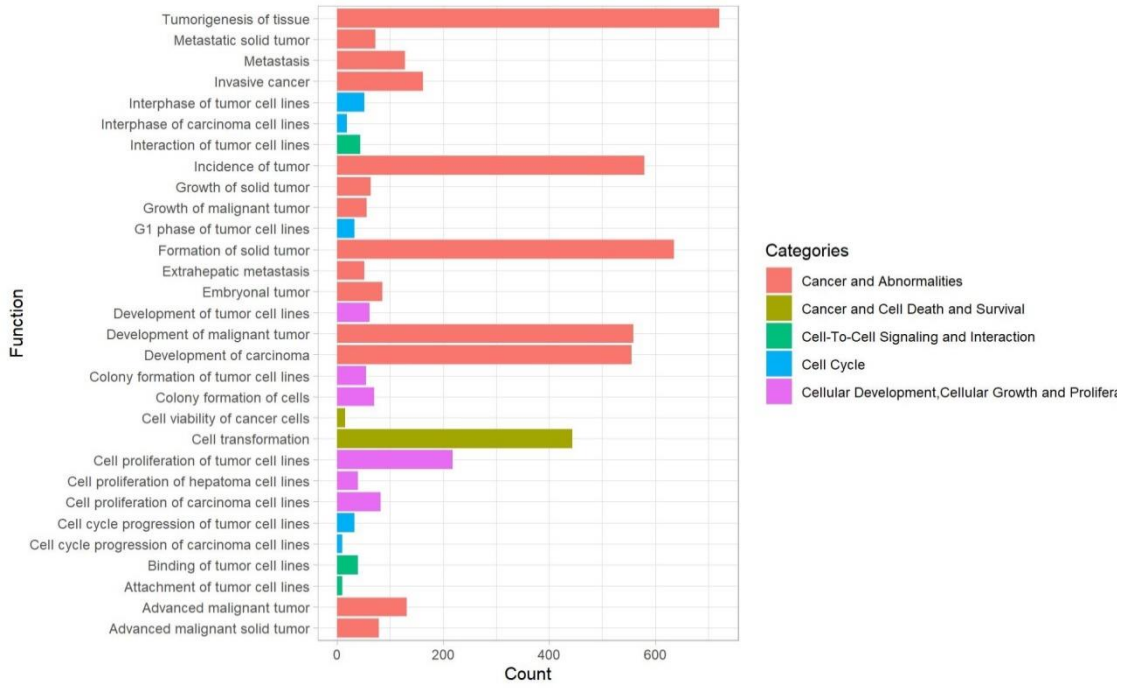


Figure 7: Barplot showing the functional distribution of the genes

3.3.3. Small RNA profile reveals oncogenic properties:

After analyzing the raw data, we specifically chose reads that precisely aligned with the complete sequences of mature miRNAs. The data statistics for each dataset, along with the counts of unique miRNAs, is provided in **Table 5**.

Sample Name	Total no of Reads	Total reads mapping to annotated miRNAs	No. of Unique annotated miRNAs
iPSC_EC_rep1	5668378	645594	180
iPSC_EC_rep2	12830672	726597	176
HUVEC_rep1	16740951	13506136	784
HUVEC_rep2	10123243	5871736	389
OVCAR3_rep1	15554033	5678382	394
OVCAR3_rep2	5411365	1412327	252
MCF7_rep1	7173617	6734868	990
MCF7_rep2	8431928	8489580	1076
MCF7_rep3	7581453	8261495	1007

Table 5: Expressed miRNA counts across the different samples

We observed that 128 miRNAs were consistently expressed in both replicates of iPSC_EC. Additionally, 346, 214, and 783 miRNAs exhibited consistent expression across replicates of HUVEC, OVCAR3, and MCF7, respectively. For the subsequent differential analysis, we identified 124 miRNAs commonly expressed in both HUVEC and iPSC_EC, 185 miRNAs in HUVEC and OVCAR3, and 285 miRNAs in HUVEC and MCF7. Following the differential analysis using Deseq2 (P-value <0.05 and fold-change cut off of ≥ 2.0), we focused on the common set of miRNAs among the three differential panels **[Figure 8]**.

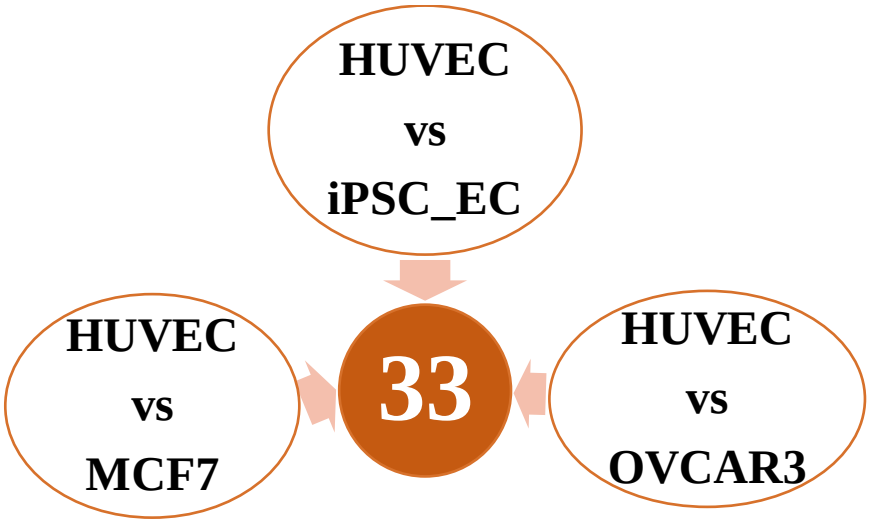


Figure 8: The overlap of the expressed miRNAs among the differential miRNA set

Further hierarchical clustering of iPSC_EC, HUVEC, MCF7 and OVCAR3 had been done using the expression profile of these common set of 33 miRNAs [**Figure 9**].

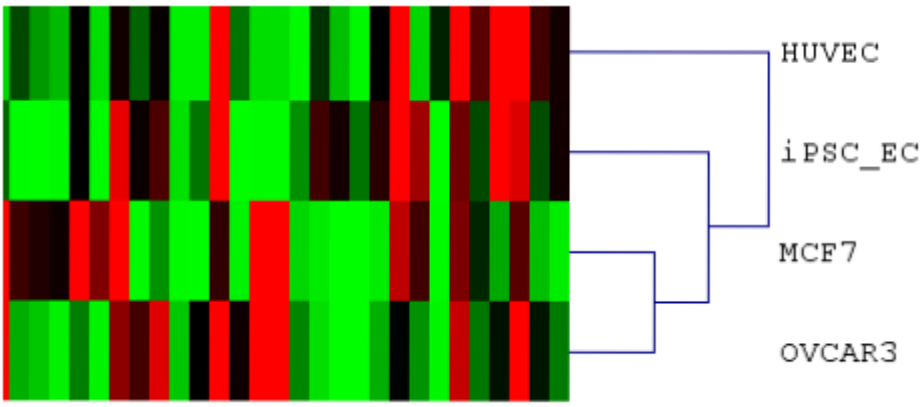


Figure 9: Heatmap showing the expression profile for the commonly expressed 33 miRNAs in the samples. Clustering shows iPSC_EC to be closest to the cancerous counterpart

The findings from Figure 11 provide clear evidence that iPSC_EC (generated via Sendai virus reprogramming) clusters more closely with the cancerous counterpart rather than HUVEC. This observation strongly suggests the presence of oncogenic contamination within iPSC derivatives, with respect to miRNA expression profiles too.

3.3.4. miRNA-mRNA target association toward oncogenicity:

To elucidate the regulatory dynamics between miRNA and mRNA governing such oncogenic contamination, a comprehensive approach was undertaken involving target prediction analysis between these molecular entities. Among the 33 DE miRNAs, it has been found that 7 miRNAs are upregulated and 6 miRNAs are downregulated in

iPSC_EC, MCF7 and OVCAR3. We predicted the negatively correlated targets of these miRNAs within the previously obtained common 122 pathway associated genes (Table 6) using TargetScan followed by a filtration using ParasoR for accessible region.

Target Prediction Steps	Up miRNAs - Down Genes		Down miRNAs - Up Genes	
	No. of miRNAs	No. of Genes	No. of miRNAs	No. of Genes
Input Set	7	202	6	85
Target Prediction	6	161	4	69
ParasoR	6	79	4	29

Table 6: No. of miRNAs and mRNAs for target prediction

Following the determination of target pairs, meticulous examination of the miRNAs' mode of action ensued, delineating their specific regulatory pathways and interactions. Subsequently, emphasis was placed on elucidating the functional significance of the targeted genes, delving into the intricacies of their regulatory responses based on the directional modulation induced by the corresponding miRNAs. Overall, this rigorous analytical approach illuminated the intricate interplay between miRNAs and their target mRNAs **(Table 7)**, unveiling the multifaceted regulatory landscape underlying cellular processes and molecular pathways.

miRNA	Mode of Regulation	miRNA Function	Targeted Gene	Mode of Regulation	Gene Association in Cancer
hsa-miR-182-5p	Up regulated	Promotes cell proliferation and suppresses cell apoptosis	RHOQ	Down Regulated	Low RhoQ levels were associated with poor overall survival with adenocarcinoma
			NFIC		A potential tumour suppressor through the epigenetic modification
hsa-miR-148a-3p		Promotes tumor angiogenesis by activating the EGFR/MAPK signalling pathway	PPP1R12B		Inhibits tumor growth and metastasisthrough regulating Grb2/PI3K/Akt signalling
			RHOQ		Mentioned earlier
hsa-miR-411-5p	Down regulated	Reported as a tumor suppressor in several cancers, regulates proliferation, migration and invasion	CASP6	Up regulated	Verified as an oncogene
hsa-miR-6869-5p		Prevented the proliferation and invasion of cancer cells	NMNAT1		Considered a poor prognostic marker

Table 7: Screened set of miRNA-mRNA target pairs

Subsequently, we pursued experimental validation through quantitative real-time polymerase chain reaction (qRT-PCR) to assess the expression levels of the identified miRNA-target gene pairs. Our experimental results showed a concordance between in silico analyses and real-time observations **[Figure 10].**

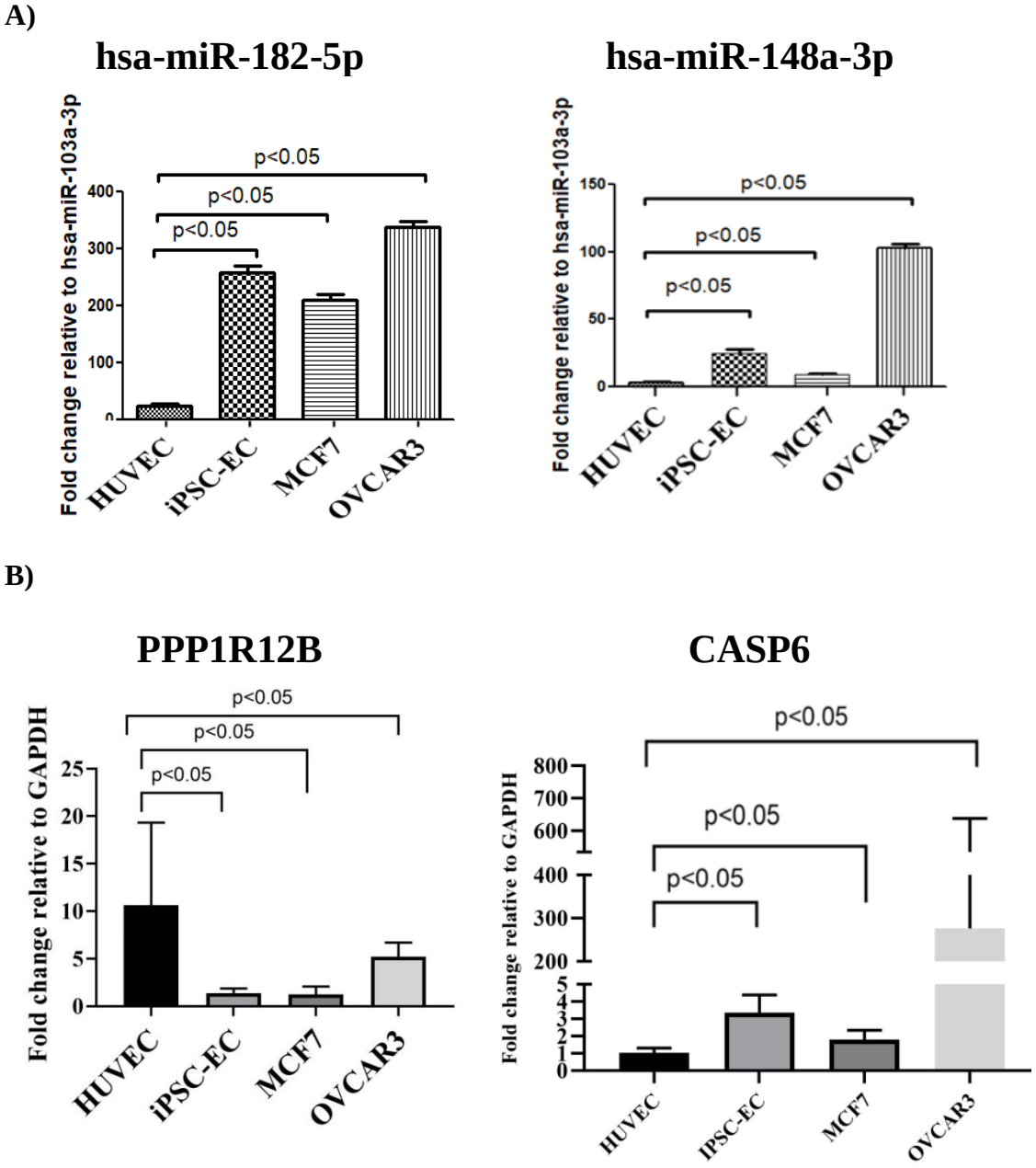


Figure 10: qRT-PCR validation for the A) miRNAs - hsa-miR-182-5p and hsa-miR-148a-3p and B) the target genes - PPP1R12B and CASP6

Specifically, the expression profiles of two miRNAs, namely hsa-miR-182-5p and hsa-miR-148a-3p **[Figure 10A]**, exhibited consistent significant upregulation in iPSC_EC, MCF7 and OVCAR3 with respect to HUVEC in both in silico predictions and qRT-PCR analyses. In addition to this, our real-time PCR expression analysis revealed a concurrent

downregulation of the target gene PPP1R12B and the upregulation of the target gene CASP6 **[Figure 10B]**.

On the contrary, qRT-PCR analyses for the downregulated miRNAs hsa-miR-411-5p and hsa-miR-6869-5p exhibited low expression levels in iPSC_EC and cancer systems compared to HUVEC **[Figure 11]** as is expected from our *in silico* analysis. But these observed downregulations were not statistically significant. Similarly, the expression patterns of the target genes NFIC, RHOQ, and NMNAT1 (data not shown) were consistent with the trends observed in the *in silico* analysis. Nevertheless, fold change calculation did not yield significant regulatory alterations in these target genes too with respect to the control. Hence we are in the process of repeating these experiments.

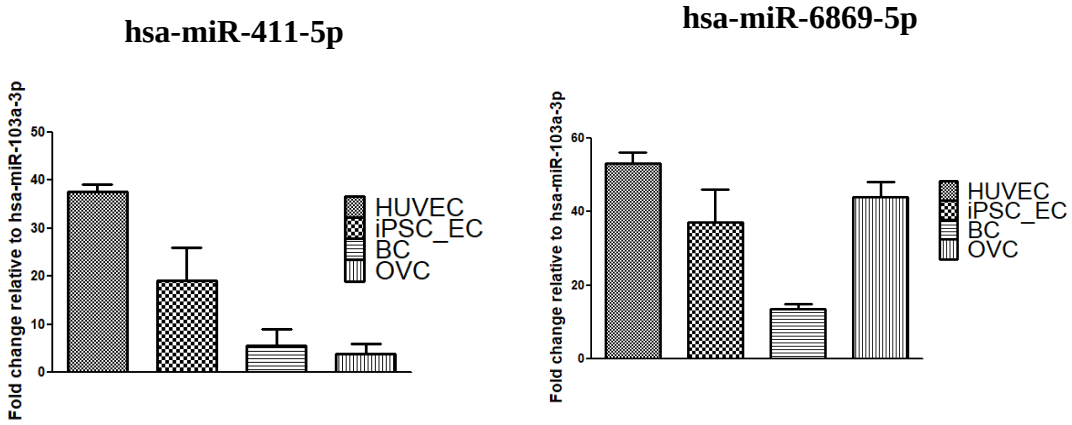


Figure 11: qRT-PCR validation for the miRNAs - hsa-miR-411-5p and hsa-miR-6869-5p

It is pertinent to highlight the oncogenic nature of miR-182-5p (miR-182), a microRNA found to be highly expressed in various tumors, including colorectal cancer (CRC) [18]. Despite its expression in the early stages of tumorigenesis, its precise role in driving cancer development remains elusive. Additionally, its upregulation in non-small cell lung cancer (NSCLC) tissues has been associated with tumor recurrence [19], where it serves as an independent prognostic factor. Functionally, miR-182-5p overexpression has been shown to promote cancer cell migration and invasion in lung cancer [19]. Another miRNA, miR-148a-3p, found in glioma and osteosarcoma, has been implicated in tumor angiogenesis through activation of the EGFR/MAPK signaling pathway [20].

The PPP1R12B axis, known to inhibit tumor growth and metastasis, exerts its regulatory influence by modulating the Grb2/PI3K/Akt signaling pathway in colorectal cancer [21]. Additionally, the caspase-6-mediated cleavage of RIPK1 [22] is significant in regulating oncogenicity. RIPK1 downregulation has been observed in iPSC derivatives and cancer systems based on our in-silico analysis. RIPK1 inactivation via caspase-mediated cleavage is crucial for development and may have implications in tumorigenesis.

These highlight the intricate molecular interplay driving cancer pathogenesis, wherein miRNAs, target genes, and signaling pathways emerge as key players governing tumor initiation, progression, and metastasis. Recognizing the regulatory networks orchestrated

by these molecular entities holds significant promise for developing strategies aimed at generating iPSC derivatives devoid of oncogenic contamination.

These findings suggest that while there may be trends suggestive of regulatory modulation, the observed changes in expression levels of miRNAs and their target genes may not reach statistical significance within the experimental context examined. This highlights the complexity of miRNA-mediated regulatory networks and underscores the importance of considering various factors that may influence gene expression dynamics in different biological systems. Further investigations employing larger sample sizes or additional experimental conditions may provide greater statistical power to discern subtle regulatory effects and shed more light on the functional significance of these miRNA-target interactions in specific physiological or pathological contexts.

By elucidating the roles of specific miRNAs, target genes, and signaling cascades implicated in cancer development, researchers gain crucial insights into the underlying mechanisms driving oncogenic transformation in iPSC derivatives. Strategies may involve fine-tuning culture conditions, optimizing differentiation protocols including duration of differentiation, and implementing stringent quality control measures to ensure the safety and integrity of iPSC-derived cell populations.

3.4.Conclusion

Understanding the nuances of these reprogramming strategies is crucial for tailoring iPSC production to specific research or clinical needs. Each method comes with its unique set of advantages and challenges, contributing to an ongoing dialogue on optimizing iPSC generation for therapeutic applications. As we delve deeper into these techniques, we not only enhance our understanding of stem cell biology but also move closer to harnessing the full potential of iPSCs for diverse and personalized applications in regenerative medicine.

Overall, our findings suggest that iPSC derived endothelial cells bear oncogenic signature irrespective of reprogramming methods which reveals that reprogramming factors are not the sole reason for oncogenic contamination in iPSC derivatives. Functional analysis also helped us to get a clear picture of the dominant regulatory hubs which gets distorted and eventually contributes to the aberrations within these cells. Ultimately, advancing our knowledge of these regulatory networks not only enhances our understanding of cancer biology but also paves the way for the development of safer and more efficacious iPSC-based therapies and regenerative medicine approaches. By mitigating the risk of oncogenic contamination, we can harness the full potential of iPSC technology for regenerative therapy.

References

1. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006 Aug 25;126(4):663-76.
2. Matrai J, Chuah MK, VandenDriessche T. Recent advances in lentiviral vector development and applications. Molecular therapy : the journal of the American Society of Gene Therapy. 2010 Mar;18(3):477-90.
3. Ebben JD, Zorniak M, Clark PA, et al. Introduction to induced pluripotent stem cells: advancing the potential for personalized medicine. World neurosurgery. 2011 Sep-Oct;76(3-4):270-5.
4. Wang AYL, Loh CYY. Episomal Induced Pluripotent Stem Cells: Functional and Potential Therapeutic Applications. Cell transplantation. 2019 Dec;28(1_suppl):112S-131S.
5. Schlaeger TM, Daheron L, Brickler TR, et al. A comparison of non-integrating reprogramming methods. Nature biotechnology. 2015 Jan;33(1):58-63.
6. Warren L, Lin C. mRNA-Based Genetic Reprogramming. Molecular therapy : the journal of the American Society of Gene Therapy. 2019 Apr 10;27(4):729-734.
7. Desponts C, Ding S. Using small molecules to improve generation of induced pluripotent stem cells from somatic cells. Methods in molecular biology. 2010;636:207-18.
8. Soman SS, Vijayavenkataraman S. Applications of 3D Bioprinted-Induced Pluripotent Stem Cells in Healthcare. International journal of bioprinting. 2020;6(4):280.
9. S. A. FastQC: a quality control tool for high throughput sequence data. http://wwwbioinformaticsbabrahamacuk/projects/fastqc. 2010.
10. Pertea M, Kim D, Pertea GM, et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature protocols. 2016 Sep;11(9):1650-67.
11. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021 Feb 16;10(2).
12. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology. 2010 May;28(5):511-5.
13. Chen Y, Ye W, Zhang Y, et al. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic acids research. 2015 Sep 18;43(16):7762-8.
14. Griffiths-Jones S. miRBase: microRNA sequences and annotation. Current protocols in bioinformatics. 2010 Mar;Chapter 12:12 9 1-12 9 10.
15. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550.
16. Agarwal V, Bell GW, Nam JW, et al. Predicting effective microRNA target sites in mammalian mRNAs. eLife. 2015 Aug 12;4.
17. Kawaguchi R, Kiryu H. Parallel computation of genome-scale RNA secondary structure to detect structural constraints on human genome. BMC bioinformatics. 2016 May 6;17(1):203.
18. Sameti P, Tohidast M, Amini M, et al. The emerging role of MicroRNA-182 in tumorigenesis; a promising therapeutic target. Cancer cell international. 2023 Jul 12;23(1):134.
19. Yang W, Yin Y, Bi L, et al. MiR-182-5p promotes the Metastasis and Epithelial-mesenchymal Transition in Non-small Cell Lung Cancer by Targeting EPAS1. Journal of Cancer. 2021;12(23):7120-7129.
20. Wang M, Zhao Y, Yu ZY, et al. Glioma exosomal microRNA-148a-3p promotes tumor angiogenesis through activating the EGFR/MAPK signaling pathway via inhibiting ERRFI1. Cancer cell international. 2020;20:518.

21. Ding C, Tang W, Wu H, et al. The PEAK1-PPP1R12B axis inhibits tumor growth and metastasis by regulating Grb2/PI3K/Akt signalling in colorectal cancer. Cancer letters. 2019 Feb 1;442:383-395.
22. van Raam BJ, Ehrnhoefer DE, Hayden MR, et al. Intrinsic cleavage of receptor-interacting protein kinase-1 by caspase-6. Cell death and differentiation. 2013 Jan;20(1):86-96.

CHAPTER 3

CHAPTER 4

CHAPTER 4| Development of a prediction model to predict the oncogenic status of an iPSC derived cell

Abstract: Assessing the safety of induced pluripotent stem cell (iPSC)-derived cells post-differentiation is paramount for their effective use in therapeutic applications. In our previous work, it has been shown that the transcriptome and the microRNome of the stem cell derivatives corresponding to all the three germ layers bear a remnant oncogenic signature irrespective of the various reprogramming methods used for generating its parent iPSCs. We were able to decipher the essential miRNA-target gene induced pathways which play a dominant role towards inducing such oncogeninc contamination. Current strategies to detect the oncogenic contamination within these differentiated counterparts of the iPSCs typically require invasive and ethically challenging animal transplantation studies. To address these limitations, we have developed a novel machine learning based prediction model based on the interesting clues revealed from the miRNA-target gene induced pathways responsible for such oncogenicity, as obtained from our previous work, to assess the oncogenicity within iPSC-derived cells using transcriptome data. Subsequently, it also provides the list of genes which act as a dominating contributory factor behind such contamination.

4.1. Introduction

In recent years, induced pluripotent stem cells (iPSCs) and their derivatives have emerged as invaluable tools in regenerative medicine, disease modeling, and drug discovery. However, one significant challenge that limits the usage of iPSCs and its derivatives is the potential risk of oncogenic contamination, which can compromise the safety and efficacy of its downstream applications. From the previous work as detailed in chapter 2 and 3, it has been clearly seen that the transcriptome and the microRNome of the stem cell derivatives corresponding to all the 3 germ layers bear a remenant oncogenic signature irrespective of the various reprogramming methods used for generating its parent iPSCs. In such a venture, we were able to decipher the essential miRNA-target gene induced pathways which play a dominant role towards inducing such oncogeninc contamination.

Hence our work reiterates the importance of adopting rigorous safety assessments to mitigate potential risk of developing cancer while using iPSC derivatives for regenerative therapy. While conventional methods like marker assays and pluripotent gene expression [1] analyses provide valuable insights into the differentiation status of iPSC-derived cells, detecting oncogenic transformation within them often requires labor-intensive and ethically challenging animal transplantation studies.

There have been previous reports regarding prediction tools like 'PluriTest' [2] which was developed to assess pluripotency in order to minimize animal sacrifice, but there exists no such prediction models to detect the presence of oncogenic contamination in iPSC-derived cell populations.

This motivated us to move a step ahead to develop a sort of robust quality control measure for checking the purity of newly generated iPSC derivatives. With such motivation along with the interesting clues obtained from our studies reported in chapter 2 and 3, we developed a prediction model to predict the oncogenic status within an iPSC derivative based on its transcriptome data. It is a machine learning (ML) model which uses features that have been obtained from the transcriptome and microRNome analysis of the iPSC derived cells previously in chapters 2 and 3. Our model offers a cost-effective approach to ensure the safety and reliability of iPSC-derived cell populations for therapeutic use. Further, it also provides the list of genes which act as a dominating contributory factor behind such contamination.

4.2. Methods

4.2.1. Dataset collection, Data Preprocessing and Feature Selection:

On analyzing the transcriptome and microRNome corresponding to the iPSC derivatives from the 3 germ layers and with a detailed knowledge regarding the miRNA targeted enriched pathways responsible for the remnant oncogenic contamination within them, the training dataset to develop a robust prediction model for predicting the quality of the iPSC derivatives have been constructed. The detail of the training dataset is provided in **Table 1**. In brief, we utilized a training dataset comprising primary cells corresponding to the three germ layers as positive training datasets and cancerous samples as negative training datasets. Positive training data include primary cells representative of the ectodermal, mesodermal, and endodermal lineages, ensuring comprehensive coverage of cell types derived from each germ layer. These primary cells served as reliable benchmarks for assessing the normal phenotype of iPSC-derived cell populations post-differentiation. Conversely, negative controls consisted of cancerous samples, chosen to represent instances of potential oncogenic contamination within iPSC-derived cell populations. By juxtaposing positive and negative data, we aimed to train the prediction model to accurately discriminate between normal and aberrant cellular phenotypes, thereby enhancing its efficacy in detecting oncogenic signatures within iPSC derivatives. All the datasets correspond to [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array.

Germ Layer	Positive Training Set	Positive Training Sample IDs - #of samples	Negative Training Set	Negative Training Sample IDs - #of samples
Ectoderm	Neuron	GSE12679 - 6 GSE19332 - 7 GSE40438 - 8	Neuroblastoma	GSE51978 - 3 GSE13273 - 2 GSE4600 - 3
Endoderm	Hepatocyte	GSE62962 - 2 GSE18269 - 6	Hepatocellular Carcionoma	GSE18269 - 3 GSE29084 - 2 GSE23031 - 3
Mesoderm	Cardiomyocyte	GSE59704 - 2 GSE32911 - 1	Sarcoma	GSE17800 - 8 GSE14975 - 5

Table 1: Training dataset corresponding to the primary cells belonging to different germ layers, along with their cancerous counterpart

Microarray data analysis had been done using affy (https://www.bioconductor.org/packages/release/bioc/html/affy.html) package. In order to remove technical variations across all samples, the datasets had been normalized using scikit-learn (https://scikit-learn.org/) [3]. Finally, data has been batch corrected using sva (https://www.bioconductor.org/packages/release/bioc/html/sva.html) package, such that model training can be employed in seamless manner. The principle feature selection criteria is based on the results obtained from our previous studies (detailed in chapter 2 and 3). Our first work resulted in 116 common pathways across ectodermal, mesodermal, and endodermal lineages which were responsible for oncogenic contamination within the iPSC derivatives. The 2nd work revealed 178 commonly enriched pathways which played a key role for oncogenic contamination within the iPSC derived cells irrespective of the different reprogramming methods considered in our analysis. We finally obtained the common set of pathways from these 2 sets of analysis in order to get the set of pathways that contributes to such contamination irrespective of the germ layers as well as reprogramming methods used for generating the parent iPSCs. We employed extensive literature mining at this stage to screen out the most important set of pathways among these that are involved highly with oncogenicity. The expression profiles of the genes involved in these screened pathways served as training features for segregating normal samples from the cancer ones.

4.2.2. Building of Prediction Model:

We have deployed both supervised classification model and dimensionality reduction based visualization technique in terms of Logistic Regression and Uniform Manifold Approximation and Projection (UMAP) models respectively for predicting the purity of an iPSC derivative. Both of these models have been trained with normal and cancerous samples. Our guiding principle behind selecting such a simplest model is to give more importance on feature set responsible for regulating the oncogenic property.

Logistic Regression: We have incorporated the simplest classification technique in forms of logistic regression [4] in segregating two classes. It is the robustness of selected features which achieve higher accuracy in separating the normal samples from the

cancerous ones linearly. Furthermore while deploying such a classification model it also determines the overall feature importance regulating the oncogenic property in iPSC derived cell.

UMAP based Visualization Technique: Furthermore, we have incorporated the dimensionality reduction based visualization technique in forms of UMAP [5] which allows the user to visualize the closeness of the samples corresponding to the normal or cancerous counterpart from training dataset. The main advantages in utilizing the UMAP plot against the Principal Component Analysis is that it is operated in non-linear mode, further we have implicated supervised UMAP procedure, which would allow user to compare the iPSC derived sample against the training. UMAP is far more advanced comparing to the other dimensionality reduction algorithms as it preserved the local structure pretty well allowing user to visualize their sample across different clusters [6].

Both R and Python programming languages have been incorporated in order to create the in-silico analysis based prediction tool. R based libraries are mainly downloaded from Bioconductor (https://www.bioconductor.org/) software packages, whereas Anaconda (https://www.anaconda.com/products/distribution) served for Python based prediction tool. Python has been utilized to create prediction tool where data has been preprocessed using Pandas (https://pandas.pydata.org/) and Numpy (https://numpy.org/) libraries. Finally we have utilized both Matplotlib (https://matplotlib.org/) from python and ggplot2 (https://ggplot2.tidyverse.org/) from R to create various plotting.

Deviation_Score, determining the set of dominant contributory factors behind such oncogenicity: One of the most significant goals associated with our analysis is determination of the most important set of genes that significantly contributes towards dictating the status of a newly generated iPSC derived cell, provided it turns out to be oncogenic. This will be provided using the 'Deviation_Score' associated with those genes. We can extract the Deviation_Score associated with those genes using the following formula.

$$\text{Deviation\_Score} = \frac{(\mu_{Normal}^{Expression} - Expression_{Input})^2 + \text{Stability Factor}}{(\mu_{Cancer}^{Expression} - Expression_{Input})^2 + \text{Stability Factor}} \qquad (1)$$

Where,

$\boldsymbol{\mu_{Normal}^{Expression}}$ is Expression value of Normal training Sample of given Gene

$\boldsymbol{\mu_{Cancer}^{Expression}}$ is Expression value of Cancerous training Sample of given Gene

$Expression_{Input}$ is the Expression value of Input sample of given Gene

Stability Factor is a small number to remove numerical instability such as division by zero. Here it is considered as 0.001

Main hypothesis of equation (1) is the determination of those genes which are maximally perturbed in terms of Deviation_Score.

For an input sample which turns out to be oncogenic, genes constituting the feature set are sorted based on the Deviation_score in descending manner so that the experimentalist can get an idea regarding the major dominating factors that contribute towards such oncogenicity.

The overall workflow depicting how the prediction model will work is presented in **Figure 1**.

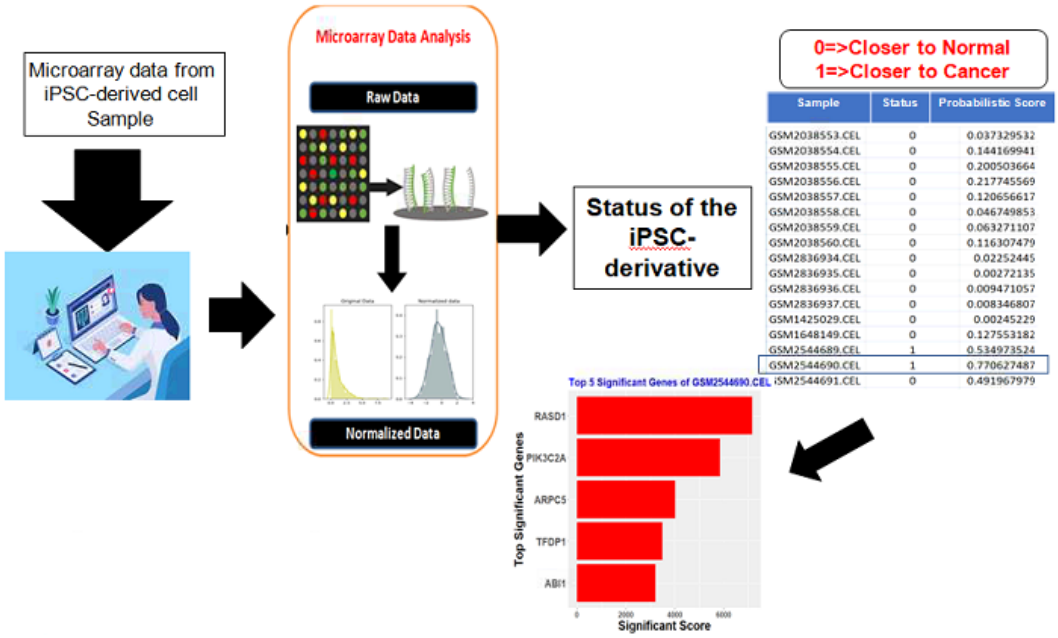


Figure 1: Working principle of the Prediction Model

The entire code (for endodermal lineage as example) for the model is provided below:

Feature_Selection_Training

Following code depicts the python based feature selection scheme mapping across various gene and probesets elucidated from manually selected pathways.

```
import pandas as pd
import numpy as np

df=pd.read_csv('train_data/hepato_norm_frma.txt',sep='\t')
df.rename({'Unnamed: 0':'PROBEID'},axis=1,inplace=True)
print(df.shape)
df.head(1)

df1=pd.read_csv('train_data/hepato pathway no regulation for
umap.txt')
print(df1.shape)
df1.head(1)

df2=pd.read_excel('train_data/significant_hepato.xlsx',sheet_n
ame='Sheet5')
print(df2.shape)
```

```
df2.head(1)

df2=df2[df2.GENE_SYMBOL.isin(df1.GENE_SYMBOL.values)]
print(df2.shape)
df2=df2[['PROBEID','GENE_SYMBOL']]
df2.head(1)

df=pd.merge(df,df2,on=['PROBEID'])
print(df.shape,df.columns)
df.head(1)

df[['PROBEID','GENE_SYMBOL']].to_csv('train_data/Hepato_Select
ed_Probes.csv',sep='\t',index=False)

df.to_csv('train_data/Hepato_Training_Expression.csv',sep='\t'
,index=False)
```

Expression_Generation_From_Unknown_Samples

We have incorporated R based packages to extract normalized intensity value using fRMA package from raw microarray samples.

```
print('Load Libraries ...')

library(affy)
library(hgu133plus2.db)
library(frma)

pd=read.AnnotatedDataFrame("test_data/Hepato_Test_Pheno.txt",h
eader=TRUE,row.names=1)

print('Read Affymetrix Array Files ...')
raw_data = ReadAffy(filenames=rownames(pData(pd)))

print('Normalize Data using fRMA ...')
eset_data=frma(raw_data)
eset_data

write.exprs(eset_data,file='test_data/Hepato_Test_Intensity.tx
t',sep='\t')
```

Batch_Correction

As training samples belong to different GSE, we have executed batch correction to remove experiment specific technical variation using R based **Combat** tool.

```
library(sva)
library(dplyr)
library(tidyverse)
library(affy)

df=read.csv('train_data/hepato_norm_frma.txt',sep='\t',row.nam
es = 1, check.names=FALSE,stringsAsFactors=FALSE)
df=df[,6:21]
```

```
head(df)

df1=read.csv('Test_Data/Hepato_Test_Intensity.txt',sep='\t',ro
w.names = 1, check.names=FALSE,stringsAsFactors=FALSE)
head(df1)

final=merge(df,df1,by = 'row.names', all = TRUE)%>%
  column_to_rownames(var = 'Row.names')

final=as.matrix(final)
dim(final)
head(final)

pheno=read.AnnotatedDataFrame('Test_Data/batch_corrected_pheno
_hepato.csv',header=TRUE,row.names=1,sep=',')
pheno=pData(pheno)
batch = pheno$Batch
head(pheno)

modcombat = model.matrix(~1, data=pheno)
combat_edata = ComBat(dat=final, batch=batch, mod=modcombat,
par.prior=TRUE,ref.batch = 1,mean.only = FALSE)

combat_edata=as.data.frame(combat_edata[,colnames(combat_edata
)%in% colnames(df1)])
write.csv(combat_edata,file='Test_Data/batch_corrected_hepato_
INTENSITIES.txt',quote=FALSE)
```

Feature Selection Test

Following snippets generates expression value from the batch corrected normalized unknown test samples.

```
import pandas as pd
import numpy as np

df=pd.read_csv('test_data/batch_corrected_hepato_INTENSITIES.t
xt')
df.rename({'Unnamed: 0':'PROBEID'},axis=1,inplace=True)
print(df.shape)
df.head(1)

df1=pd.read_csv('train_data/Hepato_Selected_Probes.csv',sep='\
t')
print(df1.shape)
df1.head(1)

df=pd.merge(df,df1,on=['PROBEID'])
df.drop(['PROBEID'],axis=1,inplace=True)
print(df.shape,df.columns)
df.head(1)
```

```
df.to_csv('test_data/Hepato_Test_Expression.csv',sep='\t',inde
x=False)
```

UMAP_Model

In order to validate the correctness of genetic features, we have deployed the UMAP based clustering technique, such that normal and cancer samples are completely separated.

```
import pandas as pd
import numpy as np
import pickle
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from umap import UMAP
import matplotlib.patches as mpatches
import pickle

df=pd.read_csv('train_data/Hepato_Training_Expression.csv',sep
='\t')
df.drop(['PROBEID'],axis=1,inplace=True)
print(df.shape)
df.head(1)

df=pd.DataFrame(df.groupby(['GENE_SYMBOL']).mean()).reset_inde
x()
print(df.shape,df.columns)
df.head(3)

train_df=df[df.columns[6:]]
train_df.columns,train_df.shape

df1=pd.read_csv('test_data/Hepato_Test_Expression.csv',sep='\t
')
print(df1.shape)
df1.head(1)

df1=pd.DataFrame(df1.groupby(['GENE_SYMBOL']).mean()).reset_in
dex()
print(df1.shape,df1.columns)
df1.head(1)

val_df=df1[df1.columns[1:]]
val_df.columns,train_df.shape

X_train=train_df.values
X_train=np.transpose(X_train)
X_train.shape

X_val=val_df.values
X_val=np.transpose(X_val)
X_val.shape
```

```
scaler1=StandardScaler()
X_train_norm=scaler1.fit_transform(X_train)
X_val_norm=scaler1.transform(X_val)


pickle.dump(scaler1,open('Hepato_StandardScaler.sav', 'wb'))

X_val_norm=np.concatenate([X_train_norm,X_val_norm])
X_val_norm.shape

umap1 = UMAP(metric='cosine',random_state=1)
mapper=umap1.fit(X_train_norm)
pickle.dump(umap1,open('Hepato_UMAP.sav', 'wb'))

X_train_umap=umap1.transform(X_train_norm)

X_val_umap=umap1.transform(X_val_norm)

label_val=['green', 'green', 'green','green', 'green',
'green', 'green','green',
           'red', 'red', 'red','red', 'red', 'red',
'red','red',
           'orange', 'orange', 'orange','orange', 'orange',
'orange','orange','orange','orange',
         'orange', 'orange', 'orange','orange','orange',
'orange','orange','orange','orange','orange',
         'orange', 'orange', 'orange','orange',

'blue','blue','blue','blue','blue','blue','blue','blue',
           ]

classes=['PRIM','HCC','HEPG_TEST','PRIM_TEST']
fig, ax = plt.subplots(1, figsize=(10, 10))
plt.scatter(*X_val_umap.T, s=90,c=np.array(label_val),
alpha=1.0)

class_colours = ['b','r','orange','g']
recs = []
for i in range(0,len(class_colours)):

recs.append(mpatches.Rectangle((0,0),1,1,fc=class_colours[i]))
plt.legend(recs,classes,loc=1,fontsize=25)
```

Logistic_Regression

Here we have deployed logistic regression based supervised classification technique using selected genetic features which are validated with unknown test samples as screened in previous steps.

```
import pandas as pd
import numpy as np
```

```
import pickle
%matplotlib inline
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression

df=pd.read_csv('train_data/Hepato_Training_Expression.csv',sep
='\t')
df.drop(['PROBEID'],axis=1,inplace=True)
print(df.shape)
df.head(1)

df=pd.DataFrame(df.groupby(['GENE_SYMBOL']).mean()).reset_inde
x()
print(df.shape,df.columns)
df.head(3)

train_df=df[df.columns[6:]]
train_df.columns,train_df.shape

df1=pd.read_csv('test_data/Hepato_Test_Expression.csv',sep='\t
')
print(df1.shape)
df1.head(1)

df1=pd.DataFrame(df1.groupby(['GENE_SYMBOL']).mean()).reset_in
dex()
print(df1.shape,df1.columns)
df1.head(1)

val_df=df1[df1.columns[1:]]
val_df.columns,train_df.shape

train_df=train_df.copy()
print(train_df.columns)
train_df=train_df.transpose()
train_df.reset_index(inplace=True)
label=[0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1]
for i in train_df.index.values:
    train_df.loc[i,'LABEL']=label[i]
train_df.LABEL=train_df.LABEL.astype(np.int16)
train_df=train_df.sample(frac=1,random_state=1)
train_df.drop(['index'],axis=1,inplace=True)
train_df

train_data=train_df.values
X_train=train_data[:,:-1]
Y_train=train_data[:,-1]
X_train.shape,Y_train.shape

X_val=val_df.values
X_val=np.transpose(X_val)
Y_val=np.array([1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
```

```
                0,0,0,0,0,0,0,0])
X_val.shape

scaler1=pickle.load(open('Hepato_StandardScaler.sav', 'rb'))
X_val_norm=scaler1.transform(X_val)
X_train_norm=scaler1.transform(X_train)


model                                                          =
LogisticRegression(random_state=1,class_weight='balanced')
model.fit(X_train_norm,Y_train)

print(str(model.score(X_train_norm,Y_train)))
print(str(model.score(X_val_norm,Y_val)))
pickle.dump(model,open('Hepato_Logistic_Regression.sav',
'wb'))

feat_names=df.GENE_SYMBOL.values
X_Train_LR_DF=pd.DataFrame(data=X_train_norm,columns=feat_name
s)
X_Train_LR_DF.head()

feat_imp                                                       =
pd.DataFrame(zip(X_Train_LR_DF.columns,model.coef_[0]*np.power
(10,2)), columns=['Feature','Value'])
feat_imp.to_csv('Hepato_Feature_Importance.csv',index=False,se
p='\t')
feat_imp.head()
```

4.2.3. Verification with Unknown Dataset:

In order to validate the efficacy of the developed prediction model, we collected additional datasets corresponding to both normal and cancerous samples of same microarray platform ([HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array). These datasets were designated as the validation dataset (**Table 2**) and served the purpose of evaluating the model's ability to accurately distinguish between cancerous and normal samples.

Germ Layer	Sample Type	Accession ID	No. of Samples
Ectoderm	Neuron from healthy donor	GSE161355	5
		GSE68605	3
		GSE19332	6
	Neuroblastoma cell line	GSE67338	3
		GSE130747	2
		GSE115406	3
		GSE65459	2
Mesoderm	Cardiomyocytes from healthy donor	GSE120895	8
		GSE76701	4
		GSE51472	5
	Sarcoma Cell line	GSE56112	6
		GSE21306	1
		GSE10021	1

Endoderm	Hepatocyte from Healthy donor	GSE154198	3
		GSE115410	3
		GSE108047	12
	Hepatocellular carcinoma cell line	GSE190473	4
		GSE183790	2
		GSE99663	1
		GSE125180	3
		GSE78736	4

Table 2: Validation Dataset

4.2.4. Application of the Prediction Model:

In the application phase, we used unknown datasets corresponding to iPSC derived neurons, hepatocytes, and cardiomyocytes (as provided in **Table 3**). These datasets were subjected to evaluation using the prediction model.

Germ Layer	iPSC derivative Type	Accession ID	No. of Samples
Ectoderm	iPSC derived Neuron	GSE76830 GSE106382 GSE59051 GSE96826	8 4 2 3
Mesoderm	iPSC derived Cardiomyocytes	GSE62203 GSE119559 GSE90000	4 3 1
Endoderm	iPSC derived Hepatocytes	GSE75888 GSE80279	6 4

Table 3: Unknown Samples used for testing the prediction model

4.2.5. Robustness of the Prediction Model:

To assess the model's versatility to predict the status of iPSC derivatives beyond neuronal, cardiac and hepatocytic lineages, we collected data from iPSC derived cells corresponding to other lineages (**Table 4**) to be tested by our prediction model. All these datasets belong to the same microarray platform ([HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array).

Sample Type	Accession ID	No. of Samples
RPE	GSE80985	9
iPSC derived RPE	GSE96853	2
	GSE43257	1
	GSE64264	2
Astrocytes	GSE15824	3
	GSE145935	3
iPSC derived Astrocytes	GSE87385	2

Table 4: Dataset comprising cell types other than neuron, cardiomyocyte and hepatocyte

4.3. Results and Discussion

4.3.1. Pathway analysis leads to feature selection:

When comparing the 116 pathways shared among ectodermal, mesodermal, and endodermal lineages with the 178 shared pathways identified through analysis of three

types of reprogramming methods, we identified 41 overlapping pathways. Upon closer examination of these pathways, we identified 12 pathways directly implicated in general carcinogenic processes. Furthermore, we extracted the genes associated with these 12 pathways corresponding to each germ layer separately **[Figure 2]**. These gene sets served as the selected input features for constructing the prediction model.

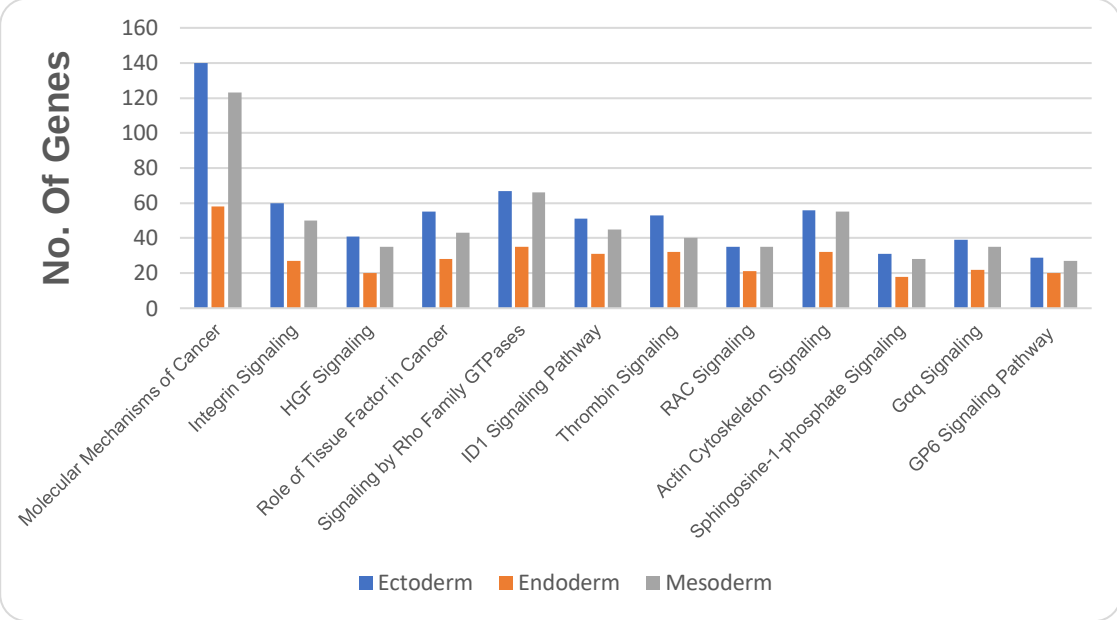


Figure 2: No. of Genes associated with selected pathways

4.3.2. Training and validation of the model for accuracy:

Utilizing the selected features, we developed a prediction model using a supervised classification approach. This model utilized the expression profiles of the identified significant genes as training features, enabling effective discrimination between normal and cancerous samples. Visualization of the training data through UMAP clustering unveiled distinct clusters corresponding to normal and cancerous samples, providing a visual representation of the model's ability to capture the underlying patterns in the data. Moreover, analysis of the validation dataset offered valuable insights into the model's performance, demonstrating its proficiency in accurately distinguishing between cancerous and normal samples across the ectodermal, mesodermal, and endodermal lineages **[Figure 3]**.

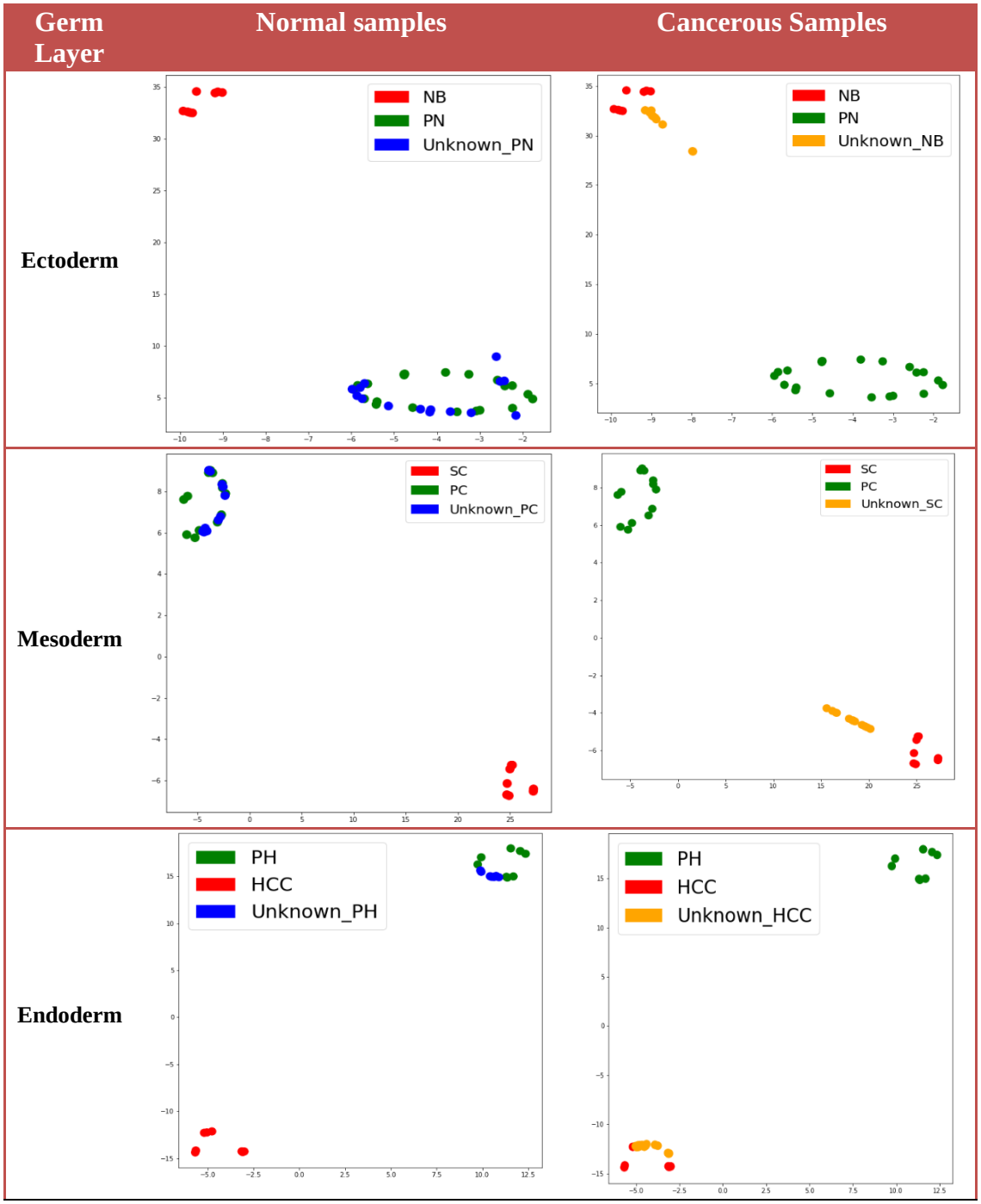


Figure 3: Validation of the model using unknown dataset; PN – Primary Neuron, NB- Neuroblastoma, PC –Primary Cardiomyocytes, SC –Sarcoma, PH – Primary Hepatocytes, HCC – Hepatocellular Carcinoma

4.3.3. Application of Prediction Model:

Our initial step involved evaluating unknown iPSC-derived datasets of neurons, hepatocytes, and cardiomyocytes. UMAP clusterings for these samples are depicted in **Figure 4**.

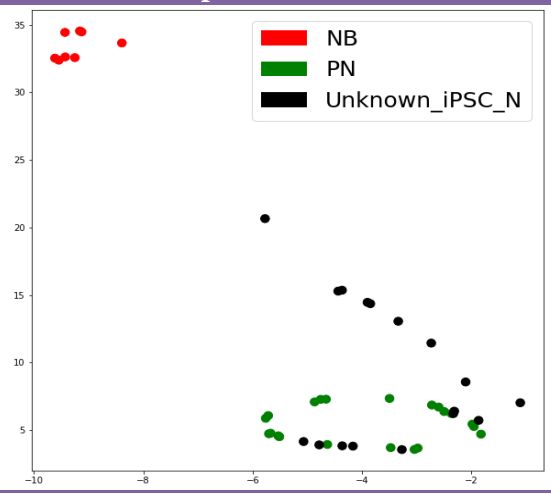
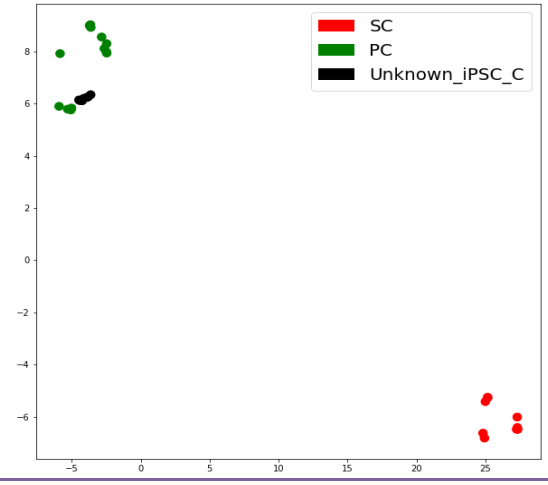
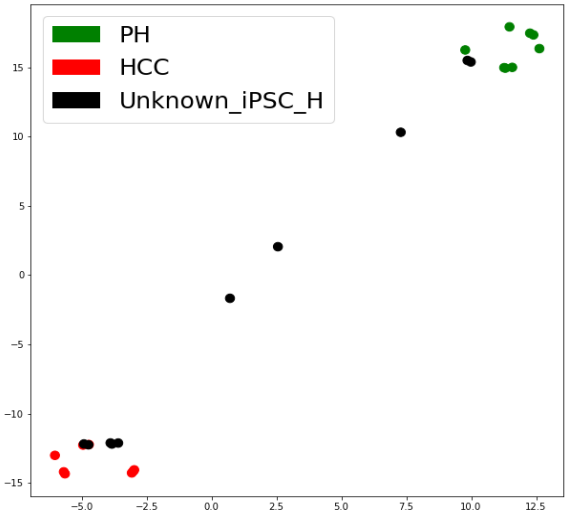
Unknown Sample Type	Sample Distribution
iPSC derived Neuron	
iPSC derived Cardiomyocytes	
iPSC derived Hepatocytes	

Figure 4: Visualization of Unknown iPSC derived samples

In order to get the exact status of the samples, logistic regression scores have been checked. It's important to note that during model training, samples having probability score to be > 0.5 are designated as cancerous and their status shows 1 whereas those

having probability score to be < 0.5 have been designated to be non-cancerous with status showing 0.

Prediction results (Table 5) for these unknown samples revealed GSM2544689 and GSM2544690 to be having oncogenic contamination as they exhibited status 1, whereas other samples displayed status 0, suggesting no oncogenic contamination within them.

Sample	Status	Probability Score
GSM2038553	0	0.037329532
GSM2038554	0	0.144169941
GSM2038555	0	0.200503664
GSM2038556	0	0.217745569
GSM2038557	0	0.120656617
GSM2038558	0	0.046749853
GSM2038559	0	0.063271107
GSM2038560	0	0.116307479
GSM2836934	0	0.02252445
GSM2836935	0	0.00272135
GSM2836936	0	0.009471057
GSM2836937	0	0.008346807
GSM1425029	0	0.00245229
GSM1648149	0	0.127553182
GSM2544689	1	0.534973524
GSM2544690	1	0.770627487
GSM2544691	0	0.491967979

Table 5: Status and Probability scores of the unknown iPSC derived neuron samples; Rows marked with same colour are the datasets coming from the same experiment

Interestingly, sample GSM2544691 from the same dataset displayed a status of 0. Upon closer examination of its probability score, it was found to be 0.49, indicating a borderline probability close to 0.5. This suggests that samples with scores around 0.5 have an equal chance of being oncogenic or non-oncogenic. Given that stem cell differentiation is a stochastic process, variations in differentiation timing and efficiency may contribute to such discrepancies. Experimentalists encountering scores close to 0.5 may need to monitor the expression of top deviated genes regularly while frequent subculturing of cells for an extended period of time, which might change the status of the cells over time.

4.3.4. Testing the Robustness of the Prediction Model:

Since no such models have been developed till today to detect the presence of oncogenicity within iPSC derived cells, we couldn't compare the efficiency of this model with other similar models. However, to ascertain the robustness of the model for providing prediction across a broader spectrum of iPSC derived cell types, other than neurons, cardiomyocytes, and hepatocytes, we expanded our investigation (based on data availability) to include additional cell types viz. RPE and astrocytes belonging to the ectoderm layer. Leveraging the same analytical platform ([HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array), we meticulously analyzed the datasets from these two cell types **[Figure 5].**

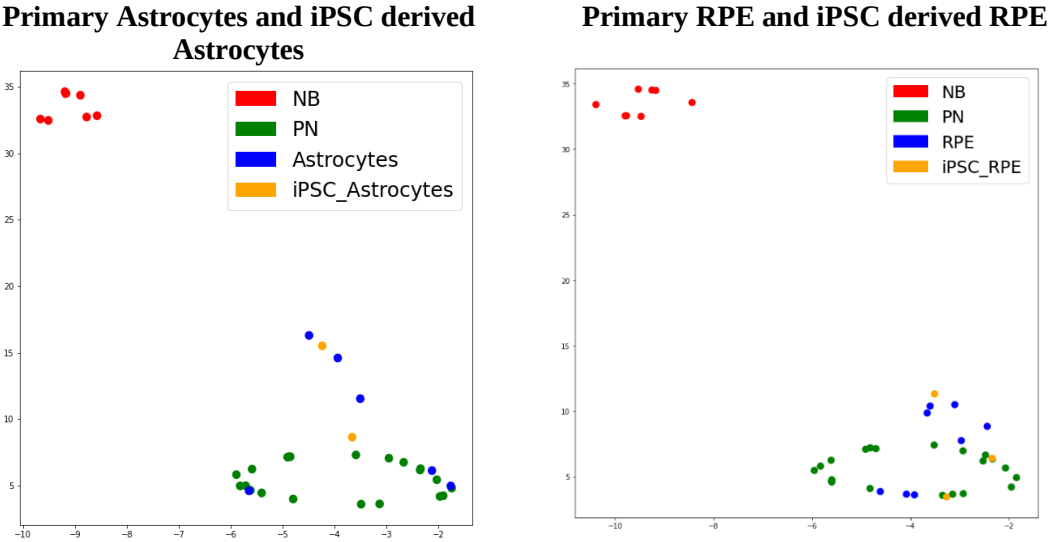


Figure 5: UMAP Clustering on primary Astrocytes and RPE along with their iPSC derivative states

Intriguingly, UMAP clustering revealed that primary astrocytes and RPE clustered with the primary samples. This once again reveals the veratility of the feature set which can distinguish between the primary cell type features and oncogenic features efficiently. The iPSC derived astrocytes and iPSC derived RPE both clusters with their primary counterparts which reveal that there are no remnant oncogenic signatures present within them.

Through this rigorous validation effort, we aim to instill confidence in researchers regarding the model's efficacy in assessing oncogenic contamination across a wide range of iPSC-derived cell types, thus facilitating safer and more reliable applications of iPSC-based therapies.

4.4. Conclusion

Our research has culminated in the development of a robust prediction model designed to assess the presence of oncogenic contamination in iPSC-derived cells. We have tested

the robustness of our model by using unknown test samples corresponding to diverse cell types other than those used for training the model. We have come across both instances where the iPSC derived cells have as well as do not have a remnant oncogenic contamination within it, revealing the variability inherent in the process of cellular differentiation.

We acknowledge that differentiation is not a deterministic process but rather a stochastic one, influenced by various factors including cell culture conditions and timing. Indeed, our observations suggest that incomplete or inadequate differentiation may contribute to the persistence of oncogenic contamination in certain samples. In our prediction model, the probability score plays a crucial role, as scores equal or close to 0.5 indicate an equal probability of oncogenic or non-oncogenic status. In such instances, it is plausible that extending the duration of differentiation or optimizing culture conditions along with monitoring the genes with high deviation scores could potentially mitigate the risk of oncogenic transformation.

There have been certain limitations in our study which needs to be addressed in future. Firstly, the reliance on microarray data limits the scope of our analysis, and future studies incorporating RNA seq as well as single cell datasets could provide a more comprehensive understanding of oncogenic contamination in iPSC derivatives as well. Additionally, our training data was constrained to three cell types each corresponding to three germ layers, due to data scarcity, highlighting the need for expanded datasets encompassing a broader range of cell types and experimental conditions. Last but not the least, due to nonavailibility of such prediction models, we were unable to compare the efficiency of our model with other similar models.

Looking ahead, further research efforts should aim to address these limitations and explore novel avenues for mitigating oncogenic contamination in iPSC-derived cells. In conclusion, despite such limitations, we expect that our developed prediction model will serve as a crucial tool for researchers to assess the status of iPSC-derived cells for potential oncogenic contamination within them prior to their subsequent application in regenerative therapy. The novelty and utility of our prediction model hold significant promise for advancing stem cell based regenerative therapy, offering researchers invaluable insights into the safety and efficacy of iPSC-based approaches. By continuing to refine and expand upon our findings, we can pave the way for safer and more effective iPSC-based therapies, ultimately benefiting patients and advancing the field of regenerative medicine.

References

1. Lemmens, M., et al., *Identification of marker genes to monitor residual iPSCs in iPSC-derived products.* Cytotherapy, 2023. **25**(1): p. 59-67.
2. Muller, F.J., et al., *A bioinformatic assay for pluripotency in human cells.* Nat Methods, 2011. **8**(4): p. 315-7.
3. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.

4. Peng, C.-Y.J., K.L. Lee, and G.M. Ingersoll, *An Introduction to Logistic Regression Analysis and Reporting.* The Journal of Educational Research, 2002. **96**(1): p. 3-14.
5. McInnes, L., J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction.* arXiv preprint arXiv:1802.03426, 2018.
6. Yang, Y., et al., *Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data.* Cell Rep, 2021. **36**(4): p. 109442.

CHAPTER 4

CHAPTER 5| piRNAQuest V.2: updating the piRNAome for silencer

Abstract:

PIWI interacting RNAs (piRNAs) have gained prominence as crucial regulators of gene expression. Following the initial release of piRNAQuest in 2014, substantial advancements in piRNA research have resulted in the annotation of numerous novel piRNAs across a diverse range of species beyond human, mouse, and rat. Responding to these developments, we have, piRNAQuest V.2, comprising 92,77,689 piRNA entries spanning 25 additional species across various phyla, alongside human, mouse, and rat. This updated database not only furnishes fundamental piRNA attributes, encompassing genomic coordinates and additional details on piRNA intersections with repeat elements, pseudogenes, and syntenic regions, but also introduces innovative features such as (i) density-based cluster prediction, (ii) piRNA expression profiles across diverse healthy and pathological systems, and (iii) piRNA target prediction. Noteworthy is the robustness of the density-based piRNA cluster identification method, eschewing parametric distribution assumptions. The piRNA expression profiles, encompassing 21 disease systems, including cancer, and 32 tissue-specific piRNA expression profiles across various species provide a comprehensive resource. Moreover, the piRNA target prediction section incorporates both predicted and curated piRNA targets within eight disease systems and developmental stages of the mouse testis. Users can visually explore the piRNA-target duplex structure and the ping-pong signature pattern for all ping-pong piRNA partners across different species. In summary, piRNAQuest V.2 stands as an updated, user-friendly database poised to serve as an invaluable resource for the exploration, retrieval, and analysis of piRNA-related information across multiple species.

Published in: RNA Biology 2022; 19(1): 12–25.

5.1. Introduction

PIWI-interacting RNAs, commonly known as piRNAs, constitute a distinctive class of small non-coding RNAs (sncRNAs) that have gained prominence for their pivotal role in regulating gene expression [1]. Operating in association with PIWI proteins, piRNAs are primarily recognized for their involvement in silencing transposable elements in the germline, thereby safeguarding genomic integrity [2]. Beyond their canonical function, piRNAs have emerged as versatile molecules implicated in various biological processes,

including epigenetic regulation, germ cell development, and potentially contributing to the etiology of certain diseases [3]. The multifaceted nature of piRNAs and their significance in diverse cellular contexts make them a subject of keen interest in contemporary RNA biology and molecular research.

piRNAs constitute a diverse category of endogenous sncRNAs [4], typically ranging in length from 25 to 33 nucleotides (nts). Distinct from other sncRNAs like miRNAs and siRNAs, piRNAs undergo biogenesis through both primary processing pathways and an amplifying ping-pong mechanism [5], independently of Dicer [6]. Primary piRNAs emanate from specific genomic loci referred to as piRNA clusters [5], which can be dual-strand clusters, producing piRNAs from both strands, or uni-strand clusters, such as the flamenco clusters in Drosophila follicle cells and murine pachytene piRNA clusters, generating piRNAs from a single DNA strand [7]. The ping-pong cycle involves the generation of sense secondary piRNAs initiated by antisense primary piRNAs, leading to the production of secondary antisense piRNAs, creating an amplifying loop in the process [5,8].

While investigations across fish, flies, and mammals have consistently revealed a conserved association between piRNAs and PIWI proteins [5,9,10], the advent of evolving sequencing technologies has unveiled variations in piRNA length across different species. In mammals, piRNAs are broadly categorized into two subclasses: pachytene (29–33 nts) and pre-pachytene (26–28 nts) [11]. Conversely, in *Caenorhabditis elegans*, these molecules are referred to as 21 U-RNA, reflecting a distinct bias toward a length of 21 nts. Despite their predominant presence in germ cells, recent studies have expanded our understanding by demonstrating piRNA expression in diverse somatic tissues, including the brain, kidney, lung, liver, stomach, testis, ovary [12-15], and notably, in various cancer types [16]. This observed breadth of piRNA expression hints at potentially broader regulatory roles beyond the traditionally recognized germline functions.

In the maintenance of genomic integrity within germ cell lineages, the robust expression of PIWI proteins in germ and stem cells [17] plays a pivotal role in the regulation of transposon activity, serving as a defensive mechanism [18]. Studies have illuminated the critical function of MIWI, a PIWI homolog in mice, in orchestrating this process, as mutations in MIWI have been linked to male infertility and the over expression of retrotransposon transcripts [19]. Similar observations have been documented in flies [6]. The collaborative action of PIWI proteins and piRNAs within piRNA-induced silencing complexes (piRISCs) constitutes the PIWI-piRNA pathway, where transposons are silenced through complementary base-pair recognition between piRNAs and transposons, culminating in the endonucleolytic cleavage of the target [20,21].

5.2. Current databases, their limitations, and the need for an updated database

Various existing databases, including piRNABank [22], piRBase [23], piRNAdb (https://www.pirnadb.org), piRTarBase [24] and the piRNA Cluster Database [25], offer valuable information on piRNAs across multiple species. Notably, piRBase stands out as

a manually curated repository providing comprehensive piRNA information for diverse species and specific disease systems. The piRNA Cluster Database (https://www.smallrnagroup.uni-mainz.de/piRNAclusterDB/) is a specialized resource dedicated to piRNA clusters, predicting clusters using proTRAC [26]. Although piRDisease V1.0 [27] contains piRNA records associated with different diseases, it is currently not accessible. Despite the extensive efforts in piRNA research, several areas remain unexplored, such as the potential associations between piRNAs and long noncoding RNAs (lncRNAs), or the influence of genomic elements within their loci on their functional roles. Addressing these gaps, the initial release of piRNAQuest [28] delves into these less-explored domains of the piRNAome, providing detailed piRNA information for three species—human, rat, and mouse.

Despite the numerous computational tools dedicated to characterizing novel piRNAs [29,30], the functional implications of these identified molecules remain largely unclear. Consequently, there is a crucial need to discern potential piRNA targets, particularly those associated with diseases. Another notable gap in existing databases is the lack of comprehensive curation for both predicted and validated piRNA targets, encompassing messenger RNAs (mRNAs) and lncRNAs. This deficiency hampers the accessibility and accuracy of information in this critical area of piRNA research. Moreover, the identification of piRNA clusters, recognized as hotspots for piRNA biogenesis, presents a formidable challenge within the current landscape of piRNA studies. Addressing these knowledge gaps is paramount for advancing our understanding of piRNA functionality and their potential roles in disease contexts.

This study introduces piRNAQuest V.2 as an updated and extended version of the original piRNAQuest, featuring several enhanced functionalities. Notable additions include: (i) an in-depth analysis covering 25 new species in addition to human, mouse, and rat, extending the scope of the database, (ii) the implementation of a density-based clustering approach [31] aimed at identifying 'hotspots of piRNA expression' or 'piRNA clusters.' Given the variability of piRNA distribution across genomic locations in different species, this approach offers a novel perspective for identifying biologically relevant piRNA clusters, (iii) the incorporation of tissue-specific piRNA expression profiles across diverse species, (iv) an exploration of piRNA expression in various disease systems, with a particular focus on different cancer types. Given the emerging recognition of piRNAs in disease progression and diagnosis [32-37], understanding their expression in different tissues and disease contexts is critical to unraveling their efficacy and potential mechanisms of action in cancer, (v) piRNA target prediction within both mRNAs and lncRNAs. This feature enhances the database's utility by facilitating the identification of key contributors to disease development through the elucidation of piRNA-mRNA and piRNA-lncRNA interactions. Overall, piRNAQuest V.2 serves as an advanced and comprehensive resource for investigating the roles of piRNAs in diverse biological contexts, including disease progression and development.

In addition to the aforementioned comprehensive features, we have revamped the 'Tools' section of the database in piRNAQuest V.2. This section now empowers users to predict

piRNA clusters with customized parameters, examine ping-pong pattern overlaps in their sequences, and predict piRNA targets using the miRanda algorithm [38].

In summary, piRNAQuest V.2 stands as a user-friendly and versatile database tailored for the multi-species exploration, search, and retrieval of piRNA-related information. The inclusion of piRNA expression profiles in both normal tissues and cancer, coupled with detailed insights into piRNA targets, positions this database as a valuable and indispensable resource for piRNA researchers, offering a platform for advanced analyses and comprehensive investigations into the intricate world of piRNAs. The database is available for free access at http://dibresources.jcbose.ac.in/zhumur/pirnaquest2.

5.3. Methods

5.3.1. Input dataset:

Sequencing data for small RNA reads corresponding to 28 species were acquired from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) [39] as well as from supplementary information of different studies. Table 1 provides details on genome builds, genome annotation and repeat annotation availability, and the count of piRNAs for each species. Updates include transitioning from hg19 to hg38 and rn5.0 to rn6.0 for human and rat genome builds, respectively. Normal tissues and disease-specific data were also obtained for expression analysis. Data are gathered in various formats such as fasta, gtf, and bed from respective sources. Repeat elements and RefSeq annotated 5'UTR, 3'UTR, exon, intron, and CDS information are sourced from UCSC [40]. miRNA information is retrieved from miRBase 22 release. Annotated piRNA sequences are downloaded in .fasta format from the NCBI [41]. LncRNA information is sourced from LncRBase V.2 [42].

Species	Phylum/Class	Genome Build	No of unique piRNAs
Human	Chordata/Mammalia	GRCh38/hg38	51509
Mouse		GRCm38/mm10	2357673
Rat		RGSC 6.0/rn6	190191
Chinese hamster		CHOK1S_HZDv1/criGriChoV2	25626
Cow		UMD_3.1.1/bosTau8	32147
Pig		Sscrofa11.1/susScr11	94818
Platypus		ASM227v2/ornAna2	1400
Rabbit		Broad/oryCun2	161564
Dog		CanFam3.1/canFam3	1036427
Horse		Broad/equCab2	1461782
Marmoset		WUGSC 3.2/calJac3	1068209
Crab-eating macaque		Macaca_fascicularis_5.0/macFas5	5380
Rhesus macaque		BCM Mmul_8.0.1/rheMac8	55341
Northern Tree Shrew		Broad/tupBel1	32701

Big Brown Bat		EptFus1.0	228382
Zebrafish	Chordata/Actinopterygii	GRCz10/danRer10	154562
Chicken	Chordata/Aves	Gallus_gallus-5.0/galGal5	118784
Xenopus laevis	Chordata/Amphibia	Xenopus_laevis_v2/xenLae2	1386083
Xenopus tropicalis		JGI 9.1/xenTro9	11466
C. elegans	Nematoda/Chromadorea	WS220/ce11	15365
Drosophila melanogaster	Euarthropoda/Insecta	BDGP Release 6 + ISO1 MT/dm6	250103
Drosophila erecta		droEre1	12055
Drosophila yakuba		WUGSC 7.1/droYak2	26892
Drosophila virilis		droVir2	20192
Aedes	Arthropoda/Insecta	Aedes_aegypti.AaegL3	253408
Silkworm		Bombyx_mori.ASM15162v1	223085
California sea hare	Mollusca/Gastropoda	Broad 2.0/aplCal1	221
Starlet sea anemone	Cnidaria/Anthozoa	ASM20922v1	2323

Table 1: Genome builds information and number of piRNA entries corresponding to different species for piRNAQuest V.2

5.3.2. Data processing and ID assignment:

The process of assigning distinct identifiers to non-redundant piRNA entries closely aligns with the methodology employed in the piRNAQuest [28] platform. Initially, sequencing data underwent alignment to their respective genomes. Subsequently, reads mapping to other sncRNAs were filtered out, and those predicted to be piRNAs were selected using an internally developed script. Following this, non-redundant reads underwent realignment with the reference genome to achieve complete alignment without mismatches. These reads were then annotated with unique piRNAQuest IDs, adopting the format [three-letter abbreviation of the species name] piRNA[number]. It is pertinent to note that the annotation IDs remain consistent for human and mouse, as established in the preceding version. The sole deviation lies in the annotation of rat compared to the prior version, where it lacked the three-letter abbreviation of the species name in its annotation. Users are directed to the Help menu's ID conversion section for human, mouse, and rat to access the previous IDs annotated in this version of the database. To investigate the genomic distribution of piRNAs, we employed in-house perl scripts, consistent with those utilized in the previous version, to explore their localization within genes, intergenic regions, introns, coding sequences (CDS), untranslated regions (UTRs), repeat elements, and pseudogenes.

5.3.3. piRNA cluster prediction using density based approach:

The previous cluster prediction protocol [43] faced a limitation by employing a fixed-length window size for all species, neglecting the variations in read distribution among different species. To address this limitation, we have implemented a novel approach using the density-based clustering algorithm DBSCAN [31]. This Python-based, in-house

protocol aims to identify piRNA clusters by taking into account the distinctive read distribution patterns of piRNAs across the genome.

Clustering parameters: Two crucial parameters, namely 'Eps' (epsilon) and 'MinReads,' play a pivotal role in identifying candidate piRNA clusters. 'Eps' is defined as the distance from a read to its nearest neighboring point, while 'MinReads' represents the minimum number of reads within this 'eps' distance. To establish optimal clustering parameters, we conducted k-dist analysis, calculating the inter-distance between annotated piRNAs. This involved determining the distance between each mapped read and its kth nearest neighboring read, referred to as 'k-dist.' The 'count versus distance' plot was then generated, revealing a distinct valley until the k-dist followed a uniform distribution. The distance at which the graph exhibited an asymptotic decrease for a given value of k is denoted as 'eps.' Essentially, 'eps' represents the distance that repeats most frequently, carrying the highest probability of defining cluster boundaries containing at least the 'MinReads' (representing the minimum number of reads within the cluster). Once 'Eps' and 'MinReads' parameters are established for each chromosome corresponding to each species, clusters are detected from the coordinate file of annotated piRNAs.

Cluster score: To quantify piRNA enrichment within each cluster, we computed a cluster score for each piRNA cluster using the following method:

$$\text{Cluster score} = \frac{\text{Total no of piRNAs in the cluster}}{\text{Minimum no of piRNAs needed to form the cluster (kth value)}} \quad (1)$$

We additionally assessed the strand specificity of the clusters by examining the directionality of the constituent piRNAs within each cluster. If a cluster comprises both sense and antisense piRNAs, it is classified as a 'dual-strand cluster.' On the other hand, if it includes exclusively sense piRNAs or antisense piRNAs, it is categorized as a 'uni-strand cluster.'

piRNA clusters locations and characteristic motifs: (i) We employed in-house Perl scripts to examine potential overlaps of piRNA clusters with coding genes, lncRNAs, and repeat regions. (ii) Notably, piRNAs exhibit a pronounced inclination to form clusters in syntenic regions of the genome [44]. To investigate this, we acquired syntenic regions from UCSC [40] and conducted a search for piRNA clusters within the corresponding syntenic regions across different species. (iii) Utilizing MEME, we conducted an analysis to identify any significant motifs present within the piRNA clusters [45].

5.3.4. Ping-pong Signatures within piRNAs:

The alternative pathway of piRNA biogenesis, known as ping pong amplification, reveals a unique sequence based characteristic in piRNAs. Specifically, there is a 10-nucleotide overlap observed between the antisense and sense piRNAs. To identify and visualize this distinctive ping-pong signature pattern, an in-house Python script has been developed for analysis.

5.3.5. piRNA expression pattern in Normal and Disease systems:

We obtained small RNA sequencing data for various tissue types from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo). The dataset comprises 243 samples, encompassing 32 types of normal tissue samples across different species. In addition to the normal dataset, we analyzed 211 samples corresponding to 21 types of disease data, including 16 different cancer datasets. To scrutinize the expression profile of piRNAs in both normal and disease systems, we employed BLAST [46] and developed in-house Perl scripts. Subsequently, the expression levels of individual piRNAs in a sample were normalized using counts per million (CPM) and screened based on the z-score [47] falling within the range of -3 to +3. Users can visualize the expression of the 200 most abundant piRNAs in each dataset. Furthermore, we examined the distribution of each piRNA across all normal tissues or disease systems, representing the data graphically to enhance understanding of their expression patterns within different biological systems.

5.3.6. Target prediction:

Predictions of piRNA target pairs have been conducted, specifically identifying interactions between upregulated piRNAs and downregulated lncRNAs and mRNAs, and vice versa. The miRanda tool [38] was employed to predict piRNA targets within lncRNAs, utilizing sequences retrieved from LncRBase V.2 [48]. For mRNA targets, sequences from the 3' untranslated region (UTR) were obtained from UCSC [40]. The criteria for target prediction involved a target score of 170 and an energy threshold of -20 kcal/mol [49]. To enhance the utility of our database, we have integrated The Human Protein Atlas [50] and Pathway Commons [51] databases, linking them to the targeted genes for subsequent pathway and pathology-based analyses. Additionally, our database encompasses a collection of experimentally validated piRNA targets for human, mouse, and *C. elegans*, meticulously curated from published reports.

The comprehensive workflow detailing these processes has been illustrated in Figure 1.

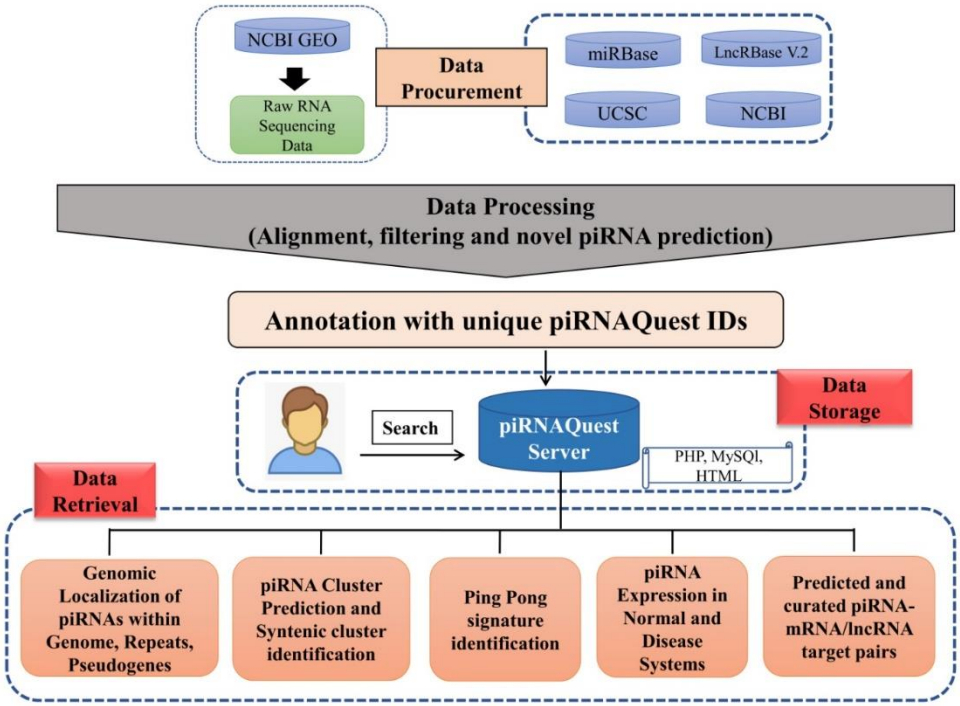


Figure 1: Workflow of piRNAQuest V.2

5.4. Results

piRNAQuest V.2, an updated version of the original piRNAQuest, consolidates information on 92,77,689 piRNAs corresponding to 28 diverse species. This set encompasses 25 newly included species, in addition to human, mouse, and rat, and spans a phylogenetic spectrum from nematodes to chordates [Figure 2].

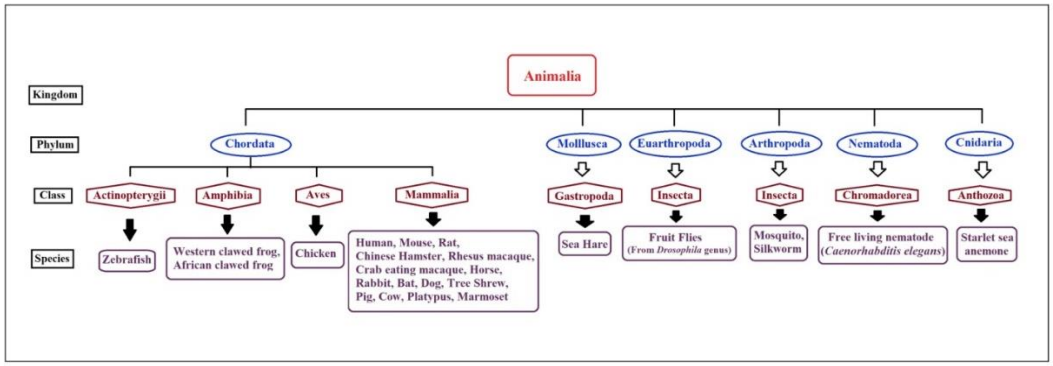


Figure 2: Taxonomical representation of species incorporated in piRNAQuest V.2, Common names are used for the species with their respective Kingdom, Phylum and Class

Beyond species coverage, this upgraded version introduces several additional features, augmenting its significance relative to other piRNA databases. A comprehensive tabulation of the updated features in piRNAQuest V.2 compared to the earlier version is presented in Table 2.

Database Content	piRNAQuest	piRNAQuest V.2
Number of species	3	28
piRNA entries	9,98,585	92,77,689
Chromosomal distribution	Yes	Yes
Association with Gene	Yes	Yes
Association with Pseudogene	Yes	Yes
Association with Repeat elements	Yes	Yes
Cluster information	Yes (Lau et al. Method), for 3 species	Yes (Density based clustering approach), for 19 species
Association of clusters with Genomic regions	Yes	Yes
Syntenic piRNA clusters	Yes	Yes
Ping-pong piRNAs	Yes	Yes
Ping-pong pattern Visualization	No	Yes
Tissue specific expression	Yes (Tissue type – 6, Dataset - 9)	Yes (Tissue type – 64, Dataset - 242)
piRNA disease association	No	Yes (21 types of diseases included)
Graphical representation of expression	No	Yes (For 64 tissue types and 18 disease systems)
Predicted piRNA - mRNA target pairs	No	Yes (For seven types of cancer, asthenozoospermia and mouse testis)
Predicted piRNA targets within lncRNAs	No	Yes (For seven types of cancers)
piRNA target genes (literature curated)	No	Yes (for Human, Mouse and *C. elegans*)
Target prediction tool	No	Yes
Ping-pong overlap prediction tool	No	Yes

Table 2: Comparison of features between piRNAQuest and piRNAQuest V.2

Additionally, a feature-wise comparative analysis between piRNAQuest V.2 and other piRNA databases is presented in Figure 3. Notably, among the 28 species, nine species, including Chinese hamster, Sea hare, Tree Shrew, Brown Bat, Silkworm, Mosquito, *Drosophila virilis*, *Drosophila erecta*, and Starlet sea anemone, lack annotations. Consequently, genomic localization and associated features are provided solely for the remaining 19 species. Of these, four species, namely Chinese hamster, Tree Shrew, *Drosophila virilis*, and *Drosophila erecta*, have repeat-associated piRNAs identified, leveraging available repeat annotations from UCSC. Visualization of this information in graphical format can be accessed through the 'Statistics' submenu under the 'Help Menu' of the database.

			Overlapping Pattern			Biogenesis					Target Prediction	
	No. Of Species	Chromosomal Distribution	Gene	Repeat	Pseudogene	Pingpong Pattern	Cluster information	Syntenic Clusters	Tissue Specific expression	Disease association	mRNA	lncRNA
piRNABank	4	Yes	Yes	Yes	No	No	Yes	No	No	No	No	No
piRBase	21	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes
piRNAdb	6	Yes	Yes	Yes	No	No	Yes	No	Yes	No	Yes	No
piRTarBase	2	No	Yes	No	No	No	No	No	No	No	Yes	No
piRNA Cluster Database	51	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No
piRDisease V1.0	3	Not Accessible database				No	No	No	No	Yes	Yes	No
piRNAQuest V.2	28	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Figure 3: Comparison of the features of different database with piRNAQuest V.2

5.4.1. Genomic Distribution of piRNAs:

piRNAQuest V.2 provides comprehensive information on piRNAs across multiple species, exhibiting a notable increase in piRNA entries compared to its predecessor. The distribution of piRNAs on chromosomes has been specifically mapped for 19 out of the 28 species covered (as indicated earlier). Notably, human chromosome 15 stands out as having the highest abundance of piRNAs, consistent with our findings in the previous version of piRNAQuest. It is noteworthy that chromosome 15 in humans is recognized for harboring a substantial number of low-copy repeats, commonly referred to as duplicons [52]. These duplicons facilitate nonhomologous recombination events, contributing to genome instability [53]. The presence of a significant piRNA population on chromosome 15 may serve as a strategic response to counteract such challenges associated with genome instability. piRNAs are well-documented for their pivotal role in safeguarding genome integrity [6], and their accumulation on this particular chromosome could be a protective mechanism against the adverse effects of nonhomologous recombination events. Moreover, the highest concentration of piRNAs is found in chromosome 7 of mice and chromosome 1 of rats. Notably, among the recently incorporated species, chromosome IV of *C. elegans* (previously reported [54]) and Chromosome 2R of *Drosophila melanogaster* exhibit the greatest abundance of piRNAs.

The prevalence of piRNAs in intergenic regions surpasses that in intronic regions across various species, as depicted in Figure 4. Notably, intergenic piRNAs play a crucial role in early embryonic development [55], given their abundance in these regions. Additionally, intergenic regions are known to harbor lncRNA loci [56]. Consequently, an investigation was conducted to identify the presence of lncRNA loci coinciding with piRNA clusters composed of intergenic piRNAs, and the results are presented later in the 'piRNA clusters overlapping with lncRNAs' section. In contrast, piRNA abundance in the 3' UTR, 5' UTR, and CDS regions is generally lower, except for specific cases such as zebrafish, where there is high piRNA abundance in the CDS region, and *C. elegans*, where there is notable piRNA abundance in the 3' UTR and 5' UTR regions [Figure 4].

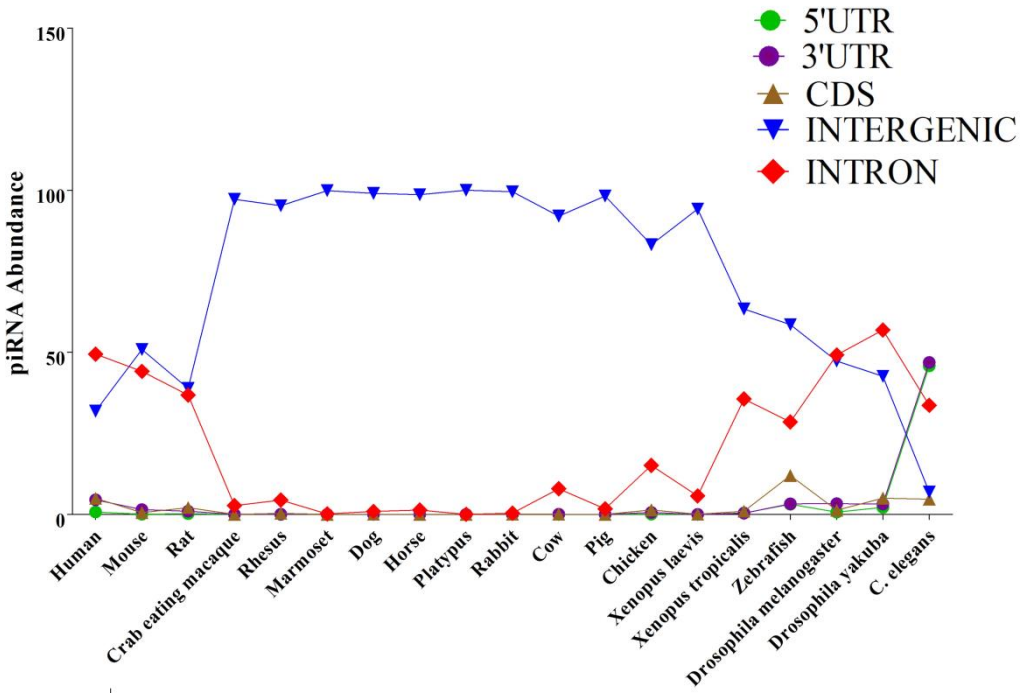


Figure 4 : The distribution of piRNAs across various genomic locations (Different genomic locations are highlighted in different colours; Abbreviations used : UTR-untranslated region, CDS - coding DNA sequence)

Furthermore, it has been discovered that pseudogenes regulate the stability of their corresponding genes through small RNA-mediated silencing [57]. Recent findings indicate that pachytene piRNAs originating from pseudogenes directly influence the regulation of their parent genes [58]. This discovery prompted an exploration of the presence of piRNAs within pseudogenes across 16 species with available pseudogene information [59]. Significant overlaps between piRNAs and pseudogenes were identified in multiple species, with human exhibiting the maximum overlap [Figure 5]. Notably, pseudogene-derived piRNAs have been detected in mature human sperm cells, suggesting their role in regulating the expression of their parent genes in male germline cells [60].

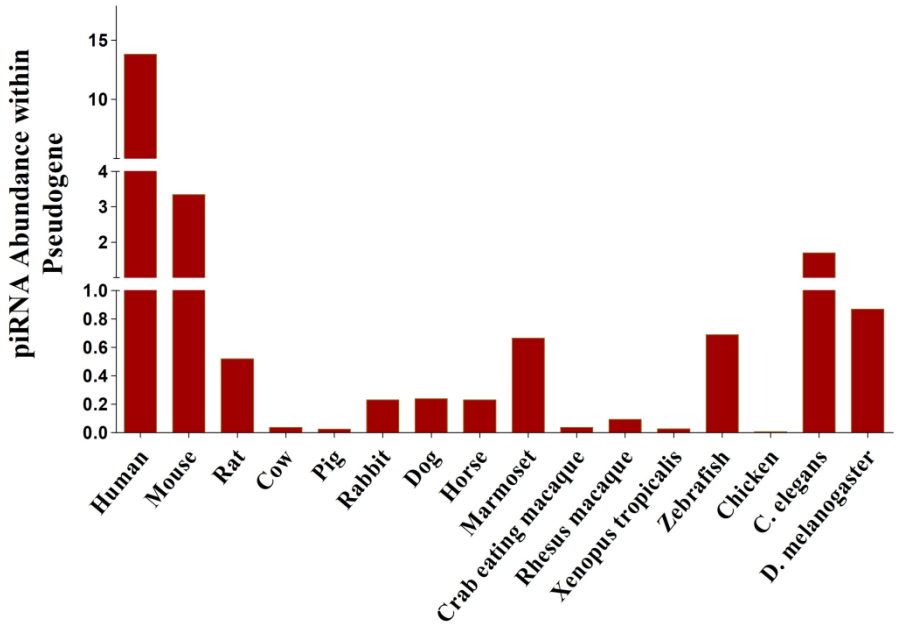


Figure 5: The distribution of piRNAs within pseudogenes

5.4.2. Repeat associated distribution of piRNAs:

piRNAs, originating from repetitive genomic regions, play a crucial role in silencing transposons and regulating global gene expression during embryonic development [61] in both insects and mice [5,62]. The genomic mapping of piRNA loci for all 19 species reveals their association with seven major categories of repeat elements: LINE, SINE, Simple repeat, DNA, Low complexity, Satellite, and LTR, as illustrated in Figure 6.

Vandewege et al. [63] documented a robust piRNA response in mammals like dogs and horses, where these piRNAs are predominantly housed within the SINE repeat regions, abundant in these species. Our database also reports the enrichment of SINE repeat-associated piRNA loci for 12 mammalian species. Furthermore, there is an overlap of piRNA loci with the LTR repeat family in human, mouse, and rat, whereas amphibian piRNAs tend to overlap with DNA and Simple repeat families.In the order Diptera, characterized by insects, Petersen et al. [64] observed an abundance of LTR repeats within genomic loci corresponding to transposable elements. Our study aligns with this observation, revealing piRNA-enriched regions in Diptera overlapping with the LTR repeat family. The presence of such repeat regions within piRNA loci holds significant implications, as demonstrated by Halbach et al. [61], who reported that satellite repeats modulate global gene expression through piRNA-mediated gene silencing, a process crucial for the embryonic development of Aedes.

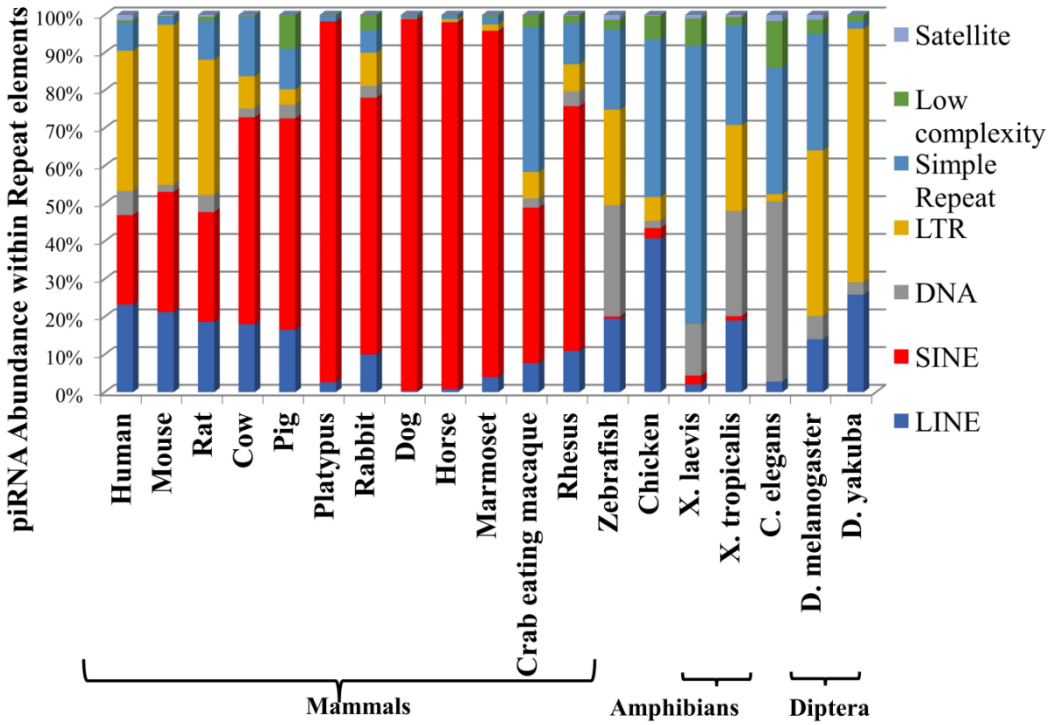


Figure 6 : Distribution of piRNAs within Repeat family (Different repeat families are highlighted in different colours; Abbreviations used : LTR - Long-terminal repeat, LINE - Long interspersed nuclear elements, SINE - Short interspersed nuclear elements)

5.4.3. piRNA Biogenesis: Clusters and ping-pong signature:

piRNA clusters, recognized as pivotal sites for piRNA biogenesis, were initially identified using a fixed window length of 20 kilobases (kb), as per the method outlined by Lau et al. [43] in the first version of piRNAQuest. However, subsequent research, notably by Rosenkranz in 2016, revealed that piRNA clusters are not uniformly distributed across chromosomes and are not directly proportional to the chromosome length [25]. The variation in piRNA read distribution across different chromosomes emphasizes the importance of adapting the methodology for detecting piRNA clusters. In response to these insights, the latest version of the database, piRNAQuest V.2, has transitioned to a density-based clustering approach [31]. This approach is designed to dynamically identify piRNA clusters, taking into account the varying piRNA read distribution across the genome. This updated strategy has proven to be effective, particularly in recognizing clusters successfully in chicken germ cells [65]. By avoiding a fixed window size, the density-based clustering approach enhances the precision and adaptability of piRNA cluster identification, contributing to a more accurate representation of piRNA biogenesis dynamics.

The highest number of piRNA clusters was identified in chromosome 15 for humans. This observation aligns with our previous findings in piRNAQuest. Additionally, for *C. elegans*, our analysis concurs with existing reports, indicating that the maximum number of piRNA clusters is located in chromosome IV. It is noteworthy that in *C. elegans*,

although the function of these clusters remains unknown, there is a documented prevalence of clusters within chromosome IV [54]. Interestingly, in the context of sex determining chromosomes, it has been reported that 'X' chromosomal piRNAs primarily originate from clusters, in contrast to 'Y' chromosomal piRNAs [25]. Our analyses extend this observation to humans, mice, and rats, revealing a higher number of piRNA clusters in the 'X' chromosomes compared to the 'Y' chromosome. This information underscores the potential significance of piRNA clusters in 'X' chromosomes, hinting at their involvement in diverse biological processes related to sex determination and gene regulation.

We investigated the arrangement of piRNA clusters within the lncRNA loci extracted from LncRBase V.2. As indicated in the initial point of this results section, a noteworthy concurrence of piRNA clusters was noted with intergenic lncRNAs [Figure 7], representing transcripts originating between two gene loci. This observation aligns with prior research findings [66]. Additionally, our examination extended to the juxtaposition of piRNA clusters with repeat regions [Figure 8], revealing a parallel trend to that identified in the distribution of piRNAs across repeat elements (depicted in Figure 6).

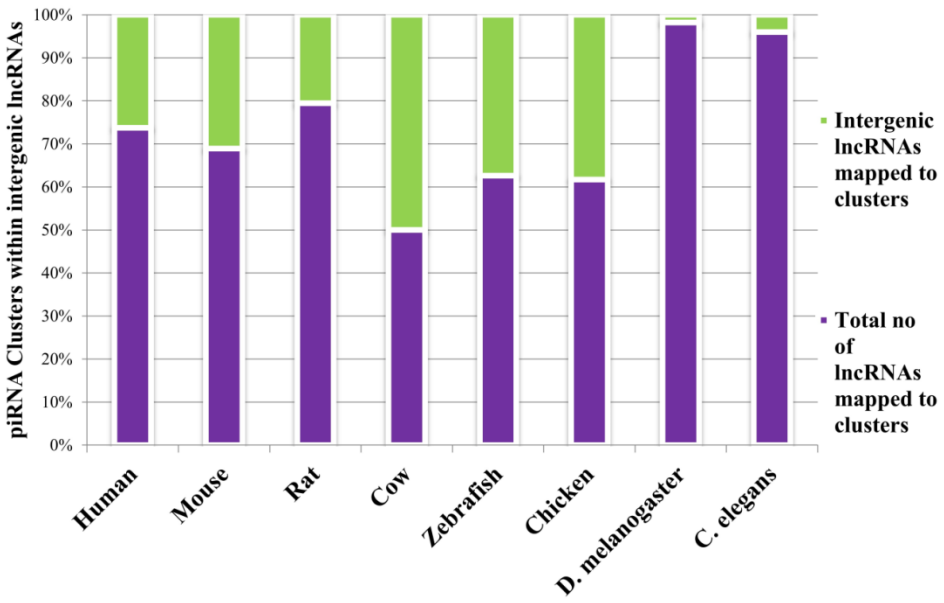


Figure 7: Distribution of piRNA Clusters within intergenic lncRNAs

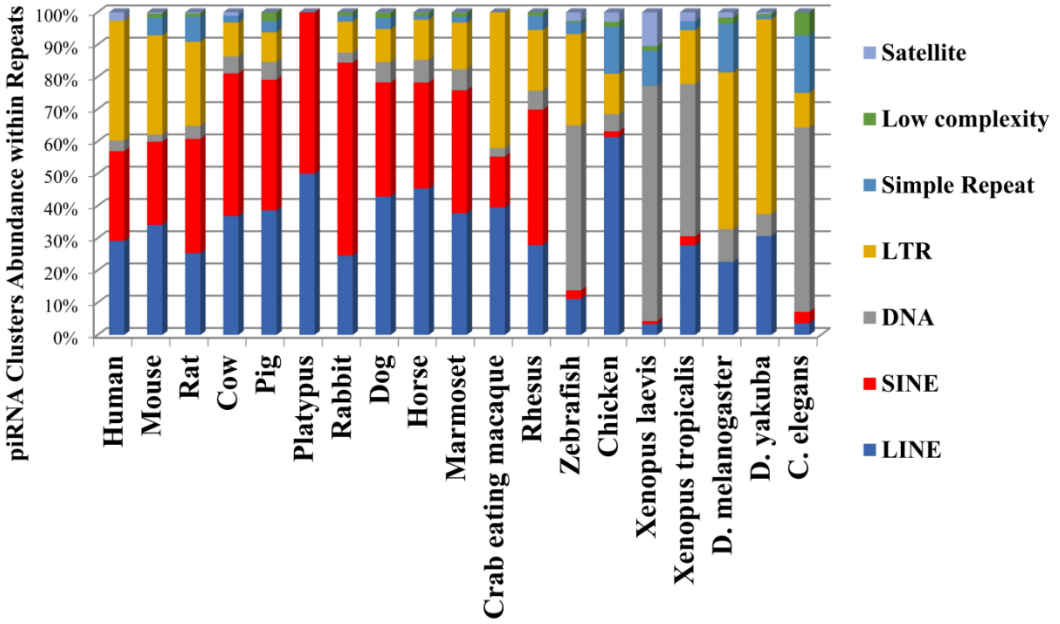


Figure 8: Distribution of piRNA Clusters within Repeat elements

Distinctive motifs have been discerned for each of the piRNA clusters. These exceptionally conserved motifs present within the piRNA clusters offer insights into potentially shared piRNA binding sites within their respective target genes. Furthermore, piRNAs originating from clusters derived from coding gene regions have the capacity to modulate the expression of their 'host' genes [67]. In numerous species, a notable percentage of total piRNA clusters have been identified to exhibit overlap with coding regions [Figure 9].

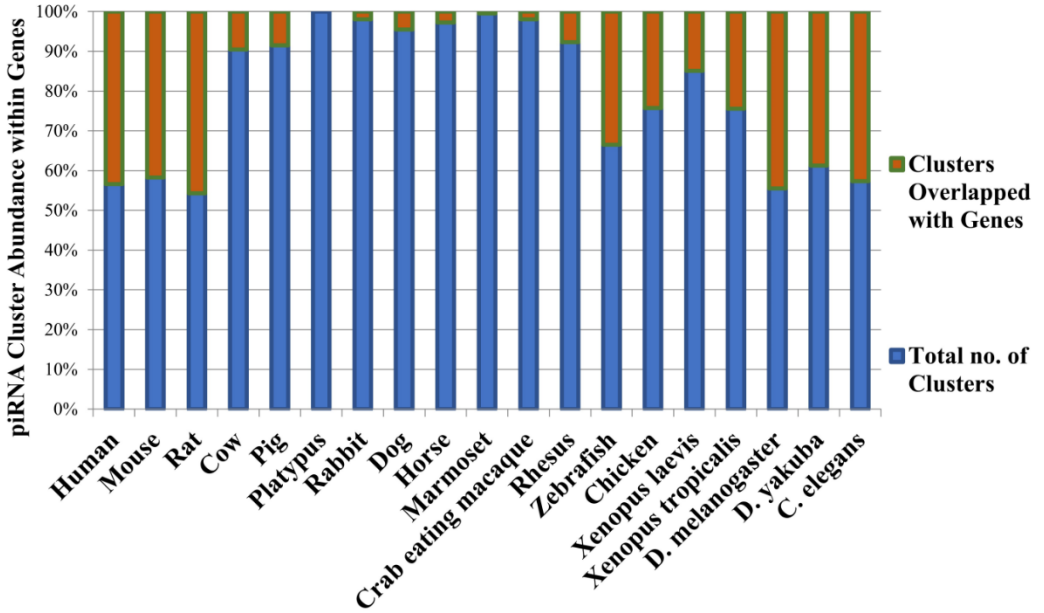


Figure 9: Distribution of piRNA Clusters within coding regions

piRNAs also undergo secondary biogenesis through the ping-pong amplification loop. Investigations in Drosophila have revealed that somatic piRNAs typically lack the ping-pong pattern, indicating that this amplification loop predominantly operates in germline cells [68,69]. We assessed the distribution of ping-pong piRNAs across different chromosomes. Notably, in humans, Chromosome 15 exhibits a pronounced prevalence of ping-pong piRNAs, a finding corroborated by recent work by Ray and Pandey [70]. Over 50% of human ping-pong piRNAs manifest overlap with protein-coding genes, suggesting a role in piRNA-mediated gene regulation [71]. Interestingly, less than 10% of these piRNAs overlap with repeat elements, with the SINE repeat family being the most prevalent. Contrary to prior findings by Das et al. [72] indicating an absence of ping-pong amplification in nematodes, our analysis unexpectedly identifies 509 ping-pong piRNAs in Chromosome IV of *C. elegans*, hinting at a potential role of the ping-pong loop in nematode biology.

5.4.4. piRNA expression across different tissue:

Initially, piRNA expression was confined to germline cells [44], but progressively, their presence has been detected in somatic cells, with the somatic piRNA pathway identified as a regulator of germline transpositions [73]. Consequently, we analyzed 243 small RNA sequencing samples encompassing 32 tissue types across 25 species, including 13 human tissue types. In humans, the highest abundance of piRNAs was observed in the brain, followed by the colon, testis, and spermatozoa, suggesting a broader role for piRNAs not only in germline cells but also in various somatic cells.

Previous research demonstrated the existence of piRNA complexes in mouse dendritic spines within the brain, and the knockdown of these piRNAs resulted in reduced spine density in the axons [55]. Recent studies also indicate that brain-associated piRNAs play a role in suppressing retrotransposons, influencing brain pathology [74]. The length distribution of piRNAs has been linked to the age of individuals within a species; for instance, in Drosophila, piRNA length decreases with age [75]. Moreover, the loss of methyltransferase results in piRNA instability, reduced piRNA length, and volume, leading to male sterility during spermatogenesis [76]. Notably, our study revealed the presence of approximately 36 nucleotide-long piRNAs in human sperm samples, a length uncommonly expressed in most somatic cells.

5.4.5. Expression of piRNAs in disease systems:

Advancements in pathological research have underscored the significance of piRNAs in the context of various diseases. Aberrant expressions of piRNAs and PIWI proteins have been identified in numerous cancer systems, emphasizing their potential as innovative biomarkers in therapeutic research [16]. Recent findings propose that dysregulation of the piRNA pathways may disrupt the genomic stability of neurons, contributing to various neurodegenerative disorders [77]. Given the pivotal role of piRNA biogenesis-associated genes in spermatogenesis, mutations in these genes may lead to male infertility [78]. Furthermore, piRNAs have demonstrated regulatory effects on Th2 cell development by suppressing IL-4, thereby inhibiting allergic inflammation and asthma [79]. Additionally, they exhibit specific binding partners in synovial fibroblasts,

suggesting their involvement in inflammatory processes such as Rheumatoid Arthritis [80]. In this study, we conducted an analysis involving 211 samples representing 21 disease types, including 16 types of cancer. The distribution of piRNAs [Figure 10A] across various cancers highlights a heightened contribution of piRNAs in germ cell cancers such as ovarian and testicular cancer. This observation aligns with the established role of piRNAs in maintaining germ cells [81].

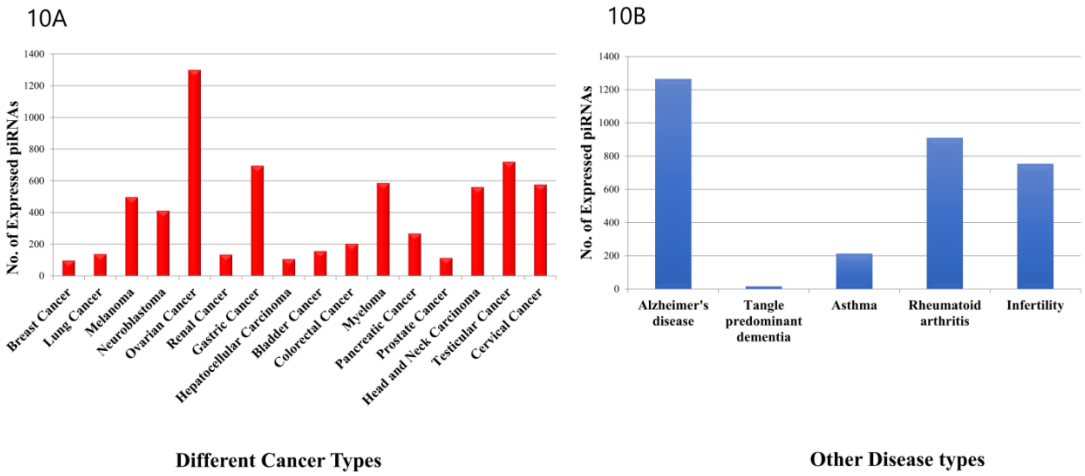


Figure 10 : piRNA expression profile in (A) different cancer systems and (B) other diseases

Among various medical conditions [Figure 10B], we identified 1274 piRNAs, with hsa_piRNA_425 exhibiting high abundance and hsa_piRNA_28207 showing comparatively low abundance in Alzheimer's disease, consistent with previous reports [34]. The expression levels of piRNAs in asthma and rheumatoid arthritis were found to be 278 and 910, respectively. Notably, our analysis revealed a distinctive observation concerning the length of piRNAs. In our investigation, we observed longer piRNAs in sperm samples, with a maximum length of 32 nucleotides in infertile samples, underscoring the relevance of piRNA length in spermatogenesis [76].

In addition to characterizing the overall expression patterns of piRNAs across diverse diseases, we conducted a differential expression analysis using DESeq [82] to investigate the distinct regulatory mechanisms of piRNAs in seven cancer types and asthenozoospermia, where both test and control datasets were available. Table 3 presents the count of differentially expressed piRNAs, and certain piRNAs exhibited expression patterns consistent with findings in the existing literature. For instance, hsa_piRNA_9871 and hsa_piRNA_27200 were found to be highly expressed in breast and lung cancer, respectively [83], aligning with previous reports and our study. Furthermore, the upregulated piRNAs, such as hsa_piRNA_7806 and hsa_piRNA_31147, known to promote proliferation and invasiveness in colon [84] and renal cancer [85], were also observed to be upregulated in our analysis. This convergence reinforces the reliability of

our differential expression analysis and underscores the potential roles of these piRNAs in the progression of specific cancers.

	Differentially expressed piRNAs		Differentially expressed Genes		Differentially expressed lncRNAs	
	Up	**Down**	**Up**	**Down**	**Up**	**Down**
			Different Cancer systems			
Breast cancer	253	133	955	1165	1087	615
Lung cancer	269	349	517	513	53	82
Ovarian cancer	208	376	3300	2376	335	1441
Renal cancer	177	724	197	162	85	40
Hepatocellular carcinoma	352	155	116	125	169	171
Colon cancer	366	475	344	361	79	119
Prostate cancer	413	328	2135	2368	504	179
			Other disease			
Asthenozoospermia	1622	2170	318	133	2957	1062
			Different developmental stages of mouse testis			
10 dpp	923	439	2173	2429	1445	1442
16.5 dpc	260	143	7281	6498	3209	2430

Table 3: Differentially expressed piRNAs, genes and lncRNAs in different cancer systems, Asthenozoospermia and in different developmental stages of Mouse Testis

5.4.6. piRNA and mRNA/lncRNA target interaction:

In addition to their well-documented role in mediating the cleavage of transposable elements, piRNAs have been recognized for their ability to target mRNAs and lncRNAs, thereby influencing the regulation of gene expression. The regulatory impact of piRNAs on mRNAs has been the subject of thorough investigation in scientific research [34,86,87]. Analogous to the cleavage mechanism observed in mRNAs, the PIWI-piRNA complex exhibits the capability to target lncRNAs, a phenomenon observed across various organisms [88]. Studies have reported a reciprocal relationship between the expression levels of the target and the corresponding targeting piRNA - a decrease in the expression of the target corresponds to an increase in the expression of the targeting piRNA, and vice versa [89]. This reciprocal regulation adds another layer to the intricate and multifaceted regulatory functions of the PIWI-piRNA complex in cellular processes. Consequently, to enhance the precision of target prediction, we systematically screened piRNAs and mRNAs, as well as piRNAs and lncRNAs, focusing on those pairs with a negative correlation in expression. This stringent criterion led us to narrow our analysis to cancer datasets where both long and small RNA sequencing datasets were available. As a result, we successfully predicted piRNA-mRNA and piRNA-lncRNA interactions for seven cancer systems, namely lung, breast, renal, hepatocellular, ovarian, prostate, and colorectal. The differential analysis was executed using the 'New Tuxedo' protocol

[90]. Sequence-based target prediction was conducted using miRanda, with separate analyses performed on tissue and cell line data. To elucidate the role of piRNAs in distinct developmental stages, small RNA data from various developmental stages of mouse testis, specifically 10 days post-partum (dpp) and 16.5 days postcoitum (dpc), were compared to those from six months old adult mouse testis. Additionally, piRNAs were analyzed in the context of asthenozoospermia, a condition characterized by reduced sperm motility in semen samples. Table 3 provides a comprehensive list of differentially expressed mRNAs, lncRNAs, and piRNAs associated with these analyses. The conclusive set of piRNA-mRNA and piRNA-lncRNA target pairs for the seven cancer types is illustrated in Figure 11(a,b), respectively. Figure 11(c,d) provides an overview of the count of piRNA targets within mRNAs and lncRNAs, respectively, across two developmental stages of mouse testis. Additionally, we have curated experimentally validated piRNA-mRNA target pairs for human, mouse, and *C. elegans*, contributing valuable insights into the conserved regulatory interactions across species. This comprehensive presentation enhances our understanding of the regulatory landscape orchestrated by piRNAs in diverse biological contexts and provides a foundation for further exploration of their functional implications.

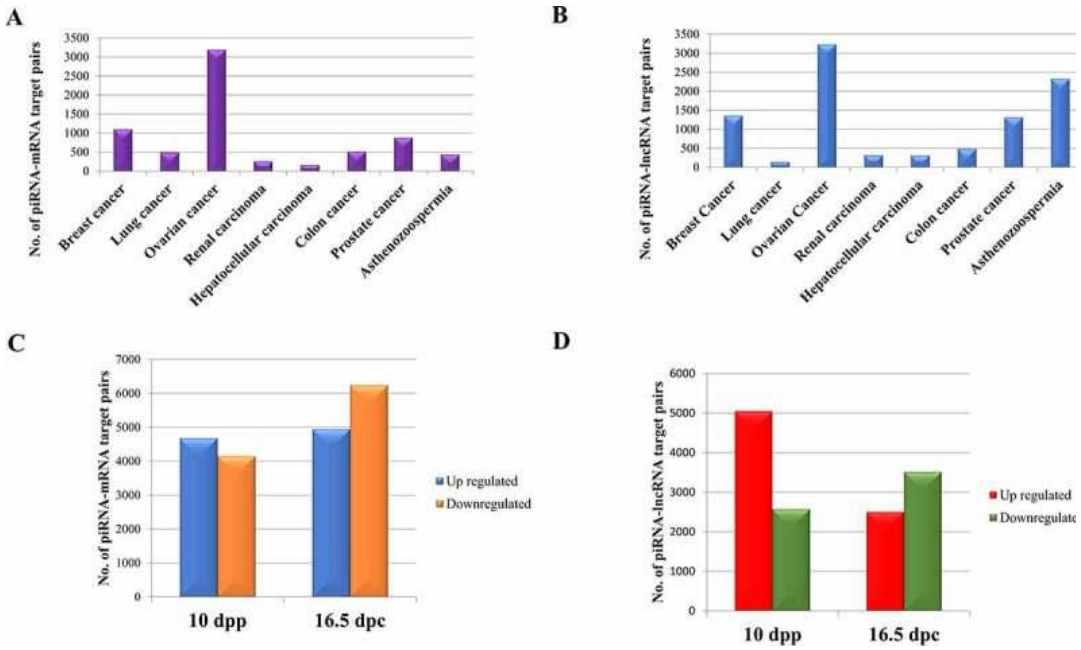


Figure 11: The predicted piRNA targets in (A) protein coding genes and (B) LncRNAs Within disease systems; (C) protein coding genes and (D) LncRNAs across different developmental stages of mouse testis

5.5. Database execution

In piRNAQuest V.2, a user's query is processed through simple searching options based on their desired selection criteria. The retrieved information is then presented on the web interface, allowing users to explore additional details. The general information page

displays basic details related to the queried piRNAs, and users have the option to delve into further genomic details, as illustrated in Figure 12.

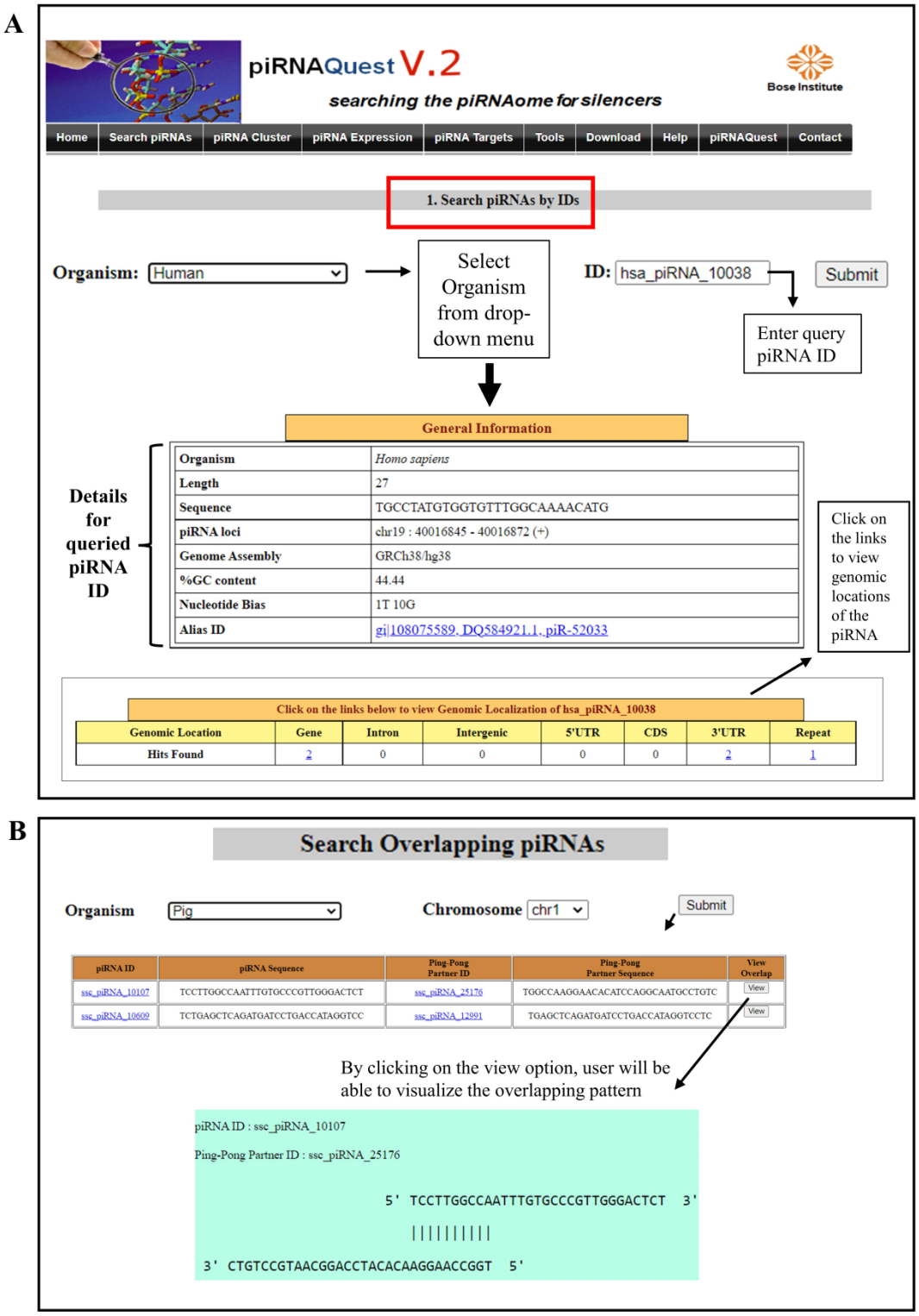


Figure 12: Web interfaces for access of piRNAQuest V.2 : (A) search options through a piRNA ID and the corresponding result page; (B) search options for pingpong piRNAs and visualizing its pattern

5.5.1. Search and output options:

(a) The following options are under the "Search piRNAs" menu

1. Search by Species Name: Users can search piRNAs by selecting a particular species name with the help of previous or next buttons

2. Search by piRNA Accession ID/Chromosomal Co-ordinates: Users can browse by piRNA accession ID for detailed information (piRNA sequence, its length, its NCBI ID (if any), %GC content, piRNA position corresponding to the genome build along with its genomic localization within genes, introns, CDS, 3' UTR, 5' UTR, intergenic regions, and repetitive elements) of selected species. Using desired chromosomal co-ordinates, users can also get above mentioned information about piRNAs.

3. Search piRNAs by Sequence: Users can obtain piRNA information by providing piRNA sequences. The sequence length should be greater than at least 20 nucleotides.

4. Search piRNA within Genes: User can browse piRNAs present within Genes by providing a Gene Name corresponding to the selected species. The result page will show the piRNAs whose loci overlap with this particular gene. User can also search for piRNAs within Genes of a particular species by providing chromosomal coordinates corresponding to that species.

To find piRNAs in pseudogene, user need to provide their desired chromosomal location only, corresponding to selected species.

5. Search piRNA within repeat regions: Users can search for piRNAs whose loci get mapped within repeats corresponding to genomic locations (viz. 3' UTR, 5' UTR, introns, CDS, intergenic regions) for a particular Repeat Family. Users can also search for repeat-associated piRNAs selecting their desired chromosomal location.

6. Search piRNAs with Ping-Pong features: User can search for overlapping piRNAs of 10nt within a particular chromosome of a particular species by selecting chromosome number corresponding to that species [Figure 12B].

(b) Search "piRNA clusters"

1. Search clusters by chromosomal co-ordinates: Users can obtain piRNA clusters by submitting a particular chromosomal location. This will fetch cluster loci, total number of piRNAs within the cluster, cluster score, cluster strandedness, prevalence of these piRNAs in minus/plus strand, and the corresponding characteristic motif of the cluster in that location. The link on the motif navigates to the website (https://meme-suite.org/meme/) where one can perform further study on the motif.

2. Search mRNAs/lncRNAs/Repeats within piRNA Clusters: Users can check piRNA clusters are overlapped with mRNA/lncRNA loci or with the repeat elements.

3. Search piRNA Clusters in Syntenic Regions: Users can search for piRNA clusters overlapping with syntenic regions by choosing a particular chromosomes of target and query organisms.

(c) Browse "piRNA Expression"

1. Search Tissue specific expression: User can browse different dataset for piRNA expression pattern by selecting tissue type and will be able to see the top 200 most abundantly expressed piRNAs corresponding to the dataset by submitting the view option.

2. Search Disease specific expression: User can also retrieve same information as above for different disease systems.
There is another search option where under user can check the expression level of a particular piRNA in normal tissues of selected species or that in human disease systems **[Figure 13A]**.

(d) Search "piRNA Targets"

1. Search Predicted Targets: Users can search piRNA targets in negatively correlated mRNA/lncRNA dataset by selecting the disease type and the mode of regulation of the piRNA. After clicking the details button, user will be able to get the detailed prediction result and visualize the piRNA-target duplex structure **[Figure 13B]**.

2. Search Curated Targets: In this part, users can find piRNA-gene target pairs which are obtained from literature.

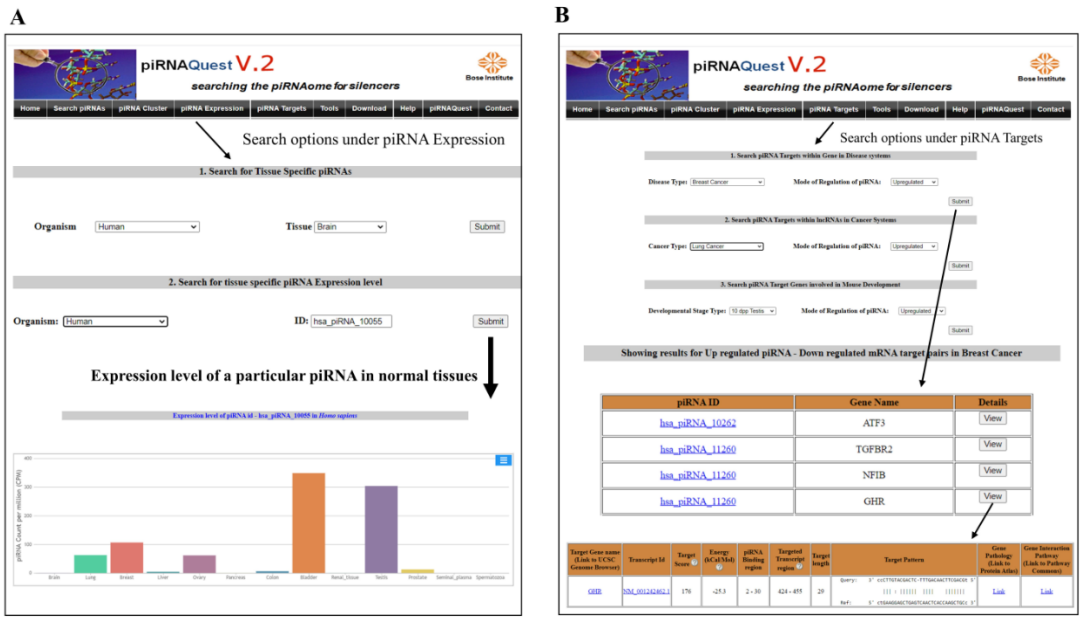


Figure 13: Web interfaces for access of piRNAQuest V.2 : (A) tissue wise expression values of individual piRNAQuest IDs with the corresponding output and (B) search options with detailed output for piRNA target prediction.

5.5.2. Tools:

1. Dynamic piRNA cluster detection: This tool can detect piRNA clusters where user can set parameters of their own, like the Eps distance, chromosomal coordinate, and MinReads.

2. piRNA Target prediction: Users can provide the piRNA and target sequences according to their data along with their desired energy parameters and threshold target score to predict piRNA targets.

3. Ping-pong signature detection: Users can visualize ping-pong signature pattern by providing piRNA sequences in .fasta format.

5.5.3. Availability:

piRNAQuest V.2 is available at http://dibresources.jcbose.ac.in/zhumur/pirnaquest2/. Files can be freely accessed and downloaded.

5.6. Discussion

The number of identified piRNAs has significantly increased across different species and cell types since the initial release of piRNAQuest in 2014. The primary objective behind developing piRNAQuest was to create a non-redundant and comprehensive catalog of piRNAs in humans, mice, and rats to enhance our understanding of piRNA genomic localization, their overlaps with genomic elements, and their associations with other lncRNAs. While initial reports primarily emphasized the main functions of piRNAs in transposon silencing [62] and maintaining gene integrity, particularly in germline cells [17], subsequent research has identified their presence and functions in somatic cells across various species [5,6]. In light of the increasing significance of the diverse functions of piRNAs, extending beyond transposon silencing to include gene expression regulation, we have introduced the new version of it as piRNAQuest V.2. This updated version expands our study to encompass 25 new species, covering different phyla or classes in addition to those included in the previous version. In piRNAQuest V.2, we have retained the features from the earlier version and introduced several new aspects. Notably, we delve into the directionality of piRNA clusters, explore piRNA expression patterns in normal tissues and disease systems, and investigate their targets among both protein-coding genes and lncRNAs. These new dimensions aim to open up novel avenues for piRNA research, providing a more comprehensive understanding of their roles and regulatory networks across diverse biological contexts.

Over time, numerous studies have elucidated the primary biogenesis mechanism of piRNAs from piRNA clusters, with several protocols developed for their identification. However, the non-uniform distribution of piRNAs across chromosomes prompted the adoption of a density-based clustering approach to identify piRNA clusters. This approach enhances our understanding of piRNA distribution throughout the genome and the formation of clusters, which serve as 'hotspots' for primary biogenesis. Additionally, secondary biogenesis through ping-pong amplification is a crucial mechanism for piRNA generation and plays a significant role in the silencing of their targets. To emphasize this aspect, we conducted an analysis of ping-pong overlap among piRNAs and have provided options for visualizing the ping-pong signature within the piRNAs. In the human genome, we observed the highest abundance of piRNAs on chromosome 15, coinciding with the presence of the maximum number of piRNA clusters and ping-pong piRNAs.

Furthermore, the analysis of piRNA expression profiles in various normal and disease systems is crucial for understanding piRNA-mediated gene regulation in those contexts. In this version, we have integrated the piRNA expression profiles of 21 disease systems alongside data from several normal tissues corresponding to different species. Given the differential regulation of piRNAs between disease and normal conditions, a decrease in the expression levels of the target should correspond to an increase in the expression levels of the targeting piRNA, and vice versa. Capitalizing on this opportunity to unveil the connection between piRNA expression and disease occurrence, we have predicted potential piRNA targets that may serve as promising biomarkers for early diagnosis and act as therapeutic targets, particularly in diseases like cancer. Additionally, to illustrate the involvement of piRNAs in different developmental stages, we have predicted piRNA targets within mRNAs and lncRNAs in various developmental stages of mouse testis.

In summary, the newly incorporated features, combined with the existing ones, make piRNAQuest V.2 a user-friendly and comprehensive database for piRNAs. Our future goal is to regularly update the database with newly annotated piRNAs, along with introducing novel features, to consistently contribute to the expanding knowledge base on piRNAs.

References:

1. Kaikkonen MU, Lam MT, Glass CK. Non-coding RNAs as regulators of gene expression and epigenetics. Cardiovascular research. 2011 Jun 1;90(3):430-40.
2. Toth KF, Pezic D, Stuwe E, et al. The piRNA Pathway Guards the Germline Genome Against Transposable Elements. Advances in experimental medicine and biology. 2016;886:51-77.
3. Zhang Q, Zhu Y, Cao X, et al. The epigenetic regulatory mechanism of PIWI/piRNAs in human cancers. Molecular cancer. 2023 Mar 7;22(1):45.
4. Han Li C, Chen Y. Small and Long Non-Coding RNAs: Novel Targets in Perspective Cancer Therapy. Current genomics. 2015 Oct;16(5):319-26.
5. Brennecke J, Aravin AA, Stark A, et al. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell. 2007 Mar 23;128(6):1089-103.
6. Siomi MC, Sato K, Pezic D, et al. PIWI-interacting small RNAs: the vanguard of genome defence. Nature reviews Molecular cell biology. 2011 Apr;12(4):246-58.
7. Czech B, Hannon GJ. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. Trends in biochemical sciences. 2016 Apr;41(4):324-337.
8. Gunawardane LS, Saito K, Nishida KM, et al. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. Science. 2007 Mar 16;315(5818):1587-90.
9. Aravin A, Gaidatzis D, Pfeffer S, et al. A novel class of small RNAs bind to MILI protein in mouse testes. Nature. 2006 Jul 13;442(7099):203-7.
10. Houwing S, Kamminga LM, Berezikov E, et al. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell. 2007 Apr 6;129(1):69-82.
11. Aravin AA, Hannon GJ, Brennecke J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science. 2007 Nov 2;318(5851):761-4.

12. Fu A, Jacobs DI, Hoffman AE, et al. PIWI-interacting RNA 021285 is involved in breast tumorigenesis possibly by remodeling the cancer epigenome. Carcinogenesis. 2015 Oct;36(10):1094-102.
13. Ortogero N, Schuster AS, Oliver DK, et al. A novel class of somatic small RNAs similar to germ cell pachytene PIWI-interacting small RNAs. The Journal of biological chemistry. 2014 Nov 21;289(47):32824-34.
14. Williams Z, Morozov P, Mihailovic A, et al. Discovery and Characterization of piRNAs in the Human Fetal Ovary. Cell reports. 2015 Oct 27;13(4):854-863.
15. Huang X, Yuan T, Tschannen M, et al. Characterization of human plasma-derived exosomal RNAs by deep sequencing. BMC genomics. 2013 May 10;14:319.
16. Liu Y, Dou M, Song X, et al. The emerging role of the piRNA/piwi complex in cancer. Molecular cancer. 2019 Aug 9;18(1):123.
17. Thomson T, Lin H. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. Annual review of cell and developmental biology. 2009;25:355-76.
18. Kalmykova AI, Klenov MS, Gvozdev VA. Argonaute protein PIWI controls mobilization of retrotransposons in the Drosophila male germline. Nucleic acids research. 2005;33(6):2052-9.
19. Reuter M, Berninger P, Chuma S, et al. Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature. 2011 Nov 27;480(7376):264-7.
20. Ishizu H, Siomi H, Siomi MC. Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. Genes & development. 2012 Nov 1;26(21):2361-73.
21. Weick EM, Miska EA. piRNAs: from biogenesis to function. Development. 2014 Sep;141(18):3458-71.
22. Sai Lakshmi S, Agrawal S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic acids research. 2008 Jan;36(Database issue):D173-7.
23. Wang J, Zhang P, Lu Y, et al. piRBase: a comprehensive database of piRNA sequences. Nucleic acids research. 2019 Jan 8;47(D1):D175-D180.
24. Wu WS, Brown JS, Chen TT, et al. piRTarBase: a database of piRNA targeting sites and their roles in gene regulation. Nucleic acids research. 2019 Jan 8;47(D1):D181-D187.
25. Rosenkranz D. piRNA cluster database: a web resource for piRNA producing loci. Nucleic acids research. 2016 Jan 4;44(D1):D223-30.
26. Rosenkranz D, Zischler H. proTRAC--a software for probabilistic piRNA cluster detection, visualization and analysis. BMC bioinformatics. 2012 Jan 10;13:5.
27. Muhammad A, Waheed R, Khan NA, et al. piRDisease v1.0: a manually curated database for piRNA associated diseases. Database : the journal of biological databases and curation. 2019 Jan 1;2019.
28. Sarkar A, Maji RK, Saha S, et al. piRNAQuest: searching the piRNAome for silencers. BMC genomics. 2014 Jul 4;15:555.
29. Monga I, Banerjee I. Computational Identification of piRNAs Using Features Based on RNA Sequence, Structure, Thermodynamic and Physicochemical Properties. Current genomics. 2019 Nov;20(7):508-518.
30. Wang K, Hoeksema J, Liang C. piRNN: deep learning algorithm for piRNA prediction. PeerJ. 2018;6:e5429.
31. M. Ester HPK, J. Sander, X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD. 1996;96:226–231.
32. Quek C, Bellingham SA, Jung CH, et al. Defining the purity of exosomes required for diagnostic profiling of small RNA suitable for biomarker discovery. RNA biology. 2017 Feb;14(2):245-258.

33. Bachmayr-Heyda A, Auer K, Sukhbaatar N, et al. Small RNAs and the competing endogenous RNA network in high grade serous ovarian cancer tumor spread. Oncotarget. 2016 Jun 28;7(26):39640-39653.
34. Roy J, Sarkar A, Parida S, et al. Small RNA sequencing revealed dysregulated piRNAs in Alzheimer's disease and their probable role in pathogenesis. Molecular bioSystems. 2017 Feb 28;13(3):565-576.
35. Li Y, Wu X, Gao H, et al. Piwi-Interacting RNAs (piRNAs) Are Dysregulated in Renal Cell Carcinoma and Associated with Tumor Metastasis and Cancer-Specific Survival. Molecular medicine. 2015 May 13;21(1):381-8.
36. Zhang W, Yao G, Wang J, et al. ncRPheno: a comprehensive database platform for identification and validation of disease related noncoding RNAs. RNA biology. 2020 Jul;17(7):943-955.
37. Zhang W, Zeng B, Yang M, et al. ncRNAVar: A Manually Curated Database for Identification of Noncoding RNA Variants Associated with Human Diseases. Journal of molecular biology. 2021 May 28;433(11):166727.
38. John B, Enright AJ, Aravin A, et al. Human MicroRNA targets. PLoS biology. 2004 Nov;2(11):e363.
39. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods in enzymology. 2006;411:352-69.
40. Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC Table Browser data retrieval tool. Nucleic acids research. 2004 Jan 1;32(Database issue):D493-6.
41. Geer LY, Marchler-Bauer A, Geer RC, et al. The NCBI BioSystems database. Nucleic acids research. 2010 Jan;38(Database issue):D492-6.
42. Das T, Deb A, Parida S, et al. LncRBase V.2: an updated resource for multispecies lncRNAs and ClinicLSNP hosting genetic variants in lncRNAs for cancer patients. RNA biology. 2021 Aug;18(8):1136-1151.
43. Lau NC, Seto AG, Kim J, et al. Characterization of the piRNA complex from rat testes. Science. 2006 Jul 21;313(5785):363-7.
44. Girard A, Sachidanandam R, Hannon GJ, et al. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature. 2006 Jul 13;442(7099):199-202.
45. Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. Nucleic acids research. 2009 Jul;37(Web Server issue):W202-8.
46. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. Journal of molecular biology. 1990 Oct 5;215(3):403-10.
47. Hazra A, Gogtay N. Biostatistics Series Module 1: Basics of Biostatistics. Indian journal of dermatology. 2016 Jan-Feb;61(1):10-20.
48. Das T, Deb A, Parida S, et al. LncRBase V.2: an updated resource for multispecies lncRNAs and ClinicLSNP hosting genetic variants in lncRNAs for cancer patients. RNA biology. 2020 Oct 28:1-16.
49. Hashim A, Rizzo F, Marchese G, et al. RNA sequencing identifies specific PIWI-interacting small non-coding RNA expression patterns in breast cancer. Oncotarget. 2014 Oct 30;5(20):9901-10.
50. Ponten F, Jirstrom K, Uhlen M. The Human Protein Atlas--a tool for pathology. The Journal of pathology. 2008 Dec;216(4):387-93.
51. Cerami EG, Gross BE, Demir E, et al. Pathway Commons, a web resource for biological pathway data. Nucleic acids research. 2011 Jan;39(Database issue):D685-90.
52. Pujana MA, Nadal M, Gratacos M, et al. Additional complexity on human chromosome 15q: identification of a set of newly recognized duplicons (LCR15) on 15q11-q13, 15q24, and 15q26. Genome research. 2001 Jan;11(1):98-111.
53. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nature reviews Genetics. 2007 Apr;8(4):272-85.

54. Ruby JG, Jan C, Player C, et al. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell. 2006 Dec 15;127(6):1193-207.
55. Lee EJ, Banerjee S, Zhou H, et al. Identification of piRNAs in the central nervous system. Rna. 2011 Jun;17(6):1090-9.
56. Nelson CE, Hersh BM, Carroll SB. The regulatory content of intergenic DNA shapes genome architecture. Genome biology. 2004;5(4):R25.
57. Pink RC, Wicks K, Caley DP, et al. Pseudogenes: pseudo-functional or key regulators in health and disease? Rna. 2011 May;17(5):792-8.
58. Hirano T, Iwasaki YW, Lin ZY, et al. Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate. Rna. 2014 Aug;20(8):1223-37.
59. Flicek P, Amode MR, Barrell D, et al. Ensembl 2012. Nucleic acids research. 2012 Jan;40(Database issue):D84-90.
60. Pantano L, Jodar M, Bak M, et al. The small RNA content of human sperm reveals pseudogene-derived piRNAs complementary to protein-coding genes. Rna. 2015 Jun;21(6):1085-95.
61. Halbach R, Miesen P, Joosten J, et al. A satellite repeat-derived piRNA controls embryonic development of Aedes. Nature. 2020 Apr;580(7802):274-277.
62. Vagin VV, Sigova A, Li C, et al. A distinct small RNA pathway silences selfish genetic elements in the germline. Science. 2006 Jul 21;313(5785):320-4.
63. Vandewege MW, Platt RN, 2nd, Ray DA, et al. Transposable Element Targeting by piRNAs in Laurasiatherians with Distinct Transposable Element Histories. Genome biology and evolution. 2016 May 9;8(5):1327-37.
64. Petersen M, Armisen D, Gibbs RA, et al. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC evolutionary biology. 2019 Jan 9;19(1):11.
65. Jung I, Park JC, Kim S. piClust: a density based piRNA clustering algorithm. Computational biology and chemistry. 2014 Jun;50:60-7.
66. Han BW, Zamore PD. piRNAs. Current biology : CB. 2014 Aug 18;24(16):R730-3.
67. Barberan-Soler S, Fontrodona L, Ribo A, et al. Co-option of the piRNA pathway for germline-specific alternative splicing of C. elegans TOR. Cell reports. 2014 Sep 25;8(6):1609-1616.
68. Saito K, Ishizu H, Komai M, et al. Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in Drosophila. Genes & development. 2010 Nov 15;24(22):2493-8.
69. Lau NC, Robine N, Martin R, et al. Abundant primary piRNAs, endo-siRNAs, and microRNAs in a Drosophila ovary cell line. Genome research. 2009 Oct;19(10):1776-85.
70. Ray R, Pandey P. piRNA analysis framework from small RNA-Seq data by a novel cluster prediction tool - PILFER. Genomics. 2018 Nov;110(6):355-365.
71. Jehn J, Gebert D, Pipilescu F, et al. PIWI genes and piRNAs are ubiquitously expressed in mollusks and show patterns of lineage-specific adaptation. Communications biology. 2018;1:137.
72. Das PP, Bagijn MP, Goldstein LD, et al. Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the Caenorhabditis elegans germline. Molecular cell. 2008 Jul 11;31(1):79-90.
73. Barckmann B, El-Barouk M, Pelisson A, et al. The somatic piRNA pathway controls germline transposition over generations. Nucleic acids research. 2018 Oct 12;46(18):9524-9536.
74. Nandi S, Chandramohan D, Fioriti L, et al. Roles for small noncoding RNAs in silencing of retrotransposons in the mammalian brain. Proceedings of the National Academy of Sciences of the United States of America. 2016 Nov 8;113(45):12697-12702.

75. Wang H, Ma Z, Niu K, et al. Antagonistic roles of Nibbler and Hen1 in modulating piRNA 3' ends in Drosophila. Development. 2016 Feb 1;143(3):530-9.
76. Lim SL, Qu ZP, Kortschak RD, et al. Correction: HENMT1 and piRNA Stability Are Required for Adult Male Germ Cell Transposon Repression and to Define the Spermatogenic Program in the Mouse. PLoS genetics. 2015 Dec;11(12):e1005782.
77. Kim KW. PIWI Proteins and piRNAs in the Nervous System. Molecules and cells. 2019 Dec 31;42(12):828-835.
78. Kamaliyan Z, Pouriamanesh S, Soosanabadi M, et al. Investigation of piwi-interacting RNA pathway genes role in idiopathic non-obstructive azoospermia. Scientific reports. 2018 Jan 9;8(1):142.
79. Zhong F, Zhou N, Wu K, et al. A SnoRNA-derived piRNA interacts with human interleukin-4 pre-mRNA and induces its decay in nuclear exosomes. Nucleic acids research. 2015 Dec 2;43(21):10474-91.
80. Plestilova L, Neidhart M, Russo G, et al. Expression and Regulation of PIWIL-Proteins and PIWI-Interacting RNAs in Rheumatoid Arthritis. PloS one. 2016;11(11):e0166920.
81. Juliano C, Wang J, Lin H. Uniting germline and stem cells: the function of Piwi proteins and the piRNA pathway in diverse organisms. Annual review of genetics. 2011;45:447-69.
82. Anders S, Huber W. Differential expression analysis for sequence count data. Genome biology. 2010;11(10):R106.
83. Reeves ME, Firek M, Jliedi A, et al. Identification and characterization of RASSF1C piRNA target genes in lung cancer cells. Oncotarget. 2017 May 23;8(21):34268-34282.
84. Mai D, Ding P, Tan L, et al. PIWI-interacting RNA-54265 is oncogenic and a potential therapeutic target in colorectal adenocarcinoma. Theranostics. 2018;8(19):5213-5230.
85. Busch J, Ralla B, Jung M, et al. Piwi-interacting RNAs as novel prognostic markers in clear cell renal cell carcinomas. Journal of experimental & clinical cancer research : CR. 2015 Jun 14;34(1):61.
86. Zuo Y, Liang Y, Zhang J, et al. Transcriptome Analysis Identifies Piwi-Interacting RNAs as Prognostic Markers for Recurrence of Prostate Cancer. Frontiers in genetics. 2019;10:1018.
87. Weng W, Liu N, Toiyama Y, et al. Novel evidence for a PIWI-interacting RNA (piRNA) as an oncogenic mediator of disease progression, and a potential prognostic biomarker in colorectal cancer. Molecular cancer. 2018 Jan 30;17(1):16.
88. Wang C, Lin H. Roles of piRNAs in transposon and pseudogene regulation of germline mRNAs and lncRNAs. Genome biology. 2021 Jan 8;22(1):27.
89. Krishnan P, Ghosh S, Wang B, et al. Profiling of Small Nucleolar RNAs by Next Generation Sequencing: Potential New Players for Breast Cancer Prognosis. PloS one. 2016;11(9):e0162622.
90. Pertea M, Kim D, Pertea GM, et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature protocols. 2016 Sep;11(9):1650-67.

CHAPTER 6| Investigating the role of piRNAs in stem cell derivatives

Abstract:
Exploring the role of PIWI-interacting RNAs in stem cell derivatives has become a focal point in current studies, shedding light on the intricate mechanisms that underlie cellular differentiation and tissue development. Recent investigations have uncovered a compelling association between piRNAs and the regulation of gene expression patterns during the differentiation process. In this study, we investigated the gene regulatory networks underlying the differentiation of iPSCs into endothelial cells, focusing on the dynamic interplay between key regulatory genes and piRNAs. Our findings reveal a fine tuned piRNA expression dynamics that modulates stemness associated genes facilitating the transition of the iPSCs to its differentiated state corresponding to endothelial lineage. Interestingly, we observed a set of tumor suppressor genes to be key piRNA targets which aid in such differentiation to the endothelial cell state. These insights shed light on the piRNA regulated intricate regulatory mechanisms orchestrating the differentiation process and highlight the critical role of piRNAs in maintaining cellular homeostasis during iPSC differentiation.

6.1. Introduction

PIWI-interacting RNAs (piRNAs) are a class of small non-coding RNAs that are known to play a crucial role in maintaining the stemness of cells, particularly in the context of stem cell biology. Stemness refers to the unique ability of stem cells to self-renew and differentiate into various cell types, making them essential for tissue homeostasis and regeneration. PIWI proteins were initially identified as key components in the self-renewal of stem cells within the Drosophila ovaries [1] and have since been observed across various animal germlines. Their presence extends to diverse potent stem cell types, encompassing hematopoietic and embryonic stem cells [2,3]. Besides, it has been noted that piRNAs play a crucial role in maintaining the self-renewal property of mammalian germ stem cells [4]. These observations imply a potentially conserved role for PIWI proteins likely in association with piRNAs in adult stem cell biology and regeneration. Despite significant advancements in understanding piRNA biogenesis and their role in transposon silencing, their specific involvement in the maintenance and differentiation of stem cells has remained enigmatic.

The intricate network of piRNA-mediated processes serves as a safeguard mechanism, ensuring the preservation of stemness features and the sustained regenerative potential of

stem cells. In a recent investigation, PIWI proteins were observed to undergo upregulation early in the reprogramming process, along with nanos and tdrd7, in induced mouse pluripotent stem (iPS) cells [5]. This observation suggests that the increased expression of piRNA pathway-related genes is not a delayed or indirect consequence of reprogramming [6]. While these findings imply potential roles for PIWI proteins and piRNAs in pluripotent stem cell reprogramming, there is currently no direct evidence regarding the presence and involvement of piRNAs in the differentiation process of induced pluripotent stem cells (iPSCs). To address this gap, we conducted an analysis on both small and long RNA expression data of iPSC and iPSC-derived endothelial cells (iPSC_EC) to gather knowledge on the transition landscape from iPSC to iPSC_EC. The reason behind such selection is because endothelial cells (ECs) are one of the major cell types derived from iPSCs and have broad relevance in both basic research and clinical applications. Further, ECs play a crucial role in vascular biology and are involved in processes such as angiogenesis, vasculogenesis, and vascular homeostasis. Overall, the selection of ECs as the focus of this study provides a platform to investigate fundamental aspects of vascular biology, disease mechanisms, and regenerative medicine using iPSC technology. Our investigation revealed the presence differential piRNA signatures between iPSC and iPSC_EC, suggesting a potential involvement of specific piRNAs in initiating and maintaining the differentiated state of endothelial lineage. Understanding the precise molecular mechanisms through which piRNAs orchestrate these processes holds significant promise for advancing our knowledge of stem cell biology and may have implications for therapeutic strategies in regenerative medicine.

6.2. Methods

6.2.1. Transcriptomic and small RNA data analysis:

The sequencing dataset utilized in this study was obtained from the Gene Expression Omnibus repository (www.ncbi.nlm.nih.gov/geo) (Table 1). It is important to note that reprogramming method used to generate iPSCs often have a long drawn effect not only on the iPSC transcriptome but also on the transcriptome of its differentiated counterpart. Hence in order to avoid such disparities we have selected both the iPSC and the iPSC_EC datasets where the iPSCs have been generated via sendai virus-mediated reprogramming method. The experimental grouping is also provided in **Table 1**.

	Trancriptomic Sequencing Data		**Small RNA Sequencing Data**	
	Accession ID	**Sample IDs**	**Accession ID**	**Sample IDs and #of samples**
iPSC (Control)	GSE107181	GSM2862282 GSM2862283 GSM2862284 GSM2862285	GSE161530	GSM4909394 GSM4909395 GSM4909396
iPSC_EC (Test)	GSE195559	GSM5840533 GSM5840534 GSM5840535	In house Generated data	2

Table 1: Input dataset

The quality assessment of the raw reads for long RNAseq data analysis was conducted using FastQC [7]. Reads were initially aligned to the reference genome using Hisat2 [8]. Samtools [9] was employed to assemble the necessary files for differential gene expression analysis, which was executed using Cuffdiff [10]. Differentially regulated genes were determined using p-value ≤0.05 and fold-change cut-off ≥2.0.

For the small RNAseq data, after initial quality assessment, adapters were eliminated from the raw sequencing reads using Cutadapt, retaining reads with lengths ranging from 18 to 40 nucleotides after removal. BLAST [11] was used to align the reads with annotated piRNAs obtained from piRNAQuest V.2 [12]. Differential analysis was conducted utilizing DESeq2 [13], followed by identification of differentially regulated piRNAs with a fold-change threshold of ≥2.0.

6.2.2. Genomic localization of the piRNAs:

The coordinate files corresponding to various genomic locations (5' UTR, 3' UTR, CDS, Intron) and repeats, including information related to different repeat families, were acquired from the UCSC Table Browser [14]. Subsequently, in-house Perl scripts were employed to ascertain the overlap between piRNAs and the aforementioned genomic elements.

6.2.3. Prediction of piRNA Targets:

To predict potential targets of differentially expressed (DE) piRNAs, we employed our targeting prediction pipelines, focusing on extensive sequence complementarity between DE-piRNAs and DE-mRNAs. 3'UTR sequences of the DEGs are obtained from UCSC table browser. We considered the negatively correalated pairs i.e., upregulated piRNAs and downregulated mRNAs and vice versa. Utilizing the miRanda algorithm [15] with alignment score (sc≥170) and energy (en≤ -20.0 KCal/mol) parameters as proposed by Hashim et al. [16], we identified putative target interactions between piRNAs and mRNAs.

6.2.4. Functional analysis of predicted targets:

In the functional analysis of predicted targets, we utilized Ingenuity Pathway Analysis (IPA) software to explore the biological significance and potential pathways associated with the identified target genes of differentially expressed piRNAs. IPA employs a comprehensive database of curated biological interactions and functional annotations to decipher the molecular pathways, biological processes, and cellular functions enriched within the target gene set.

6.3. Results and Discussion

6.3.1. Unveiling the piRNAome and its expression dynamics in due course of differentiation from iPSC to iPSC-EC state:

The differential profile of piRNAs revealed 235 piRNAs that were DE in iPSC_EC compared to iPSC. Among these, 127 piRNAs were up-regulated and 108 were down-regulated in iPSC_EC compared to iPSC. Hierarchical clustering **[Figure 1]** shows clear separate clustering of iPSC and iPSC_EC samples with respect to the differential piRNA profile between them.

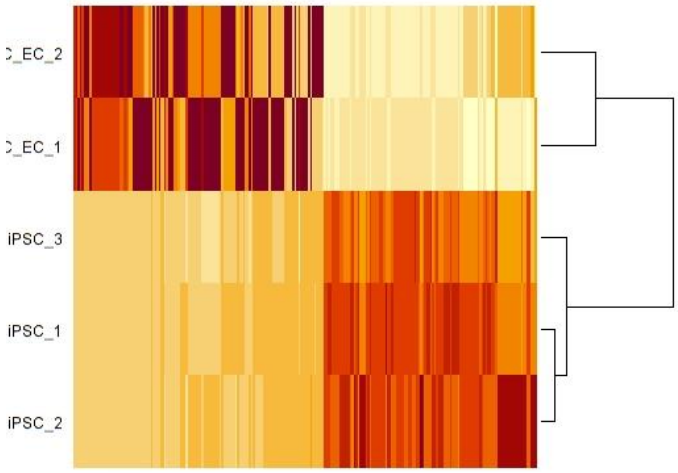


Figure 1: Hierarchical clustering of DE piRNAs

This motivated us to investigate further into their genomic localization along with their overlap with different repetitive regions of the genome which might provide us with a clue regarding their functionalities.

6.3.2. Genomic localization and overlap with repetitive region of the DE piRNAs:
The DE piRNAs were mapped to different genomic regions, including 5' UTR, 3' UTR, coding sequences (CDS), introns, and intergenic regions **[Figure 2]**.

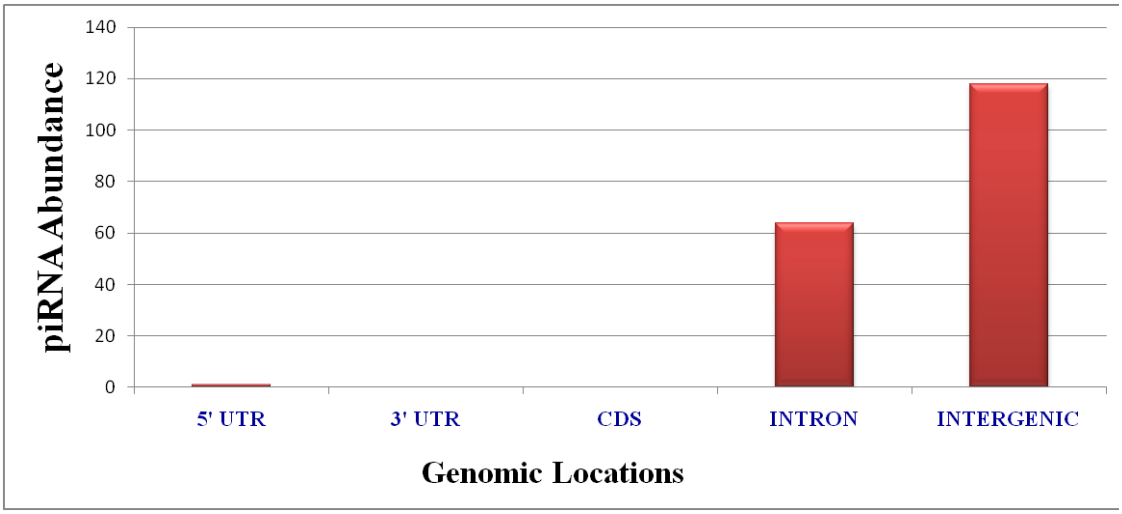


Figure 2 : Localization of DE piRNAs for iPSC and iPSC_EC within genome

The observation indicates a non-uniform distribution of piRNAs across various genomic locations, consistent with findings by Gan et al. [17]. Moreover, a significant portion of piRNAs overlapped with intronic and intergenic regions of the genome, confirming their non-coding nature.

piRNAs, primarily originating from repetitive regions of the genome, play a pivotal role in genome protection by targeting transposable elements rich in repeats [18]. Notably, our observation reveals a significant proportion of piRNAs overlapping with repeats derived from tRNAs and rRNAs **[Figure 3]**. Recent studies have identified RNA fragments as novel sources of piRNAs in various biological contexts, including mouse gametes, zygotes [19], and human somatic cells [20]. In a recent study, there is also reported a significant portion of identified piRNAs are mapped with tRNAs and rRNAs and might have role in cardiac differentiation from pluripotent stem cells [21]. The precise functions of piRNAs derived from RNA families remain elusive; however, they may be involved in epigenetic inheritance, retrotransposon silencing, and other genetic element regulation, as well as post-transcriptional mRNA regulation.

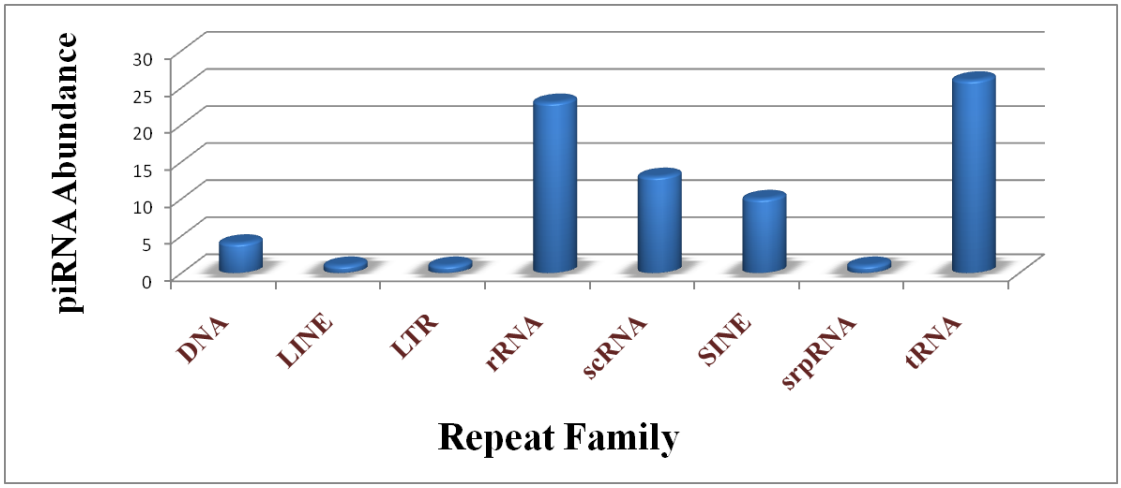


Figure 3 : Abundance of piRNAs across different Repeat families

6.3.3. Deciphering Transcriptome Dynamics: Insights into piRNA-Mediated Gene Regulation contributing to iPSC differentiation towards endothelial lineage

Transcriptomic analysis of iPSCs and iPSC_EC, revealed 6966 differentially expressed genes (3691 upregulated and 3275 downregulated) in iPSC_EC compared to iPSC. We conducted target prediction analysis to identify piRNA targets within this gene set. Specifically, we utilized miRanda to predict target interactions, focusing on negatively correlated pairs for target prediction. piRNA targeting rules vary subtly among species. However, proper piRNA targeting rules have not yet been established for most species. In the absence of proper piRNA specific target tools, miRanda [22], a popular tool for miRNA target prediction is also widely being used for piRNA target prediction [16,23-27] considering the similarity in both piRNA and miRNA seed region binding. **Table 2** presents the output of the target prediction analysis, showcasing the identified interactions between piRNAs and their predicted target genes.

piRNA-mRNA Target pairs in iPSC_EC as compared to iPSC	No. of Target Pairs	No. of Targeting piRNAs	No. of Targeting Genes
Up piRNAs - Down Genes	1012	107	235
Down piRNAs – Up Genes	861	93	225

Table 2: piRNA –Target genes

Detailed functional analysis of the targeted genes revealed certain important observations which are as follows:

piRNA modulating differentiation process: A set of the target genes constitutes essential components of endothelial cell differentiation pathways (**Table 3**). Activation of these pathways orchestrates the differentiation of iPSCs into endothelial cells, a process crucial for vascular development and angiogenesis. The precise regulation of gene expression within these pathways is essential for driving iPSCs towards the endothelial cell lineage. By modulating the expression of target genes involved in endothelial differentiation

pathways, piRNAs seems to be facilitating the transition of iPSCs along the path of differentiation into functional endothelial cells.

piRNAs	Target Genes	Gene Status in iPSC_EC	Overlapping Repeat Region within the Gene	Target Gene regulated Pathway	Significance
hsa_piRNA_21067 hsa_piRNA_35356	CAV1	Up	LINE	Vascular Endothelial Growth Factor (VEGF) Pathway	VEGF signaling plays a central role in endothelial cell differentiation by promoting cell proliferation, migration, and tube formation. Activation of VEGF receptors leads to downstream signaling events that induce endothelial cell specification and angiogenic sprouting.
hsa_piRNA_25087	FLT1		DNA		
hsa_piRNA_49281 hsa_piRNA_49283 hsa_piRNA_49284 hsa_piRNA_45457 hsa_piRNA_49282	PIK3CD	Dow	SINE		
hsa_piRNA_45831	SOX6	Up	LINE	Wnt/β-catenin Signaling Pathway	Wnt signaling pathways play diverse roles in vascular development, including endothelial cell differentiation, angiogenic sprouting, and vascular patterning. Activation of canonical Wnt/β-catenin signaling promotes endothelial cell specification and angiogenic behavior.
hsa_piRNA_28288	ADCY5	Down	-	Hedgehog Signaling Pathway	Hedgehog signaling pathways contribute to vascular development and endothelial cell differentiation by regulating endothelial cell fate determination, proliferation, and angiogenic behavior.

Table 3: piRNAs targeting genes related to pathways essential for differentiation into endothelial lineage

Studies have shown that CAV1 is involved in regulating endothelial cell proliferation, migration, and tube formation, which are essential processes during angiogenesis. Additionally, CAV1 has been implicated in modulating signaling pathways such as VEGF and TGF-β, which are key regulators of endothelial cell differentiation and angiogenesis [28]. FLT1 gene is a receptor for vascular endothelial growth factor (VEGF) and plays a crucial role in angiogenesis [29]. On the other hand, it has been found that, if the downregulated gene, PIK3CD associated with VEGF pathway, got dysregulated, it would be associated with aberrant angiogenesis and vascular dysfunction [30]. SOX6 is a member of the SOX (SRY-box) family of transcription factors involved in various developmental processes. While the specific role of SOX6 in endothelial differentiation

is not well-defined, studies have suggested its potential involvement in angiogenesis and vascular development. SOX6 has been implicated in regulating the expression of genes involved in endothelial cell function and vascular morphogenesis, although its direct role in endothelial differentiation remains to be elucidated. Studies showed that dysregulated cAMP signaling mediated by ADCY5 could potentially impact endothelial cell behavior and angiogenesis negatively.

Another set of piRNA target genes upregulated in iPSC_EC are associated with crucial signaling pathways (**Table 4**), notably the Endothelin-1 Signaling pathway and Apelin Endothelial Signaling Pathway [31,32]. These pathways play a pivotal role in regulating blood vessel tone, a fundamental aspect of vascular physiology and are involved in a variety of physiological processes including vasoconstriction and dilation, blood pressure regulation, cardiac contractility enhancement, angiogenesis, and energy metabolism modulation. These pathways are essential for proper functioning of the endothelial cells in maintaining vascular homeostasis.

piRNA	Target Genes	Gene Status in iPSC_EC	Overlapping Repeat Region with the target Gene	Target Gene regulated Pathway	Significance
hsa_piRNA_773	ADCY7	Upregulated	SINE	Endothelin-1 Signaling pathway	Endothelin-1 is a potent endogenous vasoconstrictor secreted by endothelial cells.
hsa_piRNA_27535 hsa_piRNA_27536 hsa_piRNA_27538 hsa_piRNA_27539 hsa_piRNA_42079 hsa_piRNA_48121 hsa_piRNA_48122 hsa_piRNA_48123	MEF2C		LINE	Apelin Endothelial Signaling Pathway	Apelin signaling regulates endothelial formation.

Table 4: piRNAs targeting genes and corresponding pathways essential to maintain the endothelial cell state

piRNAs targeting tumor suppressor genes: It has been reported previously that certain tumor suppressors like p53 plays a significant role in cell differentiation [33]. Further it has been seen that down regulation of p53 leads to an increased efficiency of reprogramming iPSCs [34]. Other reports have also shown that the retinoblastoma tumor suppressor controls cell proliferation and differentiation. Loss of this function in stem or progenitor cells can serve as a pivotal event in cancer initiation modifying the path of differentiation [35]. But there has been no report regarding the role of tumor suppressors influencing differentiation to a particular lineage specifically to endothelial lineage. Our analysis come up with some interesting set of tumor suppressors as piRNA targets which are significantly expressed along this pathway of differentiation to endothelial lineage. **Table 5** provides the list of the piRNAs along with their target tumor suppressor genes

which remains down in iPSCs but gradually turns up as the cell transits to the differentiated state into endothelial lineage. Subsequent future studies can be carried out towards looking further into the expression dynamics of these piRNAs and their tumor suppressor targets within the intermittent states lying between iPSC and iPSC-EC which will strengthen their role in differentiation along this lineage.

piRNA	Target Tumor Suppressor Genes	Overlapping Repeat Region within the Gene	Cancer Type
hsa_piRNA_25921	AR	SINE	Retinoblastoma
hsa_piRNA_49859	IKZF1	LINE	Leukemia
hsa_piRNA_25087, hsa_piRNA_49859	GPR161	SINE	Medulloblastoma
hsa_piRNA_48119, hsa_piRNA_48121	ZBTB16	LINE	Prostate cancer, uterine cancer
hsa_piRNA_25087, hsa_piRNA_41005	INPP4B	SINE	Ovarian cancer, cervical cancer
hsa_piRNA_25087	PRKAR1A	SINE	Lung adenocarcinoma
hsa_piRNA_24546, hsa_piRNA_24547	TNFAIP3	SINE	Primary mediastinal B-cell lymphoma
hsa_piRNA_27012, hsa_piRNA_773	PCDH9	LINE	Brain glioma, Hepatocellular carcinoma
hsa_piRNA_47867, hsa_piRNA_47868	SUFU	-	Clear cell renal cell carcinoma
hsa_piRNA_46692	SETD7	-	Stomach cancer, breast cancer

Table 5: piRNAs targeting Tumor Suppressor genes

Certain piRNA target genes that play a role in maintaining the stem-like properties of cells (NTF4, TFCP2L1, STAM, KIF14) are downregulated in iPSC_EC which is needed to facilitate the differentiation process (**Table 6**). All these once again revealed the importance of piRNA expression dynamics towards regulating stem cell fate and differentiating them along the path of endothelial lineage.

piRNA	Target Genes	Gene Status in iPSC_EC	Overlapping Repeat within the Gene	Target gene regulated Pathway	Significance
hsa_piRNA_30968 hsa_piRNA_30967	NTF4	Downregulated	SINE	Human Embryonic Stem Cell Pluripotency	Properties of pluripotency and self-renewal
hsa_piRNA_10261 hsa_piRNA_29657	TFCP2L1		LINE		
hsa_piRNA_33753 hsa_piRNA_40198	STAM		SINE	Signaling by MET	Regulator of stem cell renewal in early embryonic development
hsa_piRNA_43432 hsa_piRNA_43433	KIF14		SINE	RHO GTPase cycle	Embryonic Stem cell regulation

Table 6: piRNAs downregulating stemness related genes to facilitate differentiation of the cell

A major portion of these piRNAs playing various role in iPSC differentiation to endothelial cells showed maximum overlap with RNA Repeat families **[Figure 4]**. This once again reiterates the importance of RNA derived piRNAs as mentioned previously.

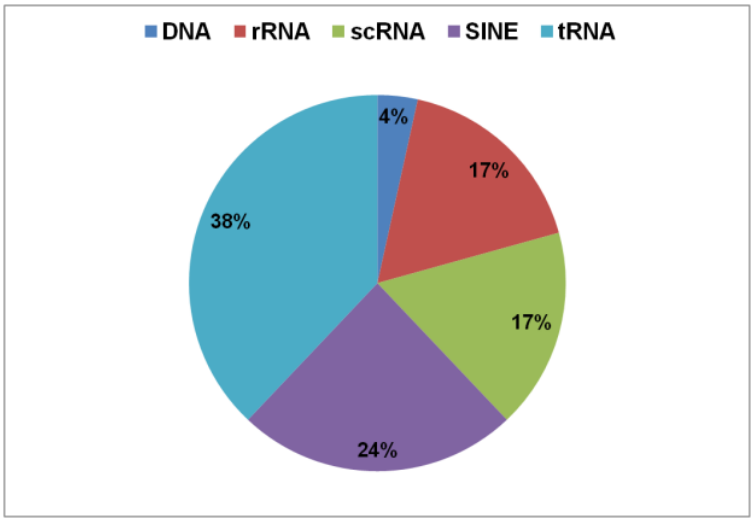


Figure 4: Distribution of the targeting piRNAs across Different Repeat Families

In summary, our findings indicate that the differentially expressed piRNAs identified in this study may exert regulatory roles in fine-tuning the expression of genes crucial for maintaining differentiated state of endothelial cells. Through their interactions with target genes, these piRNAs likely contribute to the intricate regulatory networks governing endothelial cell identity and function.

6.4. Conclusion

While the existence and diverse functions of piRNAs in pluripotent stem cells have been extensively studied, their role in iPSC-derived cells remains relatively unexplored. In this study, we have investigated the piRNA expression dynamics towards orchestrating the differentiation of iPSCs specifically into ECs as they play a crucial role in vascular biology and are involved in processes such as angiogenesis, vasculogenesis, and vascular homeostasis which are highly linked to regenerative therapy based applicative studies. Our analysis revealed various interesting aspects including the modulatory role of piRNAs on tumor suppressor genes to initiate and continue the differentiation process. Further, a large proportion of the piRNAs modulating the differentiation process have overlapping RNA repeats which reveals the importance of RNA fragments in differentiation process. Overall, it is just the tip of the iceberg. This work warrants further experimental validation to strengthen our conclusion. It also throws up open questions as to whether such piRNA targeting pattern is very specific for differentiation to a particular lineage.

References

1. Cox DN, Chao A, Baker J, et al. A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. Genes & development. 1998 Dec 1;12(23):3715-27.
2. Cheng EC, Kang D, Wang Z, et al. PIWI proteins are dispensable for mouse somatic development and reprogramming of fibroblasts into pluripotent stem cells. PloS one. 2014;9(9):e97821.
3. Sharma AK, Nelson MC, Brandt JE, et al. Human CD34(+) stem cells express the hiwi gene, a human homologue of the Drosophila gene piwi. Blood. 2001 Jan 15;97(2):426-34.
4. Bamezai S, Rawat VP, Buske C. Concise review: The Piwi-piRNA axis: pivotal beyond transposon silencing. Stem cells. 2012 Dec;30(12):2603-11.
5. Zhu Y, Fan C, Zhao B. Differential expression of piRNAs in reprogrammed pluripotent stem cells from mouse embryonic fibroblasts. IUBMB life. 2019 Dec;71(12):1906-1915.
6. Samavarchi-Tehrani P, Golipour A, David L, et al. Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. Cell stem cell. 2010 Jul 2;7(1):64-77.
7. S. A. FastQC: a quality control tool for high throughput sequence data. http://wwwbioinformaticsbabrahamacuk/projects/fastqc. 2010.
8. Pertea M, Kim D, Pertea GM, et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature protocols. 2016 Sep;11(9):1650-67.
9. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021 Feb 16;10(2).
10. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology. 2010 May;28(5):511-5.
11. Chen Y, Ye W, Zhang Y, et al. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic acids research. 2015 Sep 18;43(16):7762-8.
12. Ghosh B, Sarkar A, Mondal S, et al. piRNAQuest V.2: an updated resource for searching through the piRNAome of multiple species. RNA biology. 2022;19(1):12-25.
13. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550.
14. Haeussler M, Zweig AS, Tyner C, et al. The UCSC Genome Browser database: 2019 update. Nucleic acids research. 2019 Jan 8;47(D1):D853-D858.
15. John B, Enright AJ, Aravin A, et al. Human MicroRNA targets. PLoS biology. 2004 Nov;2(11):e363.
16. Hashim A, Rizzo F, Marchese G, et al. RNA sequencing identifies specific PIWI-interacting small non-coding RNA expression patterns in breast cancer. Oncotarget. 2014 Oct 30;5(20):9901-10.
17. Gan H, Lin X, Zhang Z, et al. piRNA profiling during specific stages of mouse spermatogenesis. Rna. 2011 Jul;17(7):1191-203.
18. Thomson T, Lin H. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. Annual review of cell and developmental biology. 2009;25:355-76.
19. Garcia-Lopez J, Hourcade Jde D, Alonso L, et al. Global characterization and target identification of piRNAs and endo-siRNAs in mouse gametes and zygotes. Biochimica et biophysica acta. 2014 Jun;1839(6):463-75.
20. Keam SP, Young PE, McCorkindale AL, et al. The human Piwi protein Hiwi2 associates with tRNA-derived piRNAs in somatic cells. Nucleic acids research. 2014 Aug;42(14):8984-95.

21. La Greca A, Scarafia MA, Hernandez Canas MC, et al. PIWI-interacting RNAs are differentially expressed during cardiac differentiation of human pluripotent stem cells. PloS one. 2020;15(5):e0232715.
22. Enright AJ, John B, Gaul U, et al. MicroRNA targets in Drosophila. Genome Biology. 2003 2003/12/12;5(1):R1.
23. Zuo Y, Liang Y, Zhang J, et al. Transcriptome Analysis Identifies Piwi-Interacting RNAs as Prognostic Markers for Recurrence of Prostate Cancer. Frontiers in genetics. 2019;10:1018.
24. Liu Y, Zhang J, Li A, et al. Prediction of cancer-associated piRNA–mRNA and piRNA–lncRNA interactions by integrated analysis of expression and sequence data. Tsinghua Science and Technology. 2018;23(2):115-125.
25. Weng W, Liu N, Toiyama Y, et al. Novel evidence for a PIWI-interacting RNA (piRNA) as an oncogenic mediator of disease progression, and a potential prognostic biomarker in colorectal cancer. Molecular cancer. 2018 Jan 30;17(1):16.
26. Roy J, Sarkar A, Parida S, et al. Small RNA sequencing revealed dysregulated piRNAs in Alzheimer's disease and their probable role in pathogenesis. Molecular bioSystems. 2017 Feb 28;13(3):565-576.
27. Singh G, Roy J, Rout P, et al. Genome-wide profiling of the PIWI-interacting RNA-mRNA regulatory networks in epithelial ovarian cancers. PloS one. 2018;13(1):e0190485.
28. de Almeida CJG. Caveolin-1 and Caveolin-2 Can Be Antagonistic Partners in Inflammation and Beyond. Frontiers in immunology. 2017;8:1530.
29. Shibuya M. Vascular Endothelial Growth Factor (VEGF) and Its Receptor (VEGFR) Signaling in Angiogenesis: A Crucial Target for Anti- and Pro-Angiogenic Therapies. Genes & cancer. 2011 Dec;2(12):1097-105.
30. Zhao Y, Qian Y, Sun Z, et al. Role of PI3K in the Progression and Regression of Atherosclerosis. Frontiers in pharmacology. 2021;12:632378.
31. Tocci P, Blandino G, Bagnato A. YAP and endothelin-1 signaling: an emerging alliance in cancer. Journal of experimental & clinical cancer research : CR. 2021 Jan 9;40(1):27.
32. Helker CS, Eberlein J, Wilhelm K, et al. Apelin signaling drives vascular endothelial cells toward a pro-angiogenic state. eLife. 2020 Sep 21;9.
33. Jain AK, Barton MC. p53: emerging roles in stem cells, development and beyond. Development. 2018 Apr 13;145(8).
34. Rasmussen MA, Holst B, Tumer Z, et al. Transient p53 suppression increases reprogramming of human fibroblasts without affecting apoptosis and DNA damage. Stem cell reports. 2014 Sep 9;3(3):404-13.
35. Sage J. The retinoblastoma tumor suppressor and stem cell biology. Genes & development. 2012 Jul 1;26(13):1409-20.

CHAPTER 7| General Conclusions and Future Perspectives

The ever-expanding repertoire of functional non-coding RNAs (ncRNAs) has revolutionized our understanding of the non-coding genome's significance in cell biology. The emergence of novel classes of ncRNAs has underscored their pivotal roles as additional layers of regulation governing diverse aspects of gene expression. In parallel, the development of advanced computational techniques has complemented experimental approaches, facilitating the detection and characterization of these elusive transcripts. Reports from the scientific literature have increasingly highlighted the regulatory contributions of ncRNAs across various biological contexts, with particular emphasis on their involvement in orchestrating gene expression dynamics in stem cells. Stem cells, including induced pluripotent stem cells (iPSCs), represent a paradigmatic model system for studying cellular identity and fate determination. The intricate gene regulatory networks governing aberrant cell fates, such as those observed in pluripotent stem cell derivatives, have emerged as focal points of investigation.

A key focus of this thesis has been to explore the role of ncRNAs in orchestrating the process of iPSC differentiation leading to certain cell fates. This study has placed emphasis on two major classes of small ncRNAs: microRNAs (miRNAs) and PIWI-interacting RNAs (piRNAs).

The investigation into the roles of miRNAs in iPSC derived cells across three germ layers—neurons, hepatocytes, and cardiomyocytes yielded intriguing findings. Notably, the clustering of iPSC-derived cells with their respective cancerous counterparts instead of clustering with the corresponding primary cells of the same lineage (based on both miRNA and mRNA expression data) suggests a potential oncogenic contamination within these iPSC derivatives. This observation underscores the regulatory role of miRNAs towards orchestrating certain cellular events which eventually leads to oncogenicity within iPSC derivatives. In the subsequent phase of my thesis, I sought to investigate whether the choice of reprogramming method for iPSC generation influences the observed oncogenic transformation within their differentiated counterparts. Surprisingly, the results revealed a consistent presence of oncogenic contamination in iPSC-derived endothelial cells, irrespective of the reprogramming method employed. This unexpected finding prompted further inquiry into the underlying mechanisms driving this phenomenon. To gain deeper insights, we conducted in-house small RNA sequencing analysis, which corroborated the observed pattern of oncogenic contamination within iPSC derivative. Subsequent analyses identified potential miRNA-mediated alterations implicated in driving this transformation. These findings highlight the critical role of miRNA dysregulation in mediating oncogenic transformation in iPSC-derived endothelial cells.

Based upon the insights obtained from the prior investigation spanning into the three germ layer based analysis and looking back into the involvement of different reprogramming methods for such remnant oncogenic contamination within iPSC derivatives we proceeded to develop a prediction model. This will serve as a quality

control measure to predict the eligibility of these iPSC derivates to be used for regenerative therapy. By leveraging machine learning techniques embedding the features involving miRNA mediated modulation of the transcriptome of iPSC derivatives, the model offers a robust tool for predicting the likelihood of oncogenic transformation within them, thus assessing their eligibility for a safe and reliable regenerative application.

One of the strengths of this study lies in its comprehensive analysis of iPSC derivatives across all germ layers along with consideration of different reprogramming methods allowing for a holistic examination of regulatory mechanisms in diverse cell types. By elucidating the oncogenic signatures associated with miRNA dysregulation in iPSC derivatives, this research provides valuable insights into the molecular underpinnings of cellular differentiation and lineage commitment. However, it is important to acknowledge certain limitations of this work. Firstly, while the clustering of iPSC derivatives with cancer systems suggests oncogenic similarities, further functional validation is required to elucidate the biological significance of these observations. Additionally, the potential constraint is posed by data availability, particularly in accessing comprehensive datasets encompassing diverse iPSC derivatives of different other cell types corresponding to each germ layer. This scarcity of data may hinder the thorough exploration of necessitate cautious interpretation of findings.

Despite these limitations, the identification of oncogenic parallels between iPSC derivatives and cancer cells opens up new avenues for future research. The implementation of the predictive model represents a significant advancement in the field of iPSC research, offering an applicative approach for quality control and risk management. By providing researchers with a means to assess the oncogenic risk associated with iPSC-derived cell populations, the prediction model stands to benefit the broader scientific community engaged in iPSC-based studies and applications. Moving forward, integrated experimental and computational approaches will be instrumental in unraveling the intricacies of oncogenic regulation in iPSC derivatives.

piRNAs are recognized for their significant roles in stem cell self-renewal and various essential processes in developmental and disease biology. To deepen our understanding of piRNA biology, we embarked on updating our previously published database on piRNAs, now known as piRNAQuest V.2. This updated version incorporates novel features associated with fundamental piRNA biology, alongside an extensive investigation into their roles in diverse disease systems. Despite ongoing efforts by researchers to explore different aspects of piRNAs in stem cells, the precise regulatory mechanisms governing their involvement in stem cell differentiation remain elusive. Leveraging the wealth of information available in piRNAQuest V.2, we conducted a comprehensive analysis on role of piRNAs in modulating the differentiation of iPSCs into endothelial cells. Our analysis uncovered several intriguing findings, including the regulatory influence of piRNAs on tumor suppressor genes, which might play a pivotal role in initiating and sustaining the differentiation process. Subsequent future studies can delve deeper into the expression dynamics of these piRNAs and their targets within the transitional states between iPSC and its derivatives. This exploration will enhance our

understanding of their role in guiding differentiation along this lineage. Further, majority of these targeting piRNAs have been found to have overlap with RNA repeat families, which once again reveals the importance of RNA derived fragments.

In summary, our research endeavors have brought to light previously unexplored and underestimated regulatory networks governed by ncRNAs in shaping the landscape of stem cell biology. Through our investigations, we have demonstrated that ncRNAs, once regarded as the "dark matter" of the genome, are indeed key players orchestrating fundamental processes within eukaryotic cellular homeostasis. By elucidating the contributions of small ncRNAs viz.miRNAs and piRNAs in iPSC differentiation, this thesis not only advances our fundamental understanding of stem cell biology but also holds implications for regenerative medicine and disease modeling. The insights gleaned from this study may inform the development of novel therapeutic strategies aimed at manipulating cellular fates for therapeutic purposes. Moreover, this research underscores the importance of integrative approaches that leverage both computational and experimental methodologies to unravel the complexities of gene regulation in stem cells and beyond. Moving forward, our findings pave the way for further exploration into the regulatory roles of these ncRNAs across diverse biological contexts. By continuing to unravel the complexities of ncRNA-mediated regulation, we can gain deeper insights into the fundamental processes governing cellular function and pave the way for the development of innovative therapeutic strategies targeting ncRNAs in various diseases and disorders.

Appendix

PEER-REVIEWED PUBLICATIONS

1. Ghosh, B., Sarkar, A., Mondal, S., Bhattacharya, N., Khatua, S., & Ghosh, Z. (2022). *piRNAQuest V.2: An updated resource for searching through the piRNAome of multiple species.* RNA Biology, 19(1), 12–25. (DOI: 10.1080/15476286.2021.2010960)

2. Ghosh, B., Das, T., Das, G., Chowdhury, N., Bagchi, A., & Ghosh, Z. (2023). *Mapping Drug-gene Interactions to Identify Potential Drug Candidates Targeting Envelope Protein in SARS-CoV-2 Infection.* Current Bioinformatics, 18(9), 760-773.
(DOI: 10.2174/1574893618666230605120640)

WORKS IN PROGRESS

1. Chakrabarty, J., Datta, S**.**, **Ghosh, B**., Parveen, R., Roy, V., Ghosh, Z., & Chaudhuri, S. *Rice ULTRAPETALA1 Regulates Developmental Reprogramming to Promote Resilience to Salinity Stress.*
Manuscript currently under revision.

2. Ghosh, B., Das, T. and Ghosh, Z. *Parental noncoding RNAs in pre and post fertilization events.*
Manuscript currently under preparation.

Curriculum vitae

BYAPTI GHOSH

byapti@jcbose.ac.in ; byaptighosh1111@gmail.com

Department of Biological Sciences
Bose Institute
Unified Academic Campus EN 80
Sector V, Bidhan Nagar
Kolkata - 700091
WB India

Ukil Para, Raiganj
Uttar Dinajpur
West Bengal, India

(+91)9046046251

CURRENT POSITION

Department of Biological Sciences, Bose Institute — Kolkata, India
Senior Research Fellow — 2017-Present
Pursuing Ph.D in Life Science and Biotechnology from Jadavpur University, India

EDUCATION

University of Kalyani — Kalyani, India
M.Sc, Microbiology. First Class, University Rank 1 — 2015

University of North Bengal — Siliguri, India
B.Sc, Microbiology. Second Class, University Rank 23 — 2013

RESEARCH EXPERIENCE

Department of Biological Sciences, Bose Institute — Kolkata, India
Research Fellow Under Dr. Zhumur Ghosh
Regulatory Noncoding RNA mediated alterations and its effects in stem cell derivatives — 2017-Present

ICMR, Summer Trainee — Kolkata, India
Studies on human cytomegalovirus infection as an opportunistic infection in immunocompromised (HIV/AIDS) patients — 2015

University of Kalyani, 3 days workshop — Kalyani, India
Laboratory methods and practices for the evaluation of wastewater — 2015

PROFESSIONAL EXPERIENCE

Raiganj University — Raiganj, India
Guest Lecturer at Department of Microbiology — 2015-2017

GRANTS AND AWARDS

Senior Research Fellowship — 2019-2022
Department of Science & Technology, Government of India

Junior Research Fellowship — 2017-2019
Department of Science & Technology, Government of India

Qualified for INSPIRE Fellowship — 2017
As 1st Rank Holder in M.Sc.

Qualified Graduate Aptitude Test in Engineering (GATE) – Life Sciences — 2016

PARTICIPATION IN WORKSHOPS

National workshop on Bioinformatics: AI in Healthcare — Kolkata, India, January, 2024
Bose Institute

One Day National Workshop on Plant Bioinformatics — Kolkata, India, December, 2023
Bose Institute

PARTICIPATION IN CONFERENCES

Indo-Japan Conference on EPIGENETICS AND HUMAN DISEASE, — 2018

International Symposium on Systems, Synthetic & Chemical Biology, Kolkata — 2017

CONFERENCE PRESENTATION

- Recent Trends in Natural Sciences, 2023, 27-29 November, 2023
  "Oncogenic contamination within stem cell derivatives: Checking their purity for safer regenerative therapy"

- Recent Trends in Natural Sciences, 2022, 26-29 November, 2022
  "Repurposed Drug candidates against COVID-19"

- Cold Spring Harbor Laboratory Conference - Regulatory & Non-Coding RNAs, 17th- 21st May, 2022
  "Investigating the non-coding RNA mediated oncogenicity in induced pluripotent stemcell derivatives"

- 41st Annual International Conference of Indian Association for Cancer Research, 2nd- 5th March, 2022
  "Regulatory noncoding RNA induced roadblocks: En route to differentiation of pluripotent stem cells"

- 4th International Conference on Translational Research: Recent Development and Innovations in Human Health and Agriculture Research, 11th- 13th October, 2018
  "Revealing the oncogenic signatures induced by miRNAs in iPSC-derivatives"

RNA BIOLOGY
2022, VOL. 19, NO. 1, 12–25
https://doi.org/10.1080/15476286.2021.2010960

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS Check for updates

piRNAQuest V.2: an updated resource for searching through the piRNAome of multiple species

Byapti Ghosh[a], Arijita Sarkar[a,b], Sudip Mondal[c], Namrata Bhattacharya[d], Sunirmal Khatua[c], and Zhumur Ghosh[a]

[a]Division of Bioinformatics, Bose Institute, Kolkata, India; [b]Present Affiliation: Department of Orthopaedic Surgery, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA; [c]Department of Computer Science and Engineering, University of Calcutta, Kolkata, India; [d]Department of Computer Science and Engineering, Indraprastha Institute of Information Technology, Delhi, India

ABSTRACT

PIWI interacting RNAs (piRNAs) have emerged as important gene regulators in recent times. Since the release of our first version of piRNAQuest in 2014, lots of novel piRNAs have been annotated in different species other than human, mouse and rat. Such new developments in piRNA research have led us to develop an updated database piRNAQuest V.2. It consists of 92,77,689 piRNA entries for 25 new species of different phylum along with human, mouse and rat. Besides providing primary piRNA features which include their genomic location, with further information on piRNAs overlapping with repeat elements, pseudogenes and syntenic regions, etc., the novel features of this version includes (i) density based cluster prediction, (ii) piRNA expression profile across various healthy and disease systems and (iii) piRNA target prediction. The concept of density-based piRNA cluster identification is robust as it does not consider parametric distribution in its model. The piRNA expression profile for 21 disease systems including cancer have been hosted in addition to 32 tissue specific piRNA expression profile for various species. Further, the piRNA target prediction section includes both predicted and curated piRNA targets within eight disease systems and developmental stages of mouse testis. Further, users can visualize the piRNA-target duplex structure and the ping-pong signature pattern for all the ping-pong piRNA partners in different species. Overall, piRNAQuest V.2 is an updated user-friendly database which will serve as a useful resource to survey, search and retrieve information on piRNAs for multiple species. This freely accessible database is available at http://dibresources.jcbose.ac.in/zhumur/pirnaquest2.

ARTICLE HISTORY
Received 12 June 2021
Revised 27 October 2021
Accepted 22 November 2021

KEYWORDS
PIWI interacting RNAs; piRNA cluster; ping-pong piRNAs; piRNA target; piRNA profile

Introduction

PIWI interacting RNAs (piRNAs) belong to a broad group of endogenous small non-coding RNAs(ncRNAs) [1], which typically ranges in length from 25 to 33 nucleotides (nts). In mammals, these ncRNAs were first reported in mouse testes [2–5]. They act as guide for PIWI proteins, which belongs to Argonaute protein family and exhibit slicer activity [6–9]. Unlike other small ncRNAs, i.e. miRNAs and siRNAs, piRNAs are biogenised from both primary processing pathway as well as the amplifying ping-pong mechanism [10] from single stranded precursor molecules [11] via Dicer independent pathway [12]. The primary piRNAs originate from individual genomic loci that are commonly known as piRNA clusters [10]. In most cases, the germline clusters generate piRNAs from both strands (known as dual-strand clusters), whereas flamenco clusters of *Drosophila* follicle cells and murine pachytene piRNA clusters generate piRNAs from only a single DNA strand (uni-strand clusters) [13]. In the ping-pong cycle, generation of sense secondary piRNAs is initiated by the antisense primary piRNAs which in turn produces secondary antisense piRNAs and the amplifying loop continues [7,10].

Although studies on fish, flies and mammals have shown a conserved association of piRNAs with PIWI proteins [2,10,11], the length variation of piRNAs have been observed with evolving sequencing technologies between different species. In general, piRNAs in mammals can be categorized into two subclasses called pachytene (29–33 nts) and pre-pachytene (26–28 nts) [14], whereas piRNAs in *Caenorhabditis elegans* are named as 21 U-RNA owing to its bias for length of 21 nts. Though the piRNAs are best seen in germ cells, several studies have shown piRNA expression in brain, kidney, lung, liver, stomach, testis and ovary [15–18] as well as in different cancers [19].

To maintain genome integrity in germ cell lineages, highly expressed PIWI proteins in germ and stem cells [9] take part in controlling transposon activity as a defensive mechanism [20]. Studies showed that, mutation in MIWI which is a PIWI homolog in mouse leads to male infertility as well as over expression of retrotransposon transcripts [21]. Similar observation has been reported in case of flies [12]. In association with piRNA forming piRNA-induced silencing complexes (piRISCs), PIWI-piRNA pathway silences transposons via complementary base-pair recognition between piRNA and

CONTACT Zhumur Ghosh zhumur@jcbose.ac.in; ghosh.jhumur@gmail.com Division of Bioinformatics, Bose Institute, P-1/12, C.I.T. Scheme-VII M, Kolkata 700 054, India

Supplemental data for this article can be accessed here

transposon followed by endonucleolytic cleaving of the target [22,23].

Existing databases like piRNABank [24], piRBase [25], piRNAdb [https://www.pirnadb.org], piRTarBase [26] and piRNA Cluster Database [27] provide information on piRNAs for multiple species. Among these, piRBase is a manually curated database which hosts piRNA information on multiple species and some disease systems. 'piRNA cluster database' is a dedicated database for piRNA clusters where the clusters are predicted using proTRAC [28]. piRDisease V1.0 [29] hosts piRNA records for different diseases but is not currently accessible. Despite such extensive work on piRNAs, there still remain several unexplored areas, such as their association with long noncoding RNAs (lncRNAs) or the presence of any genomic elements within their loci which can influence their function. We published the first version of piRNAQuest to probe deep into these lesser explored domains of piRNAome. It hosted piRNA information for three species, viz. human, rat and mouse [30].

Though various computational tools have characterized novel piRNAs [31,32] but their function remains unclear. Hence, it is important to identify potential piRNA targets and disease-related piRNAs. Further, both predicted and validated piRNA targets including mRNAs and lncRNAs are not properly curated in any of the existing databases. Moreover, identifying piRNA clusters which are hotspots of piRNA biogenesis is another big challenge in piRNA research.

In this work, we present piRNAQuest V.2, which is an extended version of piRNAQuest. This new version includes the following additional features: (i) extensive analysis on 25 new species in addition to human, mouse and rat of the previous version, (ii) density-based clustering approach [33] to identify the 'hotspots of piRNA expression', popularly known as 'piRNA clusters'. Since piRNA distribution varies with genomic locations in different species, identifying piRNA clusters based on their density in genome can provide new impetus to get biologically relevant clusters, (iii) tissue specific expression of piRNAs among different species, and (iv) expression of piRNAs in different disease systems with an emphasis on different types of cancers. Emerging evidences suggest that piRNAs have important roles in disease progression and diagnosis [34–39]. Thus, the efficacy and potential mechanism of action of a piRNA in cancer relies on its expression in various tissues and disease systems which correlate with disease progression, (v) piRNA target prediction within both mRNAs and lncRNAs that would further help to identify the key players contributing towards disease development.

In addition to these extensive features, we have updated another section of the database, viz. 'Tools', where users will be able to predict piRNA clusters using customized parameters, check ping-pong pattern overlap in their sequences and predict piRNA targets using miRanda [40].

Overall, piRNAQuestV.2 is a user friendly database for multi-species piRNA survey, search and retrieval. piRNA expression within normal tissues and cancer as well as the information about piRNA targets will serve as a valuable resource for piRNA researchers. The database is freely accessible at http://dibresources.jcbose.ac.in/zhumur/pirnaquest2.

Results

piRNAQuest V.2 (an updated version of piRNAQuest) hosts information on 92,77,689 piRNAs corresponding to 28 species (consisting of 25 new species in addition to human, mouse and rat) which are from different phylum ranging from nematode to chordate (Figure 1). Apart from the coverage of species, this new version has included several additional features which add to the significance of this database as compared to other piRNA database. The set of updated features of this new version compared to the old version has been put up in Table 1. We have also put up feature wise comparison of piRNAQuest V.2 with other piRNA database (Supplementary File S1).

Among 28 species, 9 species (viz. Chinese hamster, Sea hare, Tree Shrew, Brown Bat, Silkworm, Mosquito, *Drosophila virilis, Drosophila erecta* and Starlet sea anemone) has not been annotated yet. Hence, genomic localization and related features could only be provided for the rest of the 19 species. Among rest of the 9 species, we have been able to identify repeat-associated piRNAs for 4 species (viz. Chinese hamster, Tree Shrew, *Drosophila virilis* and *Drosophila erecta*), as their repeat annotations were available from UCSC [41] and this information can be

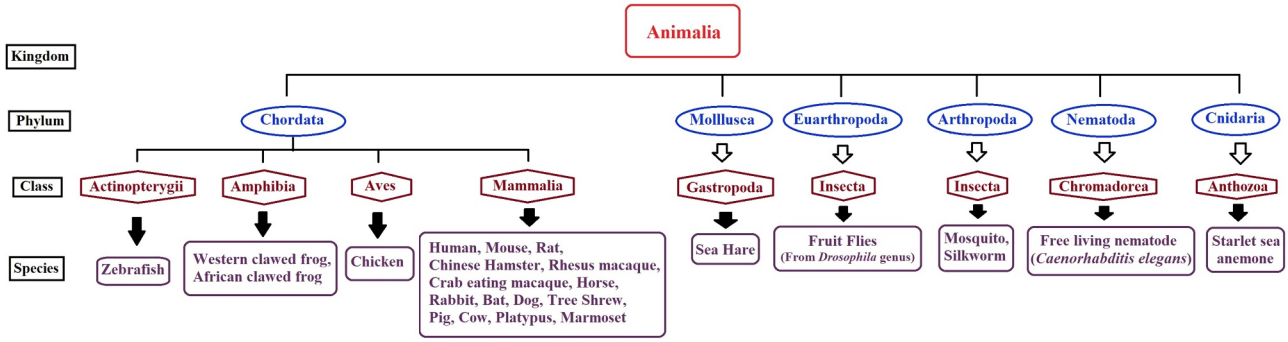


Figure 1. Taxonomical representation of species included in piRNAQuest V.2.
Common names are used for the species with their respective Kingdom, Phylum and Class.

Table 1. Comparison of features between piRNAQuest V.2 and piRNAQuest.

Database content	piRNAQuest	piRNAQuest V.2
Number of species	3	28
piRNA entries	9,98,585	92,77,689
Chromosomal distribution	Yes	Yes
Association with gene	Yes	Yes
Association with pseudogene	Yes	Yes
Association with repeat elements	Yes	Yes
Cluster information	Yes (Lau et al. method), for 3 species	Yes (Density based clustering approach), for 19 species
Association of clusters with genomic regions	Yes	Yes
Syntenic piRNA clusters	Yes	Yes
Ping-pong piRNAs	Yes	Yes
Ping-pong pattern Visualization	No	Yes
Tissue specific expression	Yes (Tissue type – 6, No. of Samples – 9)	Yes (Noraml Tissue type – 32, No. of Samples – 243)
piRNA disease association	No	Yes (16 types of cancer, 2 neurodegenerative diseases amd 3 other diseases)
Graphical representation of expression	No	Yes (For 32 normal tissue types and 16 types of cancer, 2 neurodegenerative diseases amd 3 other diseases)
Predicted piRNA – mRNA target pairs	No	Yes (For seven types of cancer, asthenozoospermia and mouse testis)
Predicted piRNA targets within lncRNAs	No	Yes (For seven types of cancers, asthenozoospermia and mouse testis)
piRNA target genes (literature curated)	No	Yes (for Human, Mouse and *C. elegans*)
Target prediction tool	No	Yes
Ping-pong overlap prediction tool	No	Yes

visualized in graphical format from the 'Statistics' submenu under 'Help Menu' of the database.

Genomic localization based distribution of piRNAs

piRNAQuest V.2 hosts multispecies piRNA information where there is a remarkable increase in the number of piRNA entries compared to that in the previous version. Among the 28 species, distribution of piRNAs across different chromosomes has been mapped only for 19 species (as mentioned above) (Supplementary Figure SF1 and SF2). Interestingly, chromosome 15 in human contains the maximum number of piRNAs which is similar to our observation reported in the earlier version of piRNAquest [30]. In this connection, it is important to note that chromosome 15 in human has been reported to harbour large number of low copy repeats popularly known as duplicons [42] which facilitate nonhomologous recombination events [42] that leads to genome instability [43]. Presence of maximum number of piRNAs in the same chromosome might be to overcome such adverse situation of genome instability, as piRNAs are known to play significant role towards maintaining genome integrity [12].

Further, chromosome 7 and 1 of mouse and rat respectively harbours the maximum number of piRNAs. Among the newly added species, chromosome IV of *Caenorhabditis elegans* (which has also been reported earlier [44]) and Chromosome 2R of *Drosophila melanogaster* contains maximum number of piRNAs.

Abundance of piRNAs in intergenic regions is mostly predominant as compared to that in intronic region for most of the species (Figure 2(a)). One of the significant functions of these intergenic piRNAs is their involvement in early embryonic development [45]. Further, it has been reported that intergenic regions harbour lncRNA loci [46]. Hence, we have checked for the presence of lncRNA loci overlapping with piRNA clusters which consists of intergenic piRNAs (results shown later under the section 'piRNA clusters overlapping with lncRNAs'). On the contrary, piRNA abundance in the 3' UTR, 5' UTR and CDS region is less, except in zebrafish (having high piRNA abundance in CDS region) and *C. elegans* (having high piRNA abundance in $3^{/}$ UTR and $5^{/}$ UTR regions) (Figure 2(a)).

Further, it has been found that pseudogenes regulate its counter gene stability via small RNA mediated silencing [47]. Recently, it has been reported that pachytene piRNAs from pseudogenes directly regulate its parent genes [48]. This motivated us to check the presence of piRNAs within pseudogenes for 16 species whose pseudogene information is available [49]. We obtained significant overlap between piRNAs and pseudogenes in several species (Figure 2(b)). Interestingly, maximum overlap of piRNAs with pseudogenes has been observed in human. Recently, pseudogene derived piRNAs have been found in mature human sperm cells which indicate their role in regulating expression of their parent gene in male germline cells [50].

Distribution of piRNA within repeat regions

piRNAs have been reported to have originated from the repetitive regions and they silence transposons in insects and mice [10,51,52] regulating global gene expression during embryonic development [52]. The piRNA loci for all the 19 species have been mapped to the genomic locations corresponding to seven major categories of repeat elements, viz. LINE, SINE, Simple repeat, DNA, Low complexity, Satellite and LTR (Figure 2(c)). Vandewege et al. [53] reported a strong piRNA response in mammals like dog and horse. These piRNAs are harboured within the SINE repeat regions which are mostly abundant within these species. In our database, we have also reported the

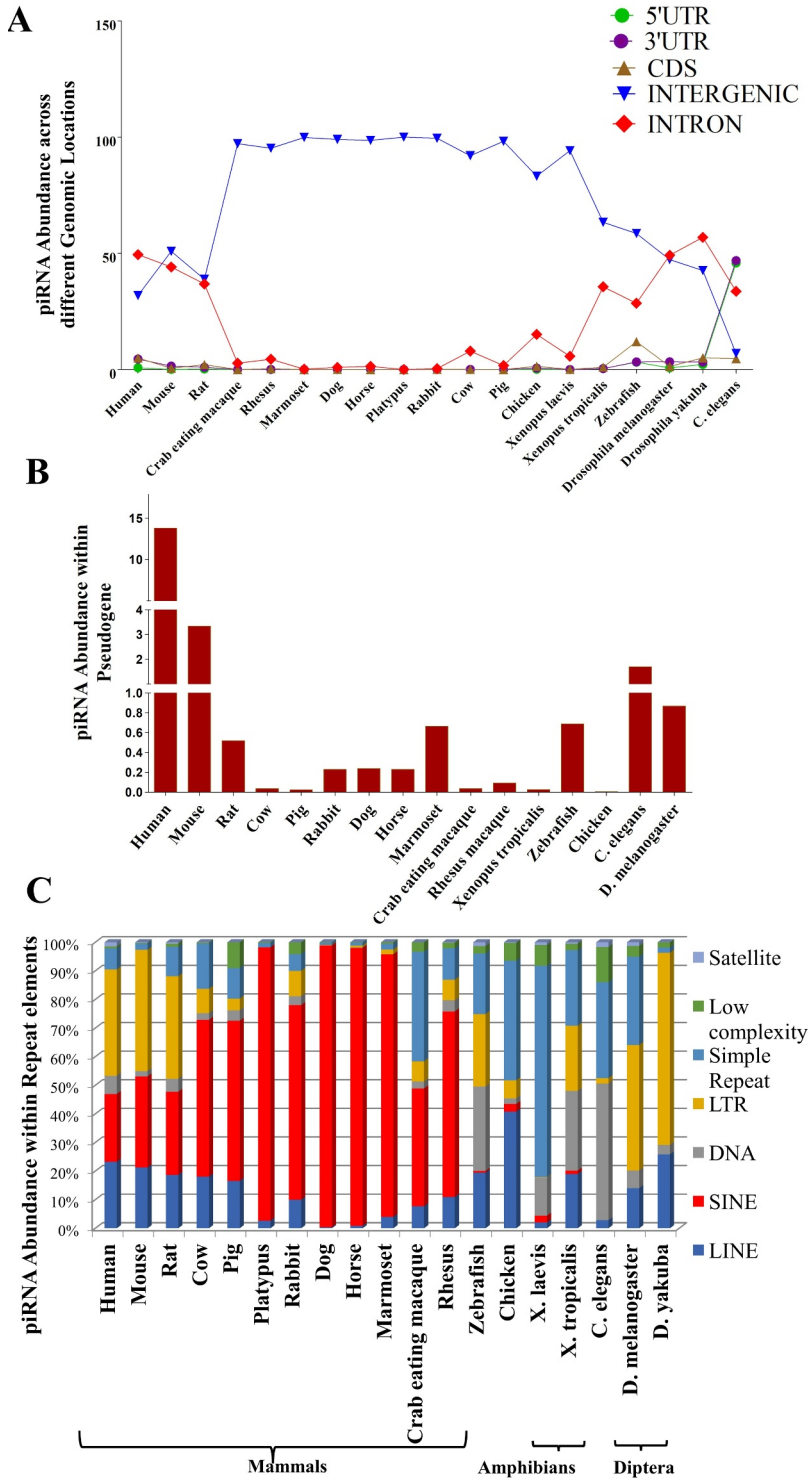


Figure 2. Distribution of multispecies piRNAs: (a) across different genomic locations, (b) within pseudogenes and (c) within repeat family.
Abbreviations used: UTR- untranslated region, CDS – coding DNA sequence, LTR – Long-terminal repeat, LINE – Long interspersed nuclear elements, SINE – Short interspersed nuclear elements.

enrichment of SINE repeat associated piRNA loci for 12 mammalian species. In addition to this, several human, mouse and rat piRNA loci overlap with LTR repeat family. On the contrary, amphibian piRNAs show a tendency to overlap with DNA and Simple repeat family. Petersen et al. [54] has reported the abundance of LTR repeats within those genomic loci corresponding to the transposable elements present in Diptera (a particular order of insect class). Our study also reveals similar observation in case of the order Diptera, where piRNA enriched regions corresponding to this order overlap with LTR repeat family. Presence of such repeat regions within piRNA loci can have important implications as is shown by Halbach et al. [52]. Here, it has been reported that satellite repeats

modulate global gene expression via piRNA-mediated gene silencing which is important for embryonic development of Aedes.

Biogenesis of piRNAs – the piRNA clusters and ping-pong amplification

piRNA clusters are also known as the hotspots of piRNA biogenesis. Initially, in the first version of piRNAQuest, the method described by Lau et al. [55] was followed to identify the piRNA clusters within a chromosome. Here a fixed window length of 20 kilobases (kb) was used to identify the clusters. Later in 2016, Rosenkranz reported that the piRNA clusters are not equally distributed across the chromosomes and is not even related to the length of the chromosome [27]. As the piRNA read distribution varies across the genome corresponding to different chromosomes, one should not fix the window size for detecting piRNA cluster. Hence, in this new version of our database, piRNAQuest V.2, we have adopted density based clustering approach [33] to identify the piRNA clusters (Supplementary Figures SF3 and SF4) which was found to be effective to recognize clusters successfully in chicken germ cell [56].

We obtained maximum no. of clusters in chromosome 15 and chromosome IV for human (Supplementary Figure SF3) and *C. elegans* respectively (Supplementary Figure SF4). We also found the same for human previously. In *C. elegans*, it is reported that maximum clusters lie within chromosome IV [44]. Though the function is still unknown, it has been found that among the sex determining chromosomes, 'X' chromosomal piRNAs mainly originate from clusters compared to the 'Y' chromosomal piRNAs [27]. Our analyses also have revealed more piRNA clusters in 'X' chromosomes than that in 'Y' chromosome of human, mouse and rat.

piRNA clusters overlapping with lncRNAs: We have studied the distribution of piRNA clusters within the lncRNA loci obtained from LncRBase V.2 [57]. As mentioned earlier, in the first point of this result section, we observed a significant overlap of piRNA clusters with the intergenic lncRNAs (Supplementary Figure SF5A) which are transcribed from in between two gene loci. This goes in line with previous reports [58]. In addition, we looked at the overlap of piRNA clusters with repeat regions (Supplementary Figure SF5B) and found similar observation as that obtained from the distribution of piRNAs in repeat elements (as shown in Figure 2(c)).

Motifs within piRNA clusters: Characteristic motifs have been identified for each of the clusters. These highly conserved motifs within the piRNA clusters may provide us information on possible common piRNA binding sites within its target gene. piRNAs from a cluster generated from coding gene regions can also regulate its 'host' gene expression [59]. A significant % of total piRNA clusters have been found to be overlapping with coding regions in many species (Supplementary Figure SF5C).

piRNAs are also generated via secondary biogenesis or the ping-pong amplification loop. Studies on fly have shown that somatic piRNAs generally do not show ping-pong pattern, suggesting that the ping-pong loop may work mainly in germline cells [60,61]. The distribution of ping-pong piRNAs among the different chromosomes was determined. Chromosome 15 in human shows predominance for ping-pong piRNAs and has also been reported very recently by Ray and Pandey [62]. More than 50% of the ping-pong piRNAs in human are found to overlap with protein-coding genes which indicates towards piRNA-dependent gene regulation [63]. Further, less than 10% of these piRNAs are found to overlap with repeat elements among which SINE repeat family is predominant. Previously, Das et al. [64] showed that ping-pong amplification does not occur in nematode, but surprisingly in our analysis, 509 ping-pong piRNAs are found in chromosome IV of *C. elegans* which may instigate the role of ping-pong loop in nematode as well.

Tissue-specific expression of piRNAs

Initially, piRNAs were observed to be expressed exclusively in germline cells [3]. But gradually they have been identified in somatic cells and the somatic piRNA pathway have been seen to regulate germline transpositions [65]. Hence, we have analysed 243 small RNA sequencing samples for 32 tissue types corresponding to 25 species (Supplementary File S2) in which 13 tissue types are from human. Supplementary Figure SF6 reveals the expression pattern of piRNAs among different tissue types corresponding to all these 25 species. For human, we have found the presence of maximum number of piRNAs in brain followed by colon, testis, spermatozoa which indicates the role of piRNAs not only in the germline cells, but also in other somatic cells. Previously, it was shown that there are piRNA complexes in mouse dendritic spines of brain and knockdown of those piRNAs resulted in lower spine density in the axons [45]. Recent studies also indicate that piRNAs in brain are associated in suppressing retrotransposons. This has a significant role in brain pathology [66]. It has been found that the piRNA length distribution is related to the age of the individual belonging to a particular species, e.g. in Drosophila the length of piRNAs becomes shorter with age [67]. Further, loss of methyltranferase result in piRNA instability and reduction in piRNA length and volume, which ultimately leads to male sterility during spermatogenesis [68]. Interestingly, in our study, we have found the presence of piRNAs, which are around 36 nts in length in human sperm samples, whereas such longer piRNAs have been seen to be expressed very less in any somatic cells.

Disease specific expression of piRNAs

With developments in pathological research, studies have highlighted the importance of piRNAs in disease systems. piRNAs and PIWI proteins are found to be expressed abnormally in several cancer systems that increases their importance as potential novel biomarkers for therapeutic research [19]. Recent evidences suggest that genomic stability of neurons may be disturbed by dysregulation of the piRNA pathways which results in various neurodegenerative disorders [69]. As genes involved in the biogenesis of piRNAs have an essential role in spermatogenesis, mutation in those genes may lead to male infertility [70]. Besides, piRNAs are shown to regulate Th2 cell development by downregulating IL-4,

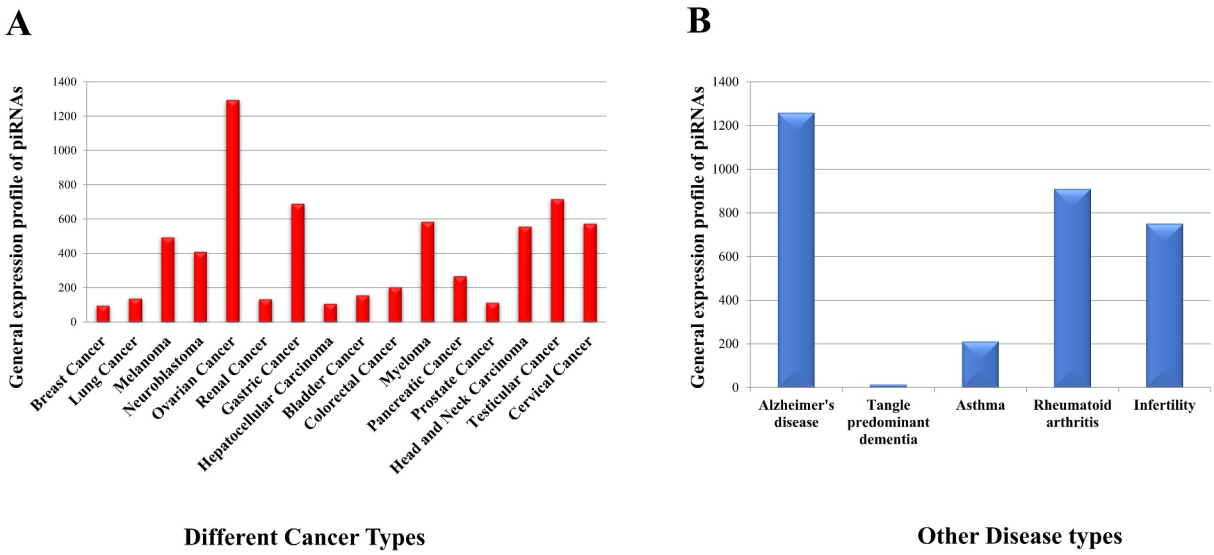


Figure 3. General expression profile of piRNAs in (a) different cancers and (b) other disease systems.

thus inhibiting allergic inflammation and asthma [71] and have specific binding partners in synovial fibroblasts, suggesting its role in inflammatory processes like Rheumatoid Arthritis [72]. Here, we have analysed 211 samples corresponding to 21 disease types (Supplementary File S3) in which 16 types of cancer are present. The distribution of piRNAs (Figure 3(a)) among different cancers shows the higher contribution of piRNAs in germ cell cancers like ovarian and testicular cancer. Here, our observation goes in line with the established role of piRNAs towards maintaining germ cells [73].

Among other diseases (Figure 3(b)), we found the presence of 1274 piRNAs among which hsa_piRNA_425 is highly abundant and hsa_piRNA_28207 is lowly abundant as compared to the abundance of other piRNAs in Alzheimer's disease. These have been reported previously [36]. The number of piRNAs expressed in asthma and rheumatoid arthritis are 278 and 910 respectively. Another interesting observation in this dataset is regarding the length of the piRNAs. In our study, we have observed the presence of longer piRNAs in sperm sample where maximum length of piRNAs is 32 nts in case of infertile samples indicating the significance of piRNA length towards spermatogenesis [68].

Beside the general expression profile of piRNAs in different diseases, differential expression analysis has been also performed using DESeq [74] to see the differential mode of regulation of piRNAs in seven cancer systems and asthenozoospermia (based on the availability of both test and control datasets). Table 2 shows the number of differentially expressed piRNAs among which the expression of some piRNAs corroborated with that obtained from literature evidences. For example, hsa_piRNA_9871 and hsa_piRNA_27200 are found to be upregulated in breast and lung cancer respectively [75]. This has also been observed in our study. The upregulated piRNAs hsa_piRNA_7806 and hsa_piRNA_31147 promote proliferation and invasiveness in colon [76] and renal cancer [77], respectively, and are also observed to be upregulated in our analysis.

piRNA-target gene interaction

Beside piRNA mediated cleavage of transposable elements, piRNAs are also known to target mRNAs and lncRNAs and subsequently regulate their expression. The involvement of piRNAs in regulating mRNAs has been studied extensively [36,78,79]. In a way, similar to the slicing of mRNAs, PIWI-piRNA complex can target lncRNAs which has been observed in multiple organisms [80]. It has been reported that a decrease in the expression levels of the target may correspond to an increase in the expression levels of the targeting piRNA, and vice versa [81]. Hence, for precise target prediction, we have screened those piRNAs and mRNAs as well as piRNAs and lncRNAs whose expression are negatively correlated. This has limited our analysis to those cancer datasets where both long and small RNA seq datasets are available. Hence, we have been able to predict piRNA-mRNA and piRNA-lncRNA interaction for 7 cancer systems (viz. lung, breast, renal, hepatocellular, ovarian, prostate and colorectal). The input dataset have been shown in Supplementary File S4 and the differential analysis was performed using the 'New Tuxedo' protocol [82]. Sequence based target prediction has been done using miRanda. Tissue and cell line data have been analysed separately. In order to highlight the role of piRNAs in different developmental stages, we have analysed the small RNA data corresponding to different developmental stages of mouse testis viz. 10dpp (days post-partum) and 16.5dpc (days postcoitum) as compared to that of six months old adult mouse testis. Further, piRNAs have been analysed corresponding to another disease system named asthenozoospermia where the sperm motility gets reduced in semen sample. The differentially expressed mRNAs, lncRNAs and piRNAs are mentioned in Table 2. The final set of piRNA-mRNA and piRNA-lncRNA target pairs for 7 cancer types is shown in Figure 4(a,b), respectively. Figure 4(c,d) shows the number of piRNA targets within mRNAs and lncRNAs respectively in two developmental stages of mouse testis. Moreover, we have also curated experimentally validated piRNA–mRNA target pairs for human, mouse and *C. elegans*.

Table 2. Differentially expressed piRNAs, genes and lncRNAs in different cancer systems, Asthenozoospermia and different developmental stages of mouse testis

	Differentially expressed piRNAs		Differentially expressed Genes		Differentially expressed lncRNAs	
	Upregulated piRNAs	Downregulated piRNAs	Upregulated genes	Downregulated genes	Upregulated lncRNAs	Downregulated lncRNAs
Different Cancer systems						
Breast cancer	253	133	955	1165	1087	615
Lung cancer	269	349	517	513	53	82
Ovarian cancer	208	376	3300	2376	335	1441
Renal cancer	177	724	197	162	85	40
Hepatocellular carcinoma	352	155	116	125	169	171
Colon cancer	366	475	344	361	79	119
Prostate cancer	413	328	2135	2368	504	179
Other disease						
Asthenozoospermia	1622	2170	318	133	2957	1062
Different developmental stages of mouse testis						
10 dpp	923	439	2173	2429	1445	1442
16.5 dpc	260	143	7281	6498	3209	2430

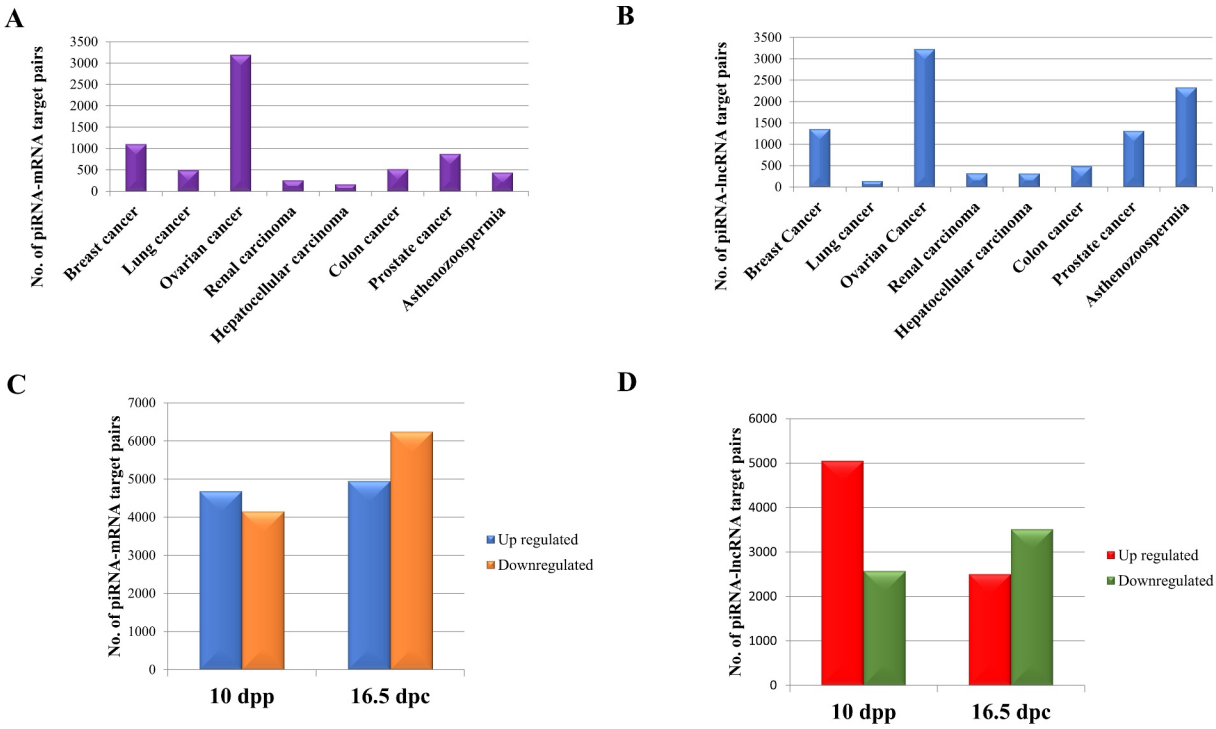


Figure 4. Predicted piRNA targets in (a) protein coding genes and (b) LncRNAs within disease systems; (c) protein coding genes and (d) LncRNAs across different developmental stages of mouse testis.

Discussion

There has been an increase in the number of piRNAs that have been identified in different species as well as in different cells since our first release of piRNAQuest in 2014. Initially, we developed piRNAQuest, with a goal to develop a non-redundant comprehensive catalogue of human, mouse and rat piRNAs so as to provide a better understanding regarding their genomic localization, overlaps with genomic elements and their association with other lncRNAs. Although, initial reports reveal the main functions of piRNAs to be transposon silencing [51] and maintenance of gene integrity mainly in germline cells [9], but later it has been identified in somatic cells as well in many species [10,12]. All these put forward, the increasing importance of its diverse functions not only in transposon silencing but also in gene expression regulation. Hence, we have come up with this new version of piRNAQuest named as piRNAQuest V.2, where we have expanded our study to 25 new species (apart from those included in previous version) covering different phylum or classes. Along with the previous features, piRNAQuest V.2 has focused on several new aspects such as directionality of piRNA cluster, piRNA expression among normal tissues and disease systems and its targets among protein coding genes and lncRNAs. These will open up novel avenues for piRNA research.

Over time, many studies have demonstrated the mechanism of primary biogenesis of piRNAs from piRNA clusters [22,23]. Several protocols have been developed to identify them. However, lack of uniform distribution of piRNAs among the chromosomes lead us to consider the density based clustering approach to identify piRNA clusters. It will help in understanding the distribution of piRNAs throughout the genome and the formation of clusters which are the 'hotspots' of piRNAs for primary biogenesis. Besides, secondary biogenesis via ping-pong amplification is also important for generation of piRNAs and its role towards silencing of its target. Emphasizing on this, we checked the ping-pong overlap among the piRNAs and have also provided options to visualize the ping-pong signature within the piRNAs. In human, we have seen the presence of maximum piRNAs in chromosome 15 where the maximum number of piRNA clusters and ping-pong piRNAs are also present.

In addition to this, analysing piRNA expression profile of various normal and disease systems will help us to understand the piRNA-mediated gene regulation in those systems. In this version, we have incorporated the piRNA expression profile of 21 disease systems along with several normal tissue data corresponding to different species. As piRNAs are differentially regulated between disease and normal conditions, a decrease in the expression levels of the target should correspond to an increase in the expression levels of the targeting piRNA, and vice versa. Taking this as an opportunity, to unravel the connection between piRNA expression and disease occurrence, we have predicted probable piRNA targets which may serve as promising biomarkers for early diagnosis and act as therapeutic targets for diseases like cancer. Further, in order to show the involvement of piRNAs in different developmental stages, we have predicted piRNA targets within mRNAs and lncRNAs in different developmental stages of mouse testis.

Overall, the newly added features along with the existing ones will make piRNAQuest V.2 a user friendly, comprehensive database for piRNAs. Our future goal is to update the database regularly with newly annotated piRNAs along with its novel features in order to continue contributing to the growing piRNA knowledgebase.

Materials and method

Improved content and new features

Input dataset

piRNA entries have been extended to 25 new species in addition to human, rat and mouse. The genome builds, availability of genome annotation and repeat annotation information and the number of piRNAs corresponding to the species has been mentioned in **Supplementary Table ST1**. The genome builds are updated from hg19 to hg38 and rn5.0 to rn6.0 for human and rat respectively. Data are collected in different formats like fasta, gtf and bed from the respective sources. Repeat elements and Refseq annotated 5′UTR, 3′UTR, exon, intron and CDS information have been downloaded from UCSC [41]. The miRNA information has been downloaded from miRBase 22 release. Annotated piRNA sequences were downloaded in.fasta format from National Centre for Biotechnology Information (NCBI) [83]. Normal tissue and disease specific small and long RNA sequencing data has been obtained from NCBI Gene Expression Omnibus (GEO) [84]. LncRNA information has been retrieved from LncRBase V.2 [57].

Data processing and refinement

Redundancy check and ID assignment: The procedure of assigning IDs to non redundant piRNA entries is similar to that followed in piRNAQuest [30]. The sequencing data were aligned to respective genome. We further filtered out those reads mapped to other ncRNAs and screened the reads predicted to be piRNAs using our in-house script. Thereafter, non-redundant reads were re-aligned with reference genome for complete alignment with no mismatches and annotated with unique piRNAQuest IDs, i.e. [three letter abbreviation of species name]_piRNA_[number]. The annotation IDs are same for human and mouse as assigned in the previous version. The only difference is in the annotation of rat from the previous one as in the last version it was not annotated as three letter abbreviation of species name. Users can find the previous IDs which are annotated in this version of the database in the ID conversion of Help menu for human, mouse and rat. To study the distribution of piRNAs within genome, we searched for the localization of piRNAs within gene, intergenic regions, intron, CDS, UTR regions, repeat elements and pseudogenes using in-house perl scripts as that followed in the first version.

Density based piRNA cluster prediction: Previously used cluster prediction protocol [55] have a disadvantage of considering window size of fixed length for all species and hence does not account for the variation in read distribution among different species. To overcome this discrepancy, we have adopted density based clustering algorithm DBSCAN [33] to develop a python based in-house protocol for identifying piRNA clusters which is based on the read distribution of piRNAs across the genome.

Clustering parameters: There are two parameters 'Eps (or epsilon)' and 'MinReads' which allow us to find candidate clusters. 'Eps' is defined as the distance of a read from a neighbourhood point and 'MinReads' are the minimum number of reads within 'eps' distance. To determine the clustering parameters, inter distance between the annotated piRNAs are calculated by performing k-dist analysis [33]. We calculated the distance between each mapped read and its kth nearest neighbouring read which is referred to as k-dist which is plotted with respect to the its counts. Eventually, a sharp valley has been observed in this 'count versus distance' plot until the k-dist follows a uniform distribution. The distance, for a given value of k, after which the graphs follows an asymptotic decrease is termed as the eps i.e. eps represents the distance which repeats itself for maximum number of times and hence has the highest probability of defining the boundaries of a cluster containing at least, the 'MinReads' which represents the number of reads within the cluster. After the 'Eps' and 'MinReads' parameters are set for each chromosome corresponding to each species, clusters are detected from the coordinate file of the annotated piRNAs.

Cluster score: In order to calculate piRNA enrichment within each cluster, we have calculated cluster score for each piRNA clusters. This has been calculated as follows:

$$\text{Cluster score} = \frac{\text{Total no of piRNAs in the cluster}}{\text{Minimum no of piRNAs needed to form the cluster(kth value)}} \quad (1)$$

We also checked for the strand specificity of the clusters based on the directionality of the constituent piRNAs within that cluster. If a cluster contains both sense and antisense piRNAs, it is considered as 'dual strand cluster' and if it contains only sense piRNAs or antisense piRNAs, it is considered as 'uni-strand cluster'.

Localization of piRNA clusters and characteristic motifs within them: (i) In-house perl scripts have been used to check for any overlap of piRNA clusters with coding genes or lncRNAs or the repeat regions. (ii) piRNAs show a strong tendency to form clusters in the syntenic regions of genome [3]. We have downloaded the syntenic regions from UCSC [41] and searched for piRNA clusters in the corresponding syntenic regions among different species. (iii) MEME have been used to find the presence of any significant motifs within the piRNA clusters [85].

Ping-pong pattern within piRNAs: The secondary mode of piRNA biogenesis, i.e ping pong amplification shows a distinct sequence based feature within the piRNAs, i.e. a 10 nt overlap is found between the antisense and sense piRNAs. An in-house python script has been developed to identify these ping-pong piRNAs and visualize this ping-pong signature pattern.

piRNA profile in Normal and Disease systems: We have downloaded small RNA sequencing data for different tissue types from GEO (https://www.ncbi.nlm.nih.gov/geo). A total of 243 samples were analysed for 32 types of normal tissue samples for different species. Along with the normal dataset, we have analysed 211 samples corresponding to 21 types of

disease data, which includes 16 different types of cancer data sets. To analyse the expression profile of piRNAs among the normal tissue and disease systems, BLAST [86] and in-house perl scripts were used. Further the expression level of each piRNA found in a sample was normalized by counts per million (CPM) and were further screened based on the z-score [87] lying between −3 and +3. Users will be able to view the expression of 200 most abundant piRNAs in each set. Additionally, we have checked the distribution of each piRNAs among all the normal tissues or disease systems which has been represented graphically to provide better understanding regarding their expression within different systems.

piRNA Target prediction: piRNA target pairs have been predicted between up regulated piRNAs and downregulated lncRNAs and mRNAs and vice versa. miRanda [40] has been used for predicting piRNA targets within lncRNAs (sequences obtained from LncRBase V.2 [57]) and 3′ UTR region of mRNAs (sequences downloaded from UCSC [41]). The target score and energy threshold are 170 and −20 kcal/mol respectively [88]. In the database, we have linked The Human Protein Atlas [89] and Pathway Commons [90] databases to the targeted genes for further pathway and pathology based analysis. Further, our database hosts several experimentally validated piRNA targets for human, mouse and *C. elegans* which have been manually curated from published reports.

The overall workflow has been outlined in (Supplementary Figure SF7).

Database execution

In piRNAQuest V.2, a query is basically processed via simple searching options using user's desired selection criterion and information are presented on the web interface after retrieving related details from the database. The general information page displays basic information related to the piRNAs and provides options to probe into its further genomic details which is shown in Figure 5(a).

Search and output options

(a) The following options are under the '**Search piRNAs**' menu

(1) ***Search by Species Name***: Users can browse all piRNAs by selecting a particular species name with the help of previous or next buttons.

(2) ***Search by piRNA Accession ID/Chromosomal Co-ordinates***: Users can search by piRNA accession ID for detailed information (piRNA sequence, its length, its NCBI ID (if any), %GC content, piRNA position corresponding to the genome build along with its genomic localization within genes, introns, CDS, 3′UTR, 5′UTR, intergenic regions, and repetitive elements) of selected species. A piRNA ID has already been provided as an example for each of the species. Using desired chromosomal co-ordinates, users can also get above mentioned information about piRNAs.

(3) ***Search piRNAs by Sequence:*** Users can retrieve piRNA information by providing piRNA sequences. The sequence length should be greater than at least 20 nucleotides.

(4) ***Search piRNA within Genes:***User can search piRNAs present within Genes by providing a Gene Name corresponding to the selected species. The result page will show the piRNAs whose loci overlappes with this particular gene. User can also search for piRNAs within Genes of a particular species by providing chromosomal coordinates corresponding to that species.

(5) ***Search piRNA within repeats:*** Users can search for piRNAs whose loci get mapped within repeats corresponding to genomic locations (viz. 3/UTR, 5/UTR, introns, CDS, intergenic regions) for a particular Repeat Family. Users can also search for repeat-associated piRNAs selecting their desired chromosomal location.

(6) ***Search piRNAs with Ping-Pong features:*** User can search for 10nt overlapping piRNAs within a particular chromosome of a particular species by selecting chromosome number corresponding to that species Figure 5(b).

(b) Search '**piRNA clusters**'

(1) ***Search clusters by chromosomal co-ordinates***: Users can obtain piRNA clusters by submitting a particular chromosomal location Figure 5(c). This will fetch cluster loci, cluster score, total number of piRNAs within the cluster, cluster strandedness, prevalence of these piRNAs in minus/plus strand, and the corresponding characteristic motif of the cluster in that location. The link on the motif navigates to the website (https://meme-suite.org/meme) where users can perform further study on the motif.

(2) ***Search mRNAs/lncRNAs/Repeats within piRNA Clusters***: Users can check if piRNA clusters are overlapped with mRNA/lncRNA loci or with the repeat elements.

(3) ***Search piRNA Clusters in Syntenic Regions***: Users can search for piRNA clusters overlapping with syntenic regions by selecting a particular chromosome for both target and query organisms.

(c) Browse **'piRNA Expression'**

(1) ***Search Tissue specific expression***: User can view piRNA expression pattern by selecting tissue type and will be able to see the top 200 most abundantly expressed piRNAs by submitting the view option corresponding to the dataset.

(2) ***Search Disease specific expression***: User can also retrieve same information as above for different disease systems.

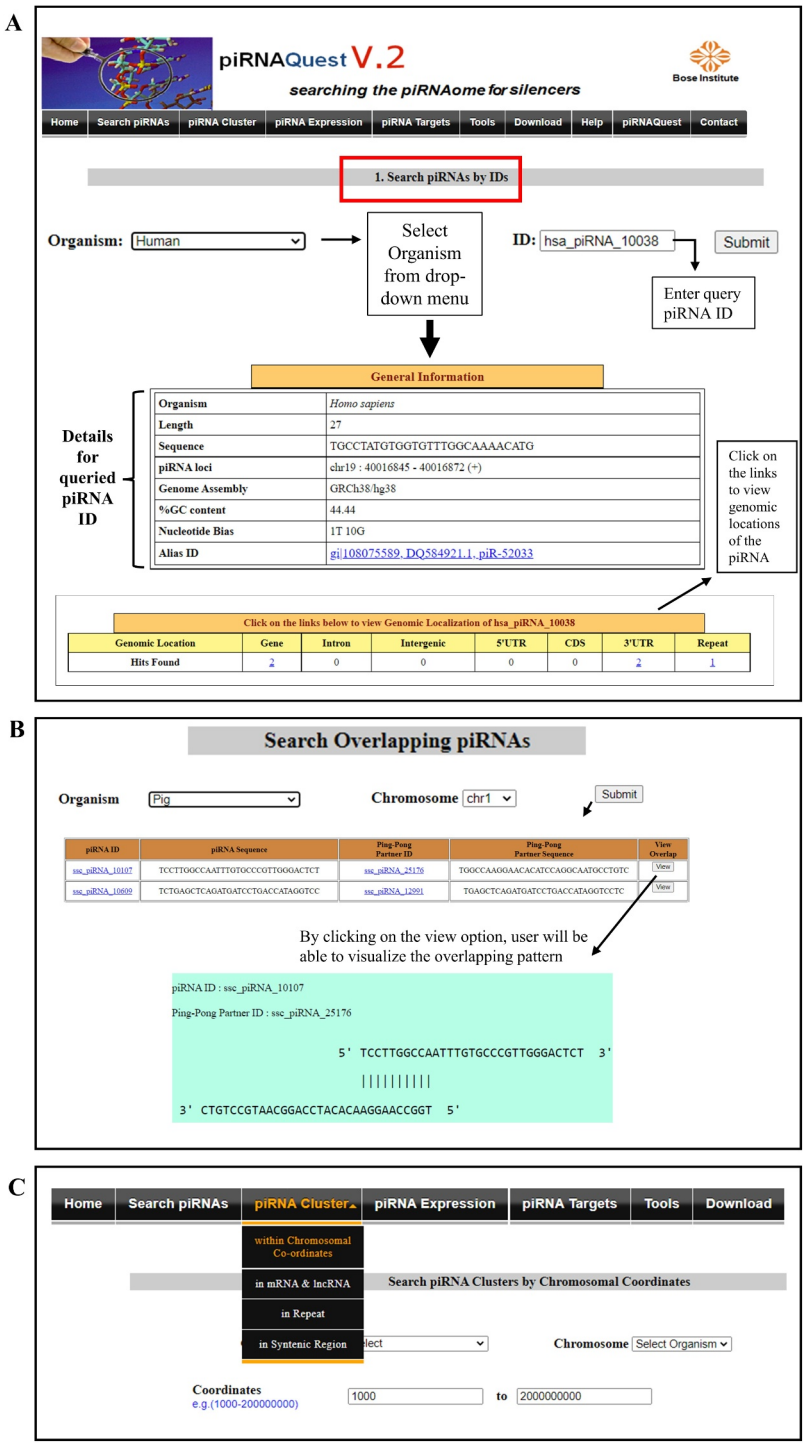


Figure 5. Web interfaces for easy access of piRNAQuest V.2 showing: (a) search options through a piRNA ID and the corresponding result page; (b) search options for pingpong piRNAs and visualizing its pattern; and (c) search options to browse piRNA clusters.

We have provided an additional option of downloading the entire set of tissue/disease wise piRNA expression information for all the samples from the download section.

An additional search option is there under both the above mentioned menus to retrieve expression level of a particular piRNA in normal tissues of selected species or that in human disease systems Figure 6(a).

(d) Search '**piRNA Targets**'

(1) ***Search Predicted Targets***: Users can search piRNA targets in mRNA/lncRNA for negatively correlated dataset by selecting the disease type/different developmental stages of mouse and the mode of regulation of the piRNA. After clicking the details option, user will be able to get the detailed prediction result and visualize the piRNA-target duplex structure Figure 6(b).

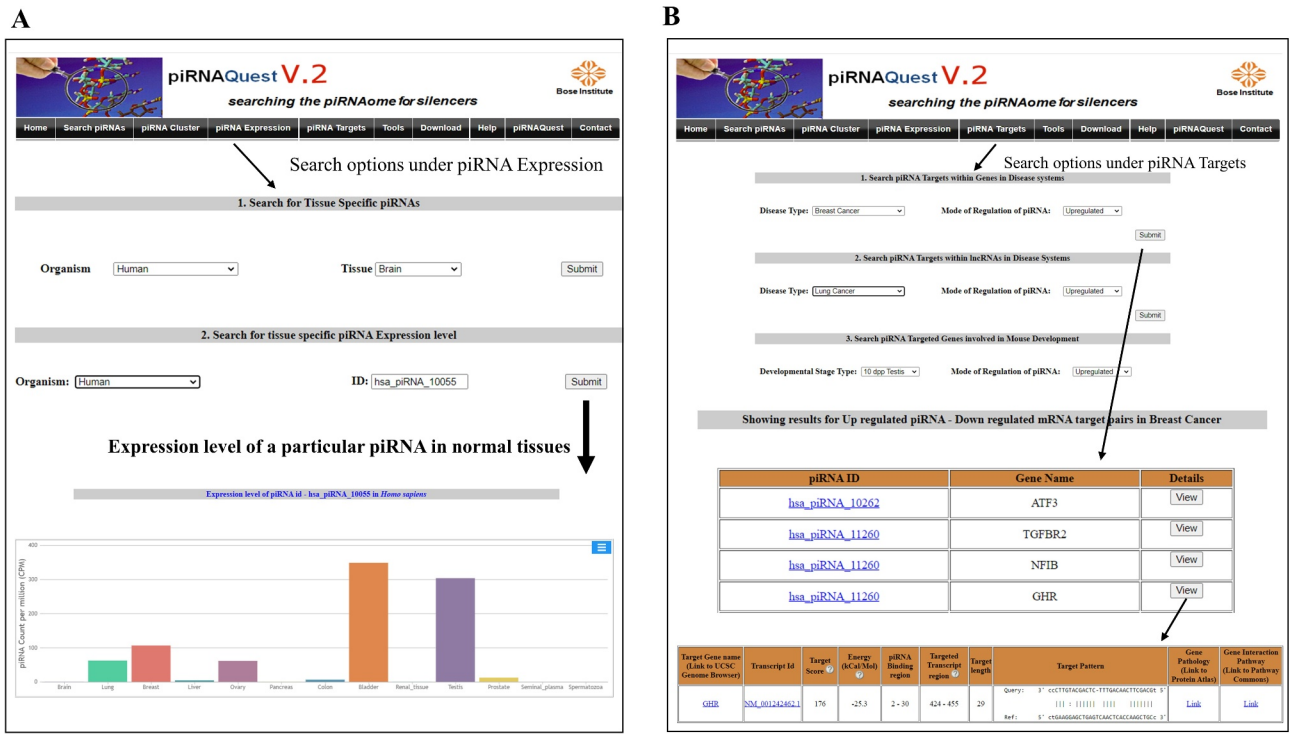


Figure 6. Web interfaces for easy access of piRNAQuest V.2 showing: (a) tissue wise expression values of individual piRNAQuest IDs and the corresponding output and (b) search options and detailed output for piRNA target prediction.

(2) ***Search Curated Targets***: In this section, users can find literature curated piRNA-gene target pairs.

Tools

(1) ***Dynamic piRNA cluster detection***: This tool can detect piRNA clusters where user can set parameters of their own, like the chromosomal coordinates, Eps distance and MinReads.

(2) ***piRNA Target prediction***: Users can provide the piRNA and target sequences of their own choice along with their desired energy parameters and threshold target score to predict piRNA targets.

(3) ***Ping-pong signature detection***: Users need to provide piRNA sequences in.fasta format to visualize ping-pong signature pattern within them.

Availability

piRNAQuest V.2 is available at http://dibresources.jcbose.ac.in/zhumur/pirnaquest2. Files can be freely downloaded and used in accordance with the GNU Public License.

Acknowledgments

We are grateful to the Department of Science and Technology (DST) for financial support. We thank and acknowledge Samarpita Sen and S. Shanmugapriya (summer trainees) for their contribution towards building the database.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the DST, India.

References

[1] Han LC, Chen Y. Small and long non-coding RNAs: novel targets in perspective cancer therapy. Curr Genomics. 2015 Oct;16 (5):319–326.

[2] Aravin A, Gaidatzis D, Pfeffer S, et al. A novel class of small RNAs bind to MILI protein in mouse testes. Nature. 2006 Jul 13;442 (7099):203–207.

[3] Girard A, Sachidanandam R, Hannon GJ, et al. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature. 2006 Jul 13;442(7099):199–202.

[4] Grivna ST, Beyret E, Wang Z, et al. A novel class of small RNAs in mouse spermatogenic cells. Genes Dev. 2006 Jul 1;20 (13):1709–1714.

[5] Grivna ST, Pyhtila B, Lin H. MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. Proc Natl Acad Sci U S A. 2006 Sep 5;103 (36):13415–13420.

[6] Saito K, Nishida KM, Mori T, et al. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. Genes Dev. 2006 Aug 15;20 (16):2214–2222.

[7] Gunawardane LS, Saito K, Nishida KM, et al. A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in Drosophila. Science. 2007 Mar 16;315(5818):1587–1590.

[8] Nishida KM, Saito K, Mori T, et al. Gene silencing mechanisms mediated by Aubergine piRNA complexes in Drosophila male gonad. RNA. 2007 Nov;13(11):1911–1922.

[9] Thomson T, Lin H. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. Annu Rev Cell Dev Biol. 2009;25(1):355–376.

[10] Brennecke J, Aravin AA, Stark A, et al. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell. 2007 Mar 23;128(6):1089–1103.
[11] Houwing S, Kamminga LM, Berezikov E, et al. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell. 2007 Apr 6;129(1):69–82.
[12] Siomi MC, Sato K, Pezic D, et al. PIWI-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol. 2011 Apr;12(4):246–258.
[13] Czech B, Hannon GJ. one loop to rule them all: the ping-pong cycle and piRNA-guided silencing. Trends Biochem Sci. 2016 Apr;41(4):324–337.
[14] Aravin AA, Hannon GJ, Brennecke J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science. 2007 Nov 2;318(5851):761–764.
[15] Fu A, Jacobs DI, Hoffman AE, et al. PIWI-interacting RNA 021285 is involved in breast tumorigenesis possibly by remodeling the cancer epigenome. Carcinogenesis. 2015 Oct;36 (10):1094–1102.
[16] Ortogero N, Schuster AS, Oliver DK, et al. A novel class of somatic small RNAs similar to germ cell pachytene PIWI-interacting small RNAs. J Biol Chem. 2014 Nov 21;289 (47):32824–32834.
[17] Williams Z, Morozov P, Mihailovic A, et al. Discovery and characterization of piRNAs in the human fetal ovary. Cell Rep. 2015 Oct 27;13(4):854–863.
[18] Huang X, Yuan T, Tschannen M, et al. Characterization of human plasma-derived exosomal RNAs by deep sequencing. BMC Genomics. 2013 May 10;14(1):319.
[19] Liu Y, Dou M, Song X, et al. The emerging role of the piRNA/piwi complex in cancer. Mol Cancer. 2019 Aug 9;18(1):123.
[20] Kalmykova AI, Klenov MS, Gvozdev VA. Argonaute protein PIWI controls mobilization of retrotransposons in the Drosophila male germline. Nucleic Acids Res. 2005;33 (6):2052–2059.
[21] Reuter M, Berninger P, Chuma S, et al. Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature. 2011 Nov 27;480(7376):264–267.
[22] Ishizu H, Siomi H, Siomi MC. Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. Genes Dev. 2012 Nov 1;26(21):2361–2373.
[23] Weick EM, Miska EA. piRNAs: from biogenesis to function. Development. 2014 Sep;141(18):3458–3471.
[24] Sai Lakshmi S, Agrawal S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Res. 2008 Jan;36(Database issue):D173–7.
[25] Wang J, Zhang P, Lu Y, et al. piRBase: a comprehensive database of piRNA sequences. Nucleic Acids Res. 2019 Jan 8;47(D1):D175–D180.
[26] Wu WS, Brown JS, Chen TT, et al. piRTarBase: a database of piRNA targeting sites and their roles in gene regulation. Nucleic Acids Res. 2019 Jan 8;47(D1):D181–D187.
[27] Rosenkranz D. piRNA cluster database: a web resource for piRNA producing loci. Nucleic Acids Res. 2016 Jan 4;44(D1):D223–30.
[28] Rosenkranz D, Zischler H. proTRAC–a software for probabilistic piRNA cluster detection, visualization and analysis. BMC Bioinformatics. 2012 Jan 10;13(1):5.
[29] Muhammad A, Waheed R, Khan NA, et al. piRDisease v1.0: a manually curated database for piRNA associated diseases. Database. 2019 Jan 1;2019. DOI:10.1093/database/baz052
[30] Sarkar A, Maji RK, Saha S, et al. piRNAQuest: searching the piRNAome for silencers. BMC Genomics. 2014 Jul 4;15(1):555.
[31] Monga I, Banerjee I. Computational identification of piRNAs using features based on RNA sequence, structure, thermodynamic and physicochemical properties. Curr Genomics. 2019 Nov;20 (7):508–518.
[32] Wang K, Hoeksema J, Liang C. piRNN: deep learning algorithm for piRNA prediction. PeerJ. 2018;6:e5429.
[33] Ester M, Kriegel HP, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD. 1996;96:226–231.
[34] Quek C, Bellingham SA, Jung CH, et al. Defining the purity of exosomes required for diagnostic profiling of small RNA suitable for biomarker discovery. RNA Biol. 2017 Feb;14(2):245–258.
[35] Bachmayr-Heyda A, Auer K, Sukhbaatar N, et al. Small RNAs and the competing endogenous RNA network in high grade serous ovarian cancer tumor spread. Oncotarget. 2016 Jun 28;7 (26):39640–39653.
[36] Roy J, Sarkar A, Parida S, et al. Small RNA sequencing revealed dysregulated piRNAs in Alzheimer's disease and their probable role in pathogenesis. Mol Biosyst. 2017 Feb 28;13(3):565–576.
[37] Li Y, Wu X, Gao H, et al. Piwi-interacting RNAs (piRNAs) are dysregulated in renal cell carcinoma and associated with tumor metastasis and cancer-specific survival. Mol Med. 2015 May 13;21 (1):381–388.
[38] Zhang W, Yao G, Wang J, et al. ncRPheno: a comprehensive database platform for identification and validation of disease related noncoding RNAs. RNA Biol. 2020 Jul;17(7):943–955.
[39] Zhang W, Zeng B, Yang M, et al. ncRNAVar: a manually curated database for identification of noncoding RNA variants associated with human diseases. J Mol Biol. 2021 May 28;433(11):166727.
[40] John B, Enright AJ, Aravin A, et al. Human microRNA targets. PLoS Biol. 2004 Nov;2(11):e363.
[41] Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC table browser data retrieval tool. Nucleic Acids Res. 2004 Jan 1;32 (Database issue):D493–6.
[42] Pujana MA, Nadal M, Gratacos M, et al. Additional complexity on human chromosome 15q: identification of a set of newly recognized duplicons (LCR15) on 15q11-q13, 15q24, and 15q26. Genome Res. 2001 Jan;11(1):98–111.
[43] Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007 Apr;8 (4):272–285.
[44] Ruby JG, Jan C, Player C, et al. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell. 2006 Dec 15;127(6):1193–1207.
[45] Lee EJ, Banerjee S, Zhou H, et al. Identification of piRNAs in the central nervous system. RNA. 2011 Jun;17(6):1090–1099.
[46] Nelson CE, Hersh BM, Carroll SB. The regulatory content of intergenic DNA shapes genome architecture. Genome Biol. 2004;5(4):R25.
[47] Pink RC, Wicks K, Caley DP, et al. Pseudogenes: pseudo-functional or key regulators in health and disease? RNA. 2011 May;17(5):792–798.
[48] Hirano T, Iwasaki YW, Lin ZY, et al. Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate. RNA. 2014 Aug;20 (8):1223–1237.
[49] Flicek P, Amode MR, Barrell D, et al. Ensembl 2012. Nucleic Acids Res. 2012 Jan;40(Database issue):D84–90.
[50] Pantano L, Jodar M, Bak M, et al. The small RNA content of human sperm reveals pseudogene-derived piRNAs complementary to protein-coding genes. RNA. 2015 Jun;21(6):1085–1095.
[51] Vagin VV, Sigova A, Li C, et al. A distinct small RNA pathway silences selfish genetic elements in the germline. Science. 2006 Jul 21;313(5785):320–324.
[52] Halbach R, Miesen P, Joosten J, et al. A satellite repeat-derived piRNA controls embryonic development of Aedes. Nature. 2020 Apr;580(7802):274–277.
[53] Vandewege MW, Platt RN 2nd, Ray DA, et al. Transposable element targeting by piRNAs in Laurasiatherians with distinct transposable element histories. Genome Biol Evol. 2016 May 9;8 (5):1327–1337.
[54] Petersen M, Armisen D, Gibbs RA, et al. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC Evol Biol. 2019 Jan 9;19 (1):11.
[55] Lau NC, Seto AG, Kim J, et al. Characterization of the piRNA complex from rat testes. Science. 2006 Jul 21;313(5785):363–367.
[56] Jung I, Park JC, Kim S. piClust: a density based piRNA clustering algorithm. Comput Biol Chem. 2014 Jun;50:60–67.

[57] Das T, Deb A, Parida S, et al. LncRBase V.2: an updated resource for multispecies lncRNAs and ClinicLSNP hosting genetic variants in lncRNAs for cancer patients. RNA Biol. 2020;18(8):1–16.
[58] Han BW, Zamore PD. piRNAs. Curr Biol. 2014 Aug 18;24(16): R730–3.
[59] Barberan-Soler S, Fontrodona L, Ribo A, et al. Co-option of the piRNA pathway for germline-specific alternative splicing of C. elegans TOR. Cell Rep. 2014 Sep 25;8(6):1609–1616.
[60] Saito K, Ishizu H, Komai M, et al. Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in Drosophila. Genes Dev. 2010 Nov 15;24(22):2493–2498.
[61] Lau NC, Robine N, Martin R, et al. Abundant primary piRNAs, endo-siRNAs, and microRNAs in a Drosophila ovary cell line. Genome Res. 2009 Oct;19(10):1776–1785.
[62] Ray R, Pandey P. piRNA analysis framework from small RNA-Seq data by a novel cluster prediction tool - PILFER. Genomics. 2018 Nov;110(6):355–365.
[63] Jehn J, Gebert D, Pipilescu F, et al. PIWI genes and piRNAs are ubiquitously expressed in mollusks and show patterns of lineage-specific adaptation. Commun Biol. 2018;1(1):137.
[64] Das PP, Bagijn MP, Goldstein LD, et al. Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the Caenorhabditis elegans germline. Mol Cell. 2008 Jul 11;31(1):79–90.
[65] Barckmann B, El-Barouk M, Pelisson A, et al. The somatic piRNA pathway controls germline transposition over generations. Nucleic Acids Res. 2018 Oct 12;46(18):9524–9536.
[66] Nandi S, Chandramohan D, Fioriti L, et al. Roles for small noncoding RNAs in silencing of retrotransposons in the mammalian brain. Proc Natl Acad Sci U S A. 2016 Nov 8;113 (45):12697–12702.
[67] Wang H, Ma Z, Niu K, et al. Antagonistic roles of Nibbler and Hen1 in modulating piRNA 3′ ends in Drosophila. Development. 2016 Feb 1;143(3):530–539.
[68] Lim SL, Qu ZP, Kortschak RD, et al. HENMT1 and piRNA stability are required for adult male germ cell transposon repression and to define the spermatogenic program in the mouse. PLoS Genet. 2015 Oct;11(10):e1005620.
[69] Kim KW. PIWI proteins and piRNAs in the nervous system. Mol Cells. 2019 Dec 31;42(12):828–835.
[70] Kamaliyan Z, Pouriamanesh S, Soosanabadi M, et al. Investigation of piwi-interacting RNA pathway genes role in idiopathic non-obstructive azoospermia. Sci Rep. 2018 Jan 9;8(1):142.
[71] Zhong F, Zhou N, Wu K, et al. A SnoRNA-derived piRNA interacts with human interleukin-4 pre-mRNA and induces its decay in nuclear exosomes. Nucleic Acids Res. 2015 Dec 2;43 (21):10474–10491.
[72] Plestilova L, Neidhart M, Russo G, et al. Expression and regulation of PIWIL-Proteins and PIWI-interacting RNAs in rheumatoid arthritis. PloS One. 2016;11(11):e0166920.
[73] Juliano C, Wang J, Lin H. Uniting germline and stem cells: the function of Piwi proteins and the piRNA pathway in diverse organisms. Annu Rev Genet. 2011;45(1):447–469.
[74] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
[75] Reeves ME, Firek M, Jliedi A, et al. Identification and characterization of RASSF1C piRNA target genes in lung cancer cells. Oncotarget. 2017 May 23;8(21):34268–34282.
[76] Mai D, Ding P, Tan L, et al. PIWI-interacting RNA-54265 is oncogenic and a potential therapeutic target in colorectal adenocarcinoma. Theranostics. 2018;8(19):5213–5230.
[77] Busch J, Ralla B, Jung M, et al. Piwi-interacting RNAs as novel prognostic markers in clear cell renal cell carcinomas. J Exp Clin Cancer Res. 2015 Jun 14;34(1):61.
[78] Zuo Y, Liang Y, Zhang J, et al. Transcriptome analysis identifies Piwi-interacting RNAs as prognostic markers for recurrence of prostate cancer. Front Genet. 2019;10:1018.
[79] Weng W, Liu N, Toiyama Y, et al. Novel evidence for a PIWI-interacting RNA (piRNA) as an oncogenic mediator of disease progression, and a potential prognostic biomarker in colorectal cancer. Mol Cancer. 2018 Jan 30;17(1):16.
[80] Wang C, Lin H. Roles of piRNAs in transposon and pseudogene regulation of germline mRNAs and lncRNAs. Genome Biol. 2021 Jan 8;22(1):27.
[81] Krishnan P, Ghosh S, Wang B, et al. Profiling of small nucleolar RNAs by next generation sequencing: potential new players for breast cancer prognosis. PloS One. 2016;11(9):e0162622.
[82] Pertea M, Kim D, Pertea GM, et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, stringtie and Ballgown. Nat Protoc. 2016 Sep;11(9):1650–1667.
[83] Geer LY, Marchler-Bauer A, Geer RC, et al. The NCBI bioSystems database. Nucleic Acids Res. 2010 Jan;38(Database issue):D492–6.
[84] Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol. 2006;411:352–369.
[85] Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009 Jul;37 (Web Server issue):W202–8.
[86] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–410.
[87] Hazra A, Gogtay N. Biostatistics series module 1: basics of biostatistics. Indian J Dermatol. 2016 Jan-Feb;61(1):10–20.
[88] Hashim A, Rizzo F, Marchese G, et al. RNA sequencing identifies specific PIWI-interacting small non-coding RNA expression patterns in breast cancer. Oncotarget. 2014 Oct 30;5(20):9901–9910.
[89] Ponten F, Jirstrom K, Uhlen M. The human protein atlas–a tool for pathology. J Pathol. 2008 Dec;216(4):387–393.
[90] Cerami EG, Gross BE, Demir E, et al. Pathway commons, a web resource for biological pathway data. Nucleic Acids Res. 2011 Jan;39(Database issue):D685–90.

Send Orders for Reprints to reprints@benthamscience.net

RESEARCH ARTICLE

Mapping Drug-gene Interactions to Identify Potential Drug Candidates Targeting Envelope Protein in SARS-CoV-2 Infection

Byapti Ghosh[1,#], Troyee Das[1,#], Gourab Das[1,#], Nilkanta Chowdhury[2,3,#], Angshuman Bagchi[2,*] and Zhumur Ghosh[1,*]

[1]*Division of Bioinformatics, Bose Institute, P-1/12, CIT Scheme VIIM, Kankurgachi, Kolkata, 700 054, India;* [2]*Department of Biochemistry and Biophysics, University of Kalyani, Kalyani, Nadia, 741235, West Bengal, India;* [3]*Department of Biotechnology, Sidho-Kanho-Birsha University, Ranchi-Purulia Road Campus, NearSainik School, Purulia, 723104, West Bengal, India*

ARTICLE HISTORY

Received: January 13, 2023
Revised: April 26, 2023
Accepted: May 11, 2023

DOI:
10.2174/1574893618666230605120640

Abstract: ***Background*****:** COVID-19 is still widespread due to the rapidly mutating disposition of the virus, rendering vaccines and previously elicited antibodies ineffective in many cases. The integral membrane Envelope (E) protein which is 75 amino acid residues long, has also acquired several mutations.

Objective**:** In this work, we have adopted a high-throughput approach incorporating patient gene expression patterns to identify drug repurposing candidates for COVID-19. We have come up with a list of FDA-approved drugs that can not only prevent E protein oligomerization in both its wild type and a mutational state but can also regulate gene targets responsible for inducing COVID symptoms.

Methods**:** We performed an exhaustive analysis of the available gene expression profiles corresponding to a spectrum of COVID patient samples, followed by drug-gene interaction mapping. This revealed a set of drugs that underwent further efficacy tests through *in silico* molecular docking with the wild-type E-protein.

We also built the molecular models of mutant E-protein by considering the important non-synonymous mutations affecting E-protein structure to check the activities of the screened set of drugs against the mutated E-protein. Finally, blind molecular docking simulations were performed to obtain unbiased docking results.

Results**:** Interestingly, this work revealed a set of 8 drugs that have the potential to be effective for a wider spectrum of asymptomatic to severely symptomatic COVID patients.

Conclusion**:** The varied stages of infection and rapid rate of mutation motivated us to search for a set of drugs that can be effective for a wider spectrum of asymptomatic to severely symptomatic COVID patients. Further, the efficiency of these drugs against mutated E-protein increases another level of confidence to fight against this rapidly changing deadly RNA virus and subsequently needs to be validated in clinical settings.

Keywords: COVID-19, drug-gene interaction, envelope protein, mutation, transcriptome, docking.

1. INTRODUCTION

The outbreak of COVID-19 driven by SARS-CoV-2, with estimated cases of 669,357,330 and subsequent fatalities of 6,717,841 worldwide reported to date, is still spreading widely [1]. In this infection, the virus spreads due to droplet or aerosol transmission, *i.e*., inhaling or coming in contact with the respiratory release of the infected individual, and the extent of pathogenicity largely depends on the immune response of the host [2]. Nowadays, in most cases, the patients are asymptomatic or display mild symptoms [3, 4]. However, hospitalization is required for patients with moderate to severe reactions where the most dominant cause of the patient's demise is acute respiratory distress syndrome (ARDS) and pulmonary injury, coupled with co-morbid conditions adding more complications [5, 6].

Our main focus is to identify novel drugs against the Envelop (E) protein, a short, integral membrane protein of 75 amino acid residues of SARS-CoV-2 [7], and our area of interest lies mainly within the region between the 15th to 37th amino acid residues of the Transmembrane Domain (TMD)

*Address correspondence to these authors at the Department of Biochemistry and Biophysics, University of Kalyani, Kalyani, Nadia, 741235, West Bengal, India; E-mail: angshu@klyuniv.ac.in; and Division of Bioinformatics, Bose Institute, P-1/12, CIT Scheme VIIM, Kankurgachi, Kolkata, 700 054, India; E-mail: zhumur@jcbose.ac.in (Zhumur Ghosh)

[#]*All these authors contributed equally to this work. All these authors are to be considered as the first authors of the paper.*

1574-8936/23 $65.00+.00

of the E-protein. This region is involved in the formation of the pentameric viroporin structure, making an ion channel and contributing to viral pathogenicity [8]. Our previous work on COVID-19 [9] focused on repurposing FDA-approved drugs and pharmaceuticals against this E-protein using a machine learning-based approach.

The innate immune response of the infected host rises with the spread of the infection from the upper to the lower respiratory tract when the symptoms of the patient soar [10]. The diverse range of clinical manifestations from asymptomatic to severe disease may be explained by this steady spread. The innate immune response is also responsible for the characteristic 'cytokine storm' of severe cases of COVID-19 [11]. There are also conclusive evidences of the infection not being restricted to the respiratory tract but its dissemination to other organs. The viral load has also been detected in blood and feces [12, 13], and there have been reported cases of damage to the eyes, skin, liver, heart, and kidneys, among other organs [14-18]. The time-point of viral spread has been estimated to take about 4-9 days in the lungs and 8-15 days in the blood [19]. Hence, detailed information on gene expression changes corresponding to such a spectrum of clinical specimens from covid patients can yield significant insights into the disease pathogenesis and detect other possible modes of viral transmission apart from respiratory droplets [20]. All these tempted us to adopt a high-throughput approach incorporating patient gene expression data to refine the drug screening process [21, 22]. The main hypothesis behind this approach is that an effective drug reverses the gene expression signature of the patient [23, 24].

Hence, in this work, we analyzed the available gene expression profiles from (i) blood samples of asymptomatic patients, (ii) swabs and blood samples of hospitalized patients and (iii) lung samples of deceased individuals and elucidated the corresponding drug-gene interactions. Interestingly, gene expression analysis followed by drug-gene interaction mapping revealed a set of drugs that have the potential to be effective for the treatment of a wider spectrum of patients suffering from COVID-19 infections with asymptomatic to severe symptoms.

However, one of the major challenges, which lie ahead in the selection of drugs against COVID-19, is the mutational capability of the SARS-CoV-2. Owing to the rapidly changing predisposition of this RNA virus, mutated new strains are stemming every now and then [25, 26]. Although most mutations are inconsequential, some can change viral characteristics [27]. In such cases, antibodies elicited against earlier virus strains could be rendered less effective by some of these mutations [28]. Thus, the immune response generated due to prior infection or after vaccination can be easily escaped by the mutated strain [29]. The genotyping analysis revealed a higher rate of mutations in several viral proteins like Spike, Nucleocapsid and RNA polymerase [30-32], compared to genes encoding for E-protein [33, 34]. However, E-protein has also undergone mutations. Hence, it became a matter of concern whether the drugs screened against the wild-type variant of E-protein would work as efficiently against the mutated protein. Hence, we considered the important non-synonymous mutations affecting E-protein structure and checked the activities of the screened set of drugs against wild type as well as the mutant E-proteins.

2. MATERIALS AND METHODS

2.1. Transcriptomic Data Analysis

Raw long RNA sequence data of COVID patients from blood, swab and lung, as well as their respective normal counterparts, were downloaded from NCBI GEO [35]. A summary of the input dataset is provided in Table **1.** The lung dataset consists of post-mortem samples of hospitalized deceased patients who received treatments [36]. The Naso-pharyngeal (NP) swab data are categorized into (i) young adults (aged 19-40) and (ii) aged adults (aged above 40) [37]. Blood datasets consist of (i) asymptomatic individuals [38] and (ii) hospitalized patients on day 0 with mild or moderate symptoms, and (iii) hospitalized patients on day 0 with severe symptoms [39]. Altogether, there are six groups that were analyzed separately. The grouping information is provided in Table **2**. Detailed information on the GEO datasets considered in this analysis and selected samples for both patients and healthy controls are provided in Supplementary File **SF1**. For all the groups, read qualities were checked with FastQC (http://wwwbioinformaticsbabrahamacuk/projects/fastqc) followed by trimming of adapters and low-quality reads (quality<35) with Cutadapt3.4 [40]. Trimmed reads were mapped to the Human hg38 reference genome downloaded from GENCODE (https://www.gencodegenes.org/) using Hisat2 2.1.0 [41]. Sorting and indexing of BAM files were performed using SAMtools0.1.19 [42], followed by differential analysis against their respective normal counterparts with Ballgown [41]. The stattest function of Ballgown takes care of several concerns, including the batch effect that may arise due to multiple data sources. Subsequently, a significant set of differentially regulated genes was screened with an FC cut-off of 1.5 and q-value ≤ 0.05.

2.2. Drug-gene Interaction and Selection of a Potential Set of Drugs

We categorized all the six patient groups into three categories (Table **2**)

- Asymptomatic condition, *i.e.*, infected patients showing no symptoms of the disease. Our asymptomatic patient data from blood samples belong to this category.
- Hospitalized untreated condition – where patients displayed symptoms and needed to be hospitalized but received no medication at the time of sample collection. This includes (a) Swab young patient dataset, (b) Swab old patient datasets, (c) mild to moderate cases, and (d) severe cases. The common set of DEGs with the same mode of regulation in all these datasets was considered for further analysis.

Table 1. Summary of input Long RNA sequence dataset.

Sample Type	Tissue Type	GEO Accession ID	No. of Samples
CONTROL	Swab	GSE97668 GSE129959	5 4
	PBMC	GSE173670	11
	Whole blood	GSE167000	15
	Lungs	GSE158752	17
COVID	Swab	GSE172274	14
	PBMC	GSE159678	13
	Whole blood	GSE166424	15
	Lungs	GSE150316	24

Table 2. Experimental grouping of patient-specific input datasets.

Category	Patient Datasets	No. of Patient Samples	Control Datasets	No. of Healthy Controls
Asymptomatic	Whole Blood	15	COVID -ve Whole Blood	5
Hospitalized (Untreated cases, sample collected at Day 0 of admission)	Naso-pharyngeal Swab young (Aged 19-40 years)	6	Healthy Control Nasal Epithelial cells	9
	Naso-pharyngeal Swab old (Aged above 40 years)	8		9
	PBMC (Mild/Moderate case)	4	Healthy control CD14+ monocytes	11
	PBMC (Severe case)	9		11
Deceased (post-mortem samples of treated patients)	Lung Formalin fixed paraffin embedded tissue	24	Healthy Control Bronchial Epithelial Cells	17

- Treated deceased patients – Hospitalized patients who died despite medication. This includes the lung post-mortem dataset.

Target drugs corresponding to the DEGs were screened out for each of the patient datasets. For each of the 3 Differentially Expressed (DE) groups, the list of their target drugs was retrieved from the following databases:

DGIDb [43], a database of druggable genomes to interpret drug-gene interactions with probable therapeutic benefits and DrugBank [44], a free online database holding information on drugs and their targets, were used.

Next, the target drugs, common across all three categories of patients, were filtered.

These drugs were screened further based on the following criteria:

1. FDA status– only approved drugs and nutraceuticals were considered
2. Side effects – drugs with reported severe side effects were discarded
3. Chemical structure of the drug molecules – At this stage, our main target was to select those drugs which would be able to disrupt the formation of the active pentameric E-protein by interacting with it in such a way that the viable E-protein pentamer was not generated. In this context, our main area of interest is the membrane-spanning region of the E-protein located between the amino acid residues 15 to 37.

   This region, containing 3 aromatic Phe residues (at positions 20, 23 and 26), participates in E-protein oligomerization to make a viable and active pentameric structure necessary for its host cell activity. In order to stop the oligomerization process, the optimum drug

candidate must have aromatic rings to block the said region of E-protein by stacking interactions of the aromatic rings in the ligands with the side chains of the Phe residues. Thus, we screened those sets of drugs which have at least 2 aromatic rings in them.

4. Pathophysiological properties of the drug molecules – So far, the pathophysiological properties are concerned, we omitted the set of drugs which have the potential to be used as chemotherapeutic agents [45], such as antipsychotic drugs [46], as they have extreme side effects. Besides, we excluded non-edible drugs, *viz*., topical drugs.

Overall, we obtained the final set of screened FDA-approved drugs based on their beneficial chemical structure, with minimal side effects. These sets of drugs were considered for subsequent molecular dockings. The detailed workflow is summarized in Fig. (**1**).

2.3. Mutational Analysis of SARS-CoV-2 E-protein

In order to check for the sequence conservation of E-protein across the globe, amino acid sequences of the E-

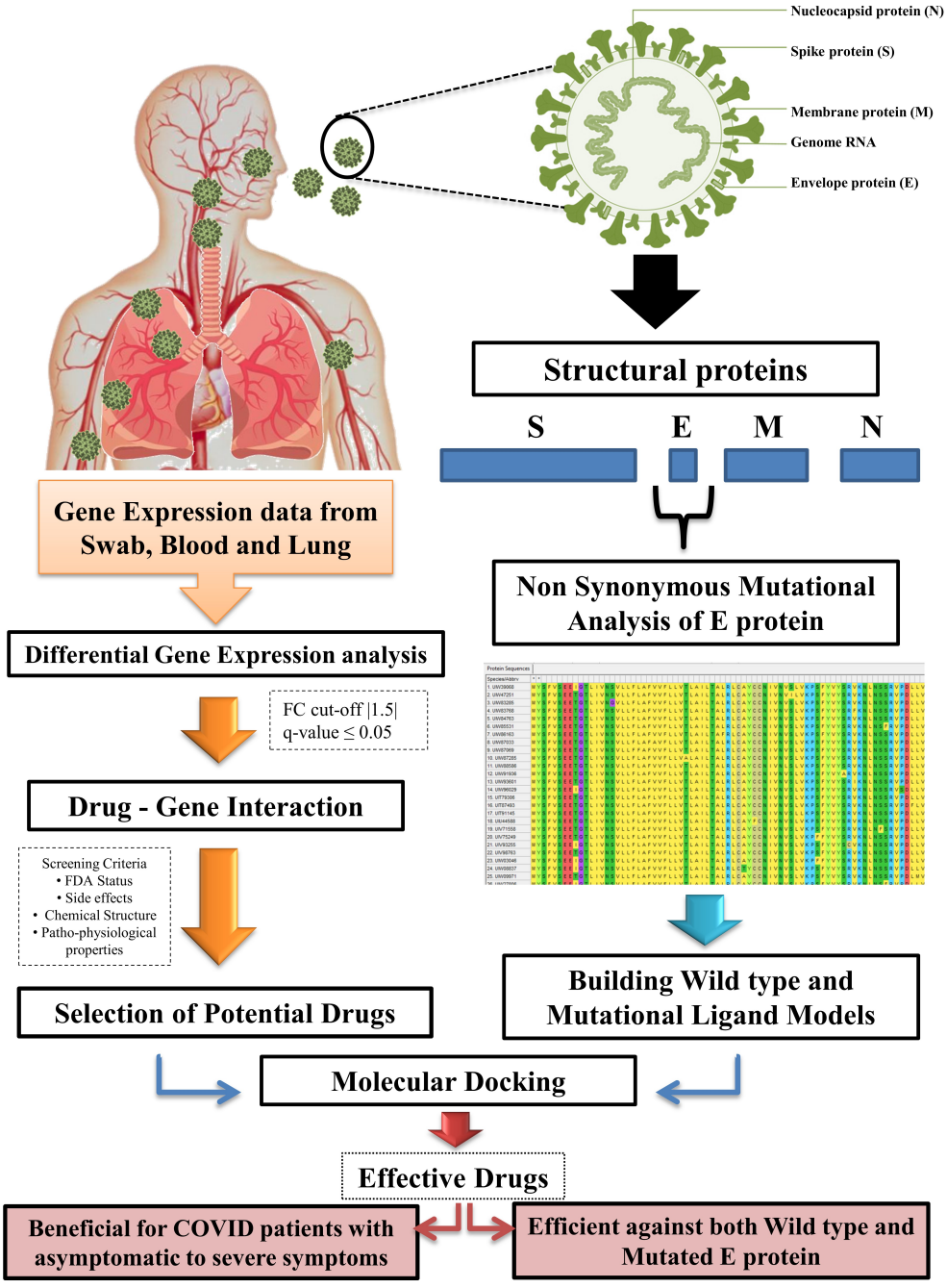


Fig. (1). Workflow summarizing the entire work. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

protein were downloaded from NCBI SARS-CoV-2 resources submitted till 18th January 2022. Custom scripts were used to check for sequence conservation and mutation at the amino acid level against their wild-type E-protein sequence (accession no. YP_009724392.1) obtained from Wuhan, China. We also derived all the unique amino acid substitutions at the N-terminal domain (NTD), Transmembrane domain (TMD), and C-terminal domain (CTD) separately.

2.4. Molecular Docking

2.4.1. Building the Structures of the Wild-type and Mutant E-protein

We have used the same model of the wild-type E-protein (which was built in our previous work), [9] where the amino acid residues spanned the region between 8 to 65, encompassing the transmembrane hydrophobic region of the E-protein, and were involved in its oligomerization to make a viable and active pentameric structure necessary for exerting its activity in the host cell. We considered the mutations within the aforementioned region only, as they will only affect the function of the E-protein to form the active pentameric structure.

We developed an in-house Python pipeline for analyzing sequence features in order to find out the mutations present within the E-protein. The mutants of the E-protein were built using this pre-optimized structure of E-protein in Discovery Studio 2.5 (DS2.5) using the 'Build Mutant' protocol. The mutated structures were subsequently energy minimized in DS2.5 using the steepest descent algorithm, followed by the conjugate gradient algorithm until the RMS gradient reached 0.001kcal / mole. The energy minimization was done using the CHARMm force field (inbuilt in DS2.5). We used GBSW implicit solvent model for energy minimization [47].

2.4.2. Building of the Structures of the Ligands

We used the SMILES information of the ligands to prepare their 3D structures and optimized the structures, maintaining the physiological pH (7.35-7.45) using DS2.5.

2.4.3. Virtual Screening of Ligands

We made a library of ligands, which was used to perform virtual screening with the help of AutoDock Vina [48]. We performed the virtual screening of the selected set of ligands with wild-type E-protein and all the unique mutants of E-protein, which we have built.

3. RESULTS

3.1. Drug-gene Pairs Act as Potential Players During Different Stages of the Disease Progression

3.1.1. Gene Expression Data Analysis Corresponding to the Patient's Swab, Blood, and Lung Samples

The pathophysiology of COVID infection ranges from asymptomatic condition to severe reaction and death. The host response to SARS-CoV-2 infection is different among asymptomatic, mild-moderate, to severe cases. Hence it is important to look into the gene expression changes corresponding to the clinical samples from covid patients, which are expected to bear the reflection of the pathophysiology of such infection.

We have analyzed the bulk transcriptomic data from NCBI GEO of individuals with asymptomatic (whole blood), untreated mild to moderate (NP Swab and PBMC), and severe (PBMC) symptoms as well as deceased patients (Lung) post-treatment. The control dataset was of healthy volunteers or patients with negative COVID outcomes. It is important to mention that among all the samples, lung samples corresponded to patients who had received medication (which includes Azithromycin, Atorvastatin, Cefepime, Ceftriaxone, Hydroxychloroquine and Vancomycin) [36]. Also, all these patients suffered from co-morbidity. Hence, the gene expression data of treated samples should reflect a different story as compared to those of untreated infected lung samples. Although it is possible that some patients benefited from the treatment, they still did not survive because of the severity of their condition, and the treatment was not effective in all cases. Hence, the expression of certain crucial genes could not be reversed despite medication, and these genes would be interesting to work on. The Whole Blood dataset was from asymptomatic individuals[38], while Swab and PBMC samples were from Day 0 hospitalized patients with mild to severe symptoms [37, 39]. Hence, a similar pattern of gene regulation was not expected in all these clinical conditions. We grouped our DE datasets based on (i) Asymptomatic (WB), (ii) Hospitalized (Common PBMC and Swab DE genes), and (iii) Deceased (Lung) condition (as given in Table **2**). The number of DE genes for the 6 datasets are provided in Table **3**.

Table 3. Differentially regulated genes across various COVID patient-specific samples.

Grouping	Datasets	UP	DOWN
ASYMPTOMATIC	Whole Blood	3443	2013
HOSPITALIZED	Swab Old	2678	1639
	Swab Young	2040	963
	PBMC Mild/Moderate	1173	2752
DECEASED	PBMC Severe	1350	2407

3.1.2. The Potential Set of Selected Drugs Targets Effective in All Stages of Disease Progression

There remains a need to identify potential drugs which could benefit patients regardless of the stage of the disease progression as well as the degree of severity of the disease symptoms. Hence, it is necessary to select those drugs that would correspond to the genes across all the stages of the

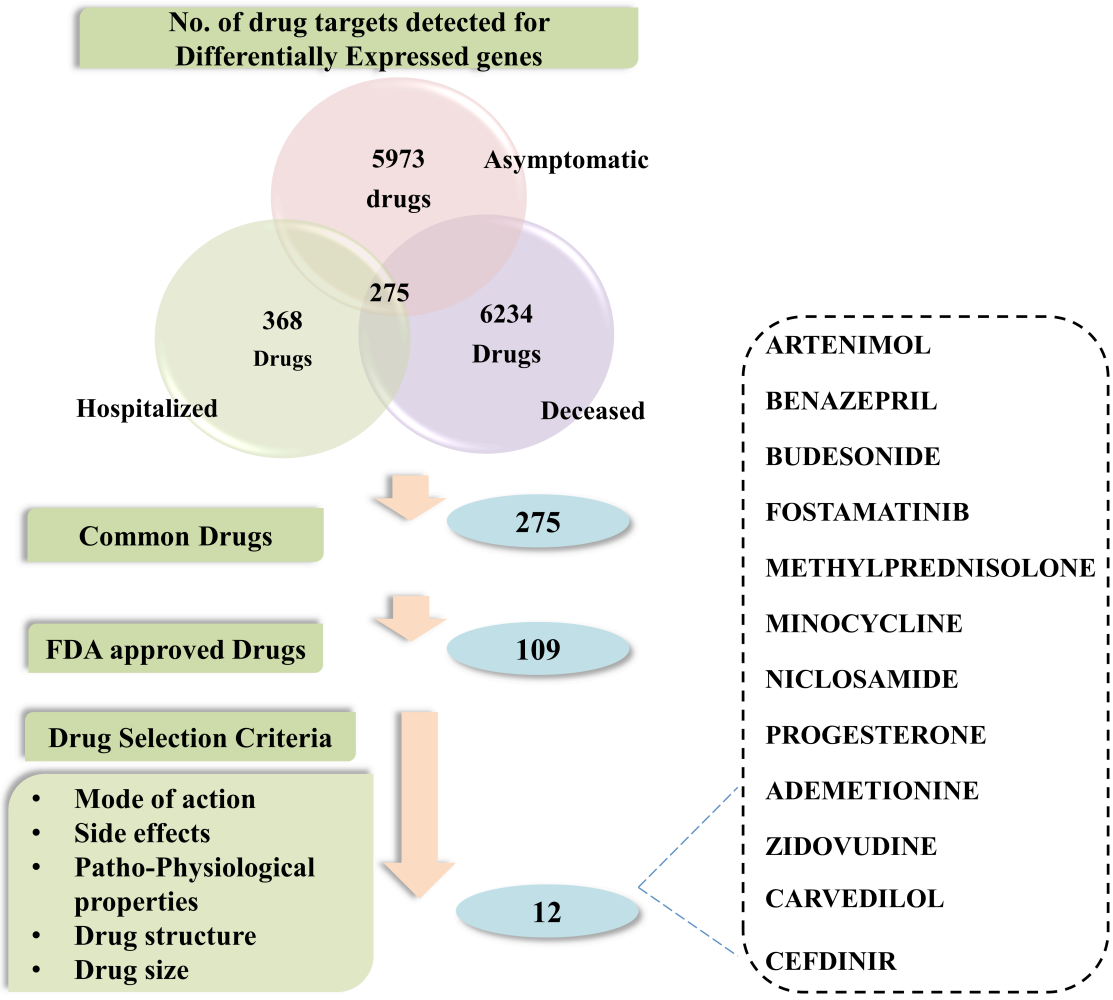


Fig. (2). Flow chart describing screening criteria of drugs. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

disease. Our analysis detected 5973, 368 and 6234 drugs from the three groups, respectively, with 275 common drugs among them.

Among them, we considered only FDA-approved drugs and nutraceuticals, which reduced the number of drugs to 109, followed by other filtration criteria like mode of action, severity of side effects, drug structure, and size for maximum binding efficiency. Lastly, the functions of their corresponding genes were checked in the literature for possible involvement in COVID pathophysiology. A detailed flow chart for the entire screening process is provided in Fig. (**2**). Finally, a selected list of 12 drugs (shown in Fig. **2**) was considered for molecular docking.

3.2. Sequence Conservation and Mutation of SARS-CoV-2 E-protein at Amino Acid Level

In light of the rapidly changing nature of the virus in the form of mutation, it is a matter of concern that the selected set of drugs would bind as efficiently to any mutated form of the same protein. Hence, it is important to find out the mutations in E-protein in the form of amino acid substitution and confirm whether these substitutions would largely impact the structure of the E-protein and interfere with the activity of the selected drugs.

A comparison of the amino acid sequences of the structural proteins of SARS-CoV-2 strains collected from various geographical locations revealed around 90% conservation of E-protein sequences. This is the highest level of conservation among any structural proteins, followed by M (50%), while N and S proteins account for less than 5% of sequence conservations in them. There were a total of 4,91,600 amino acid sequences of E-proteins in NCBI till 18th January 2022. However, after the removal of partial sequences (41,740 in number) from it, 280 unique amino acid sequences corresponding to E-protein could be retrieved for our analysis. We compared the 280 unique amino acid sequences with the reference sequence of SARS-CoV-2 (bearing the ID: YP_009724392.1), which is the one obtained from Wuhan, China. A summary of sequence conservation and mutations is provided in Table **4**.

Table 4. Summary of sequence conservation and mutations of Envelope (E) protein.

No of Envelope (E) protein Sequences	491600
No. of full-length Sequences	449860
No. of WT protein sequences	403572
No. of Mutations	46288
Percentage of mutation	10.28
No of Unique mutations	280
No. of mutations at 15-37th position (core region of E protein)	2403
Percentage of mutation within the core region	0.5
Unique no. of mutations within the core region	59

On deeper probing, we detected 25, 94, and 129 unique mutations in the N-terminal, Transmembrane and C-terminal domains of the E-protein, respectively. In our previous work, we targeted the region of E-protein spanning the amino acid residues 15 to 37 since this region forms the scaffold to build the active pentameric structure of the E-protein necessary for the functionality of the virus. Here, we used the same model of the E-protein (which was built in our previous work [9]), where the amino acid residues spanned the region between 8 to 65. We eliminated those mutations outside the aforementioned region, as they will not affect the function of the E-protein to form the active pentameric structure.

Our analysis revealed that less than 1% of the global viral strains bear a mutation in this zone of interest. We concentrated on the mutations present in this region with a mutation frequency of 1% or more based on previous reports [49]. Details of the mutations present within the entire E-protein of SARS-CoV-2 from different parts of the world, along with their mutational frequencies (≥1%), are presented in Table **5**. From Table **5**, it was observed that there were two amino acid residues, Leu21 and Thr30, which lie within the zone of interest with the desired mutational frequencies for our analysis. They were (i) L21P with a mutational frequency of 3.78%, (ii) L21V with a mutational frequency of 2.10%, and (iii) T30I with a mutational frequency of 2.52%. We finally selected 11 unique mutations present within our region of interest for further analysis, as presented in Table **6**.

3.3. Molecular Docking-based Virtual Screening

We calculated the binding interactions of the 12 ligands with the wild-type and mutant E-proteins and the energy profile of the interactions. Based on the interaction score (as shown in Table **7**) and zone of binding, 8 FDA-approved drugs were finally selected. The detailed binding interactions of all the ligands with the wild-type E-protein are provided in Supplementary File **SF2**.

Table 5. Details of the mutations present in the SARS-CoV-2 E-protein from different parts of the world, along with their mutational frequencies (>1%).

Residue Number	Mutation	Mutation Occurrence	Mutation Frequency (%)
9	9:Thr->Ile	6	2.52%
21	**21:Leu->Phe**	**9**	**3.78%**
21	**21:Leu->Val**	**5**	**2.10%**
30	**30:Thr->Ile**	**6**	**2.52%**
50	50:Ser->Gly	3	1.26%
55	55:Ser->Phe	5	2.10%
62	62:Val->Phe	5	2.10%
68	68:Ser->Phe	6	2.52%
71	71:Pro->Leu	6	2.52%
71	71:Pro->Ser	5	2.10%
72	72:Asp->Gly	3	1.26%
73	73:Leu->Phe	9	3.78%

Note: Different residues from the zone of interest (spanning the amino acid residues 15 to 37) are marked in bold.

Table 6. Details of the selected mutations spanning amino acids residues of E-protein from 8 to 65.

Sequence_ID	Mutations
QWU54674.1	L21F
QWO17723.1	L21F, V62F
QWF08470.1	L21F, V58F
QVL74519.1	L21F, T9I
QUX04199.1	L21F, S50G
QUP12063.1	L21F, V14I
QWS07306.1	T30I
QUQ61256.1	T30I, T9I
QTQ59159.1	T30I, S55F
QWU56593.1	L21V
QVK81428.1	L21V, T9I

Artenimol: This is the only ligand that could target the Asn15 of the E-protein. The presence of a hydroxyl group in the ligand could favour the formation of polar interactions with the side chain of Asn15. The structure of the ligand is such that it could fit into the cavity of the E-protein, where it is able to interact with other polar and non-polar amino acids like Glu8, Thr11, Leu12, Val14 *etc*.

Fostamatinib: This is the only ligand that could target the Val25 of the E-protein. The presence of a long hydrophobic carbon backbone in the ligand could favour the formation of non-polar binding interactions with the side chain of Val25.

Table 7. Binding scores of the selected 8 drugs with the wild type and the mutant E-protein models.

Ligand	Wild Type	L21F	L21F S50G	L21F V14I	L21F V58F	L21F V62F	L21F T9I	L21V	L21V T9I	T30I	T30I S55F	T30I T9I
ARTENIMOL	-5.3	-5.3	-5.3	-5.3	-5.5	-5.3	-5.3	-5.3	-5.3	-5.3	-5.3	-5.3
BENAZEPRIL	-5.4	-5.4	-5.5	-4.9	-5.2	-5.3	-5.4	-5.5	-4.9	-4.9	-5.4	-5.1
BUDESONIDE	-5.6	-5.4	-5.6	-5.6	-5.4	-5.6	-5.6	-5.2	-5.4	-5.6	-5.6	-5.6
FOSTAMATINIB	-5.5	-5.6	-5.6	-5.2	-6.3	-5.4	-5.3	-5.8	-5.2	-5.2	-5.6	-5.3
METHYLPREDNISOLONE	-5.8	-5.7	-5.8	-5.8	-5.7	-5.9	-5.8	-5.8	-5.8	-5.8	-5.7	-5.6
MINOCYCLINE	-5.8	-5.4	-5.8	-5.8	-5.8	-5.8	-5.8	-5.7	-5.7	-5.5	-5.4	-5.5
NICLOSAMIDE	-5.2	-5.2	-5.1	-5.2	-5.3	-5.2	-5.1	-5.2	-5	-5	-5	-5
PROGESTERONE	-6.2	-6.2	-6.2	-6	-6.1	-6.2	-6.2	-6.3	-6.2	-6	-6.1	-6
ADEMETIONINE	-4.7	-4.8	-4.2	-4.1	-4.2	-4.8	-4.2	-3.9	-4.7	-4.6	-4.3	-4.3
ZIDOVUDINE	-4.3	-4.3	-4.3	-4.5	-4.9	-4.5	-4.5	-4.3	-4.3	-4.2	-4.3	-4.5
CARVEDILOL	-5.3	-4.9	-5.1	-5.5	-4.8	-5.2	-4.9	-5.2	-5.6	-5.5	-5.2	-5.5
CEFDINIR	-5	-4.9	-5	-4.9	-5.1	-5	-5	-5	-4.9	-5	-4.9	-5

The presence of the long hydrophobic carbon skeleton in the molecule would make it a good candidate to bind with the hydrophobic core of the E-protein, like Leu18, Leu19, Leu21, Phe23 *etc*.

Progesterone: The structure of Progesterone is appropriate to fit inside a hydrophobic cavity. There are a couple of cyclic rings in Progesterone, which allow it to be stacked onto the hydrophobic amino acid residues present in the core of the E-protein. This would be conducive for the ligand to fit into the hydrophobic cavity of the E-protein. The presence of polar carbonyl groups would help the ligand to form suitable polar interactions with the side chain of Thr 30 as well as the main chains of the other amino acid residues present in the core of the E-protein. Such polar binding interactions would help to neutralize the charges associated with the polar carbonyl groups present in the ligand. This would make the ligand a good candidate to bind with the E-protein. The ligand was found to have strong binding interactions with the mutants like Thr30Ile. The conversion of polar Thr to non-polar Ile would create a more hydrophobic environment in the core of the E-protein. This would allow the hydrophobic backbone of the ligand to bind strongly.

Benazepril: This ligand has aromatic rings in its backbone as well as polar groups. Therefore, this ligand is able to make strong interactions with aromatic amino acid residues like Phe20, Phe23, and Phe26 in the core of the E-protein. Due to this, the ligand has a good binding affinity towards the mutant Thr30Ile, as the conversion of polar Thr to non-polar Ile makes the core of the E-protein conducive to binding with the non-polar backbone of the ligand.

Budesonide, Methylprednisolone, and Minocycline: These ligands have hydrophobic cores bound with many polar groups. The ligands were found to be interacting with non-polar amino acid residues present at the core of E-protein. The polar groups present in the ligand were found to be interacting with the polar peptide linkages.

Niclosamide: This ligand has many charged groups like Cl^-, $-NO_2^-$,–CONH bound to aromatic rings. It has binding interactions with aromatic amino acid residues present at the core of the E-protein.

4. DISCUSSION

The basic aim of our work was to disrupt the formation of the active pentameric E-protein by suitable ligands, and the selected ligands would be able to do so by interacting with the E-protein in such a way that the viable E-protein pentamer is not generated. The interaction between the ligands and the E-protein occurs mainly through hydrophobic stacking. Such interactions are mediated through the hydrophobic backbones of the ligands. The ligands tend to organize themselves properly onto the part of the E-protein which is responsible for creating the hydrophobic scaffolding necessary to create the active pentameric structure of the E-protein.

We tested the efficiency of the FDA-approved drugs against mutated E-protein and checked the corresponding drug-gene interactions, which revealed a set of 8 drugs that can be effective for a wider spectrum of COVID patients with asymptomatic to severe symptoms. This set of drugs had high interaction score with the wild type as well as mutant E-proteins (Table **7**). Drug-gene target details for these selected drugs are summarized in Table **8**. Further, Fig. (**3**) shows the important modes of action of these drugs during different stages of the viral life cycle. Detailed information regarding the COVID-specific functions of these targeted genes is provided in Supplementary document **S1**.

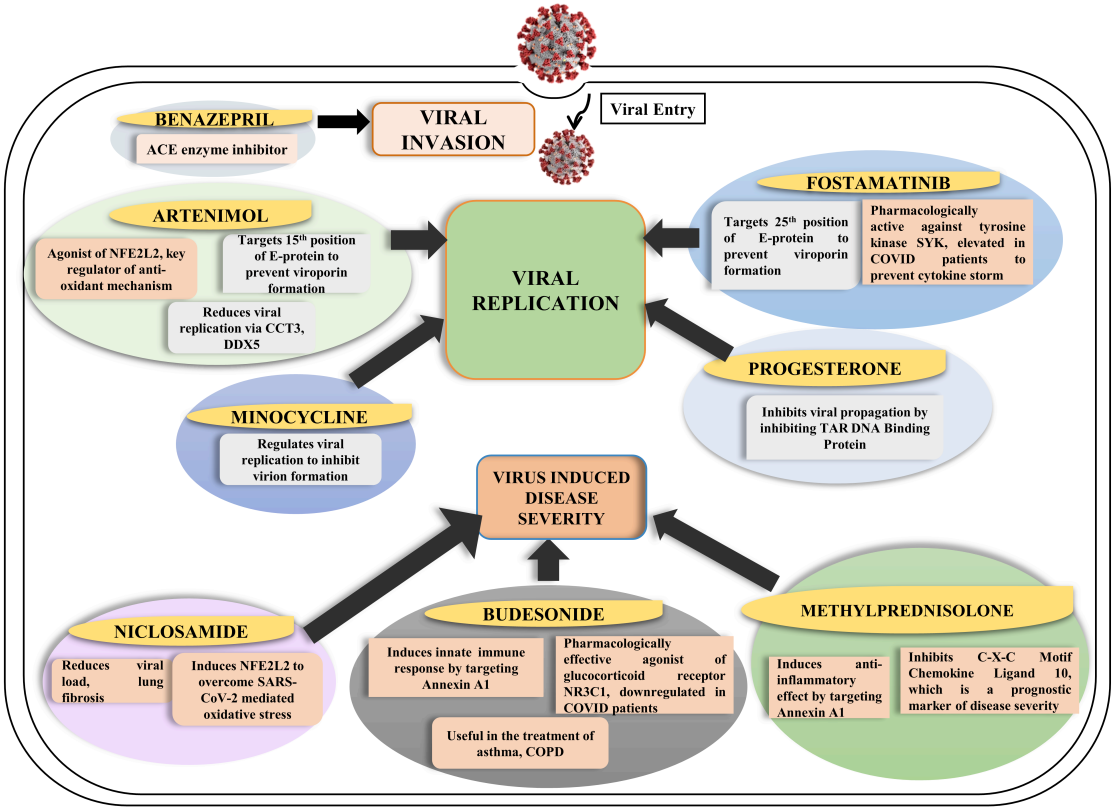


Fig. (3). Schematic diagram depicting the important modes of action of the selected set of drugs during the viral lifecycle. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Table 8. Drug-gene target interaction for the selectedset of drugs.

Drug	Drug Function	Condition	Gene Status	Gene Name	Gene Functions Relevant to COVID Scenario	Drug Action on Genes
Artenimol	Anti-malarial drug. Known for its anti-inflammatory activity and is considered a potential repurposed COVID-19 drug. It not only prevents cytokine storm by reducing the level of inflammatory cytokines but also interferes with the viral replication cycle post-entry.	Asymptomatic	UP	ACTG1	Actin ACTG1 regulates Cell motility. High serum levels reported in COVID-19 patients.	Ligand
				ANAX2	A pro-inflammatory receptor and Ca^{2+}-dependent lipid raft binding protein. Facilitates SARS Cov entry, cytokine storm, and thrombosis.	
				CCT3	A molecular chaperone. Higher expression was observed in the transfection group of SARS-CoV-2 N-protein.	
				DDX5	This Dead Box helix plays a dual role by augmenting viral RNA transcription and suppressing the innate immune response.	
				RPL10	Enhance viral translation.	
		Hospitalized	UP	MAP4	Cleavable target by the poliovirus 3CLpro proteinase in HeLa cells.	
				MYH9	ACE2 co-receptor that facilitates SARS-CoV-2 infection by interacting with the S2 subunit and the S1-NTD subunit.	
		Deceased	UP	HSPB1	Stress protein that induces the expression of pro-inflammatory cytokines like IL6, IL8, and COX2.	
				VIM	An intermediate filament protein. A crucial factor for Spike protein-mediated viral entry.	
			DOWN	NFE2L2	Key regulator of anti-oxidant mechanisms activated by Artenimol for transcriptional Activity.	Agonist

(Table 8) Contd…

Drug	Drug Function	Condition	Gene Status	Gene Name	Gene Functions Relevant to COVID Scenario	Drug Action on Genes
Benazepril	A prodrug used in the treatment of hypertension along with congestive heart and chronic kidney failure. In the liver, Benazepril is converted into its active form, benazeprilat, a potent ACE enzyme inhibitor.	Asymptomatic	UP	ARDB2	Beta-2-adrenergic receptors were observed to be significantly elevated, with polymorphisms in the gene contributing to asthma severity.	Regulator
				MTHFR	MTHFR polymorphism has been linked to COVID-related mortality, and MTHFR-synthesized homocysteine can lead to thrombosis and eventual D-dimer production.	
		Hospitalized Deceased	DOWN	PRCP	Reduced PRCP levels resulted in a prothrombic state and elevated production of ROS.	
Budenoside	A glucocorticoid used in the treatment of allergies, asthma, hay fever and COPD, among others.	Asymptomatic Deceased	DOWN	NR3C1	Glucocorticoid receptor NR3C1 expression is found to be upregulated in milder cases but reduced in severely diseased patients of COVID-19.	Agonist with confirmed pharmacological action
		Hospitalized Deceased	DOWN	ANXA1	Encode a membrane-localized protein that inhibits phospholipase A2 and exerts anti-inflammatory action, and plays an effective role in glucocorticoid-mediated innate immune response.	Agonist
Fostamatinib	A spleen tyrosine kinase inhibitor used in the treatment of Immune Thrombocytopenic Purpura and Rheumatoid Arthritis. Recently, Fostamatinib has gained importance as a repurposed COVID-19 drug for its potential to combat acute respiratory distress syndrome (ARDS)	Asymptomatic	UP	ITK1	Elevated Tyrosine Kinase ITK is associated with SARS-related lymphopenia and pro-inflammatory cytokine production.	Inhibitor
				JAK1, JAK2, JAK3	Janus Kinase1,2 and 3 of the protein tyrosine kinase family are involved in pro-inflammatory cytokine production, and JAK inhibitors are suggested to be promising candidates in COVID treatment.	
				MAP3K2, MTOR	Acts as a hub of inflammatory signalling, has been suggested as a beneficial approach in COVID treatment.	
				PI4KB	Plays an important role Spike protein-mediated viral entry during or before viral fusion.	
				SYK	A non-receptor type protein kinase that can modulate fc receptor signalling and immune complex-mediated inflammation. Also found to have high expression levels in COVID-19 patients.	Inhibitor with confirmed pharmacological action
		Hospitalized	UP	NEK4	Associated with ARDS-related Endothelial barrier dysfunction (EBD), inflammation and sepsis.	Inhibitor
				SYK	As discussed before.	
				JAK2, JAK3	As discussed before.	
		Deceased	UP	BTK	Involved in multiple signalling pathways and induce the production of pro-inflammatory cytokines.	Inhibitor
				CTSL	A lysosomal cysteine protease is essential for the functional cleavage of S protein and host entry.	
				FTL1	Higher expression of the soluble form of VEGFR family gene FLT1 was reported in COVID19 associated pneumonia compared to healthy controls.	
				MYLK	Inhibition of MYLK genes resulted in a dose-dependent reduction of SARS-CoV-2 titers and has been linked to ARDS.	

(Table 8) Contd…

Drug	Drug Function	Condition	Gene Status	Gene Name	Gene Functions Relevant to COVID Scenario	Drug Action on Genes
Methylprednisolone	A glucocorticoid prescribed for its anti-inflammatory effects. There have been reports of methylprednisolone in the successful treatment of COVID19 associated pneumonia in a patient with long-term immunosuppression.	Asymptomatic	UP	IL2RA	Immune activation markers also reported in COVID-related complications.	Inhibitor
		Hospitalized	UP	IL2RG	Immune activation markers also reported in COVID-related complications.	Inhibitor
			DOWN	ANXA1	As discussed before.	Agonist with confirmed pharmacological action
		Deceased	UP	CXCL10	Elevated serum levels of CXCL10 have been reported to be positively correlated with disease severity and increased mortality risk.	Inhibitor
				MYC	MYC regulates adaptive immune response by inducing T cells generation upon viral entry.	
			DOWN	ANXA1, NR3C1	As discussed before.	Agonist
Minocycline	A second-generation antibiotic with broad-spectrum activity.	Asymptomatic	UP	IL1B	IL1B has been reported to induce cytokine storm in COVID patients.	Inhibitor
		Asymptomatic Hospitalized	UP	CASP3	ORF6 of SARS-CoV-2 induces apoptosis *via* Caspase-3 mediated pathways, and Caspase inhibitor reportedly blocked ORF6 induced apoptosis.	Negative modulator
		Desceased	UP	VEGFA	A potential vascular permeability factor, VEGF induces vascular leakiness leading to inflammation, plasma extravasation, and pulmonary edema. Hypoxia-induced elevated VEGF expression *via* HIF1 in COVID-19 disease.	Inhibitor
Niclosamide	An anti-helminthic drug used to treat tapeworm infections.	Asymptomatic	UP	MTOR	As discussed earlier.	Inhibitor
		Asymptomatic Hospitalized	UP	STAT3	STAT-3 is known to induce an unbalanced antiviral immune response *via* Th17-, Th1-, Treg-, and B cell-mediated pathways. It can cause the polarization of M2 macrophage, cytokine storm, thrombosis (along with PAI-1) and lung fibrosis.	Inhibitor
		Asymptomatic Deceased	UP	HIF1A	HIF1A has reported dysregulated expression in COVID patients with a pro-inflammatory response and associated with high mortality in elderly patients.	Inhibitor
		Deceased	DOWN	NFE2L2	As discussed earlier.	Agonist
Progesterone	An endogenous steroid with multiple functions involving contraception, ovulation, pregnancy support and uterine bleeding control, among others.	Asymptomatic	UP	TARDBP	An RNA binding protein that holds the potential to facilitate the translation of viral proteins and effective viral propagation. C241T, a possible 5'UTR mutation derived from sequence analysis, creates a TARDBP1 binding site in the 5'UTR of SARS-CoV.	Inhibitor
		Hospitalized	DOWN	NME2	Master suppressor of metastasis.	Agonist
		Deceased	DOWN	NFE2L2	AS discussed before.	Agonist
				NR3C1	As discussed before.	
				SLC2A1 (GLUT1)	Major glucose transporter and member of the solute carrier family have reports of being down-regulated in deceased COVID-19 patients with co-morbidity but not in mild or severely affected patients who survived, hinting at altered pH regulation and cellular ion handling.	
				PROS1	A cofactor for anticoagulant protease can activate the TAM family of receptor tyrosine kinases (RTKs).	

CONCLUSION

We analysed the gene expression dataset of COVID-19 patients where the samples were taken from swabs, blood and lungs, and the condition of the patients was asymptomatic, untreated, hospitalized, and even treated deceased one. Further, we checked drug-gene interaction to see the effect of these drugs at different stages of the infection. Interestingly, the drug-gene interaction search for the set of genes across the wide spectrum of clinical samples(with respect to controls) taken from COVID-19 patients (specific to different stages of infection) revealed a novel set of drugs that can be effective in terms of their actions.

Among these, Minocycline, has already been shown to interact with important genes responsible for COVID-19 pathogenicity [50]. The TMD region of the E protein plays a very important role in forming the pentameric ion channel capabilities regulating the virulence of SARS-CoV-2. Mutation within this region spanning between 15 and 25 residues of E protein, disrupts homopentameric structure formation [51]. Interestingly, 2 of the drugs, Artenimol and Fostamatinib, were found to be interacting with the amino acid residues present at the 15^{th} and 25^{th} positions of E-protein and are expected to block the virulent activity of the virus. One of the drugs, Progesterone which is a predominantly female hormone, has achieved the highest docking score against the E protein and is very much hydrophobic in nature. It has been reported that COVID-19 affects more male individuals as compared to females [52]. The presence of high levels of Progesterone hormone in females might be one of the reasons for this. Niclosamide, with its hydrophobic nature, remains tightly coupled at the TMD domain of E protein, which goes well with that reported by Baitha *et al*. [53] regarding the utilization of Niclosamide in treating COVID-19 patients.

In order to make our analysis robust, we have considered RNAseq data from a wide spectrum of patient samples which include swab, blood and lungs, where the condition of the patients vary between asymptomatic, untreated hospitalized and even treated deceased one. The versatility of this wide range of samples brings in both advantages and limitations within our analysis.

As per the advantage is concerned, the set of differentially expressed genes obtained from various patients makes our dataset robust, which allows us to identify the gene set associated with COVID-19 pathogenicity across diverse samples. On the contrary, we have not been able to consider the effect of age group, sex, ethnicity or past clinical history of the patients in the analysis due to the scarcity of data with such information. The availability of patient-specific gene expression datasets with such detailed information will help to strengthen our findings in the future.

Overall, the selection of drugs based on gene expression changes as well as docking score strengthens our prediction and also provides a magnified view of the effect of these drugs based on disease progression. Further, the efficiency of these drugs against the mutated E-protein increases another level of confidence to fight against this rapidly changing deadly RNA virus. It is, henceforth, important to validate the action of these drugs in clinical settings so as to keep ourselves prepared to fight against the deadly variants of COVID-19.

AUTHORS' CONTRIBUTIONS

BG, TD, GD, and NC contributed to the conception and design, collection and/or assembly of data, data analysis and interpretation, and manuscript writing, AB and ZG contributed to the conception and design, data analysis and interpretation; they also assisted in drafting the manuscript and revising it critically for important intellectual content. All authors read and approved the final manuscript.

LIST OF ABBREVIATIONS

ARDS	=	Acute Respiratory Distress Syndrome
DE	=	Differentially Expressed
TMD	=	Transmembrane Domain

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used for studies that are basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

Not applicable.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

We are grateful to the Council of Scientific and Industrial Research (CSIR), the Department of Science and Technology (DST), and the Indian Council of Medical Research (ICMR) for financial support.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Covid-19 coronavirus pandemic. 2022. Available from: Worldometers.info/coronavirus

[2] Wang CC, Prather KA, Sznitman J, *et al.* Airborne transmission of respiratory viruses. Science 2021; 373(6558): eabd9149. http://dx.doi.org/10.1126/science.abd9149 PMID: 34446582

[3] Fan Y, Li X, Zhang L, Wan S, Zhang L, Zhou F. SARS-CoV-2 Omicron variant: recent progress and future perspectives. Signal Transduct Target Ther 2022; 7(1): 141. http://dx.doi.org/10.1038/s41392-022-00997-x PMID: 35484110

[4] Shao W, Zhang W, Fang X, Yu D, Wang X. Challenges of SARS-CoV-2 Omicron Variant and appropriate countermeasures. J Microbiol Immunol Infect 2022; 55(3): 387-94. http://dx.doi.org/10.1016/j.jmii.2022.03.007 PMID: 35501267

[5] Tzotzos SJ, Fischer B, Fischer H, Zeitlinger M. Incidence of ARDS and outcomes in hospitalized patients with COVID-19: a global literature survey. Crit Care 2020; 24(1): 516. http://dx.doi.org/10.1186/s13054-020-03240-7 PMID: 32825837

[6] Liu H, Chen S, Liu M, Nie H, Lu H. Comorbid chronic diseases are strongly correlated with disease severity among COVID-19 patients: A systematic review and meta-analysis. Aging Dis 2020; 11(3): 668-78. http://dx.doi.org/10.14336/AD.2020.0502 PMID: 32489711

[7] Kuo L, Hurst KR, Masters PS. Exceptional flexibility in the sequence requirements for coronavirus small envelope protein function. J Virol 2007; 81(5): 2249-62. http://dx.doi.org/10.1128/JVI.01577-06 PMID: 17182690

[8] Nieto-Torres JL, DeDiego ML, Verdiá-Báguena C, *et al.* Severe acute respiratory syndrome coronavirus envelope protein ion channel activity promotes virus fitness and pathogenesis. PLoS Pathog 2014; 10(5)e1004077 http://dx.doi.org/10.1371/journal.ppat.1004077 PMID: 24788150

[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): 1129-40. http://dx.doi.org/10.1016/j.ygeno.2020.11.009 PMID: 33189776

[10] Mason RJ. Thoughts on the alveolar phase of COVID-19. Am J Physiol Lung Cell Mol Physiol 2020; 319(1): L115-20. http://dx.doi.org/10.1152/ajplung.00126.2020 PMID: 32493030

[11] Ragab D, Salah Eldin H, Taeimah M, Khattab R, Salem R. The COVID-19 cytokine storm; what we know so far. Front Immunol 2020; 11: 1446. http://dx.doi.org/10.3389/fimmu.2020.01446 PMID: 32612617

[12] Wang W, Xu Y, Gao R, *et al.* Detection of SARS-CoV-2 in different types of clinical specimens. JAMA 2020; 323(18): 1843-4. http://dx.doi.org/10.1001/jama.2020.3786 PMID: 32159775

[13] Wu Y, Guo C, Tang L, *et al.* Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. Lancet Gastroenterol Hepatol 2020; 5(5): 434-5. http://dx.doi.org/10.1016/S2468-1253(20)30083-2 PMID: 32199469

[14] Jamiolkowski D, Mühleisen B, Müller S, Navarini AA, Tzankov A, Roider E. SARS-CoV-2 PCR testing of skin for COVID-19 diagnostics: a case report. Lancet 2020; 396(10251): 598-9. http://dx.doi.org/10.1016/S0140-6736(20)31754-2 PMID: 32798450

[15] Bacherini D, Biagini I, Lenzetti C, Virgili G, Rizzo S, Giansanti F. The COVID-19 pandemic from an ophthalmologist's perspective. Trends Mol Med 2020; 26(6): 529-31. http://dx.doi.org/10.1016/j.molmed.2020.03.008 PMID: 32470381

[16] Dong Z, Xiang BJ, Jiang M, Sun M, Dai C. The prevalence of gastrointestinal symptoms, abnormal liver function, digestive system disease and liver disease in COVID-19 infection. J Clin Gastroenterol 2021; 55(1): 67-76. http://dx.doi.org/10.1097/MCG.0000000000001424 PMID: 33116063

[17] Topol EJ. COVID-19 can affect the heart. Science 2020; 370(6515): 408-9. http://dx.doi.org/10.1126/science.abe2813 PMID: 32967937

[18] Akilesh S, Nast CC, Yamashita M, *et al.* Multicenter clinicopathologic correlation of kidney biopsies performed in covid-19 patients presenting with acute kidney injury or proteinuria. Am J Kidney Dis 2021; 77(1): 82-93.e1. http://dx.doi.org/10.1053/j.ajkd.2020.10.001 PMID: 33045255

[19] Zhou F, Yu T, Du R, *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 2020; 395(10229): 1054-62. http://dx.doi.org/10.1016/S0140-6736(20)30566-3 PMID: 32171076

[20] Platt J. Fast training of support vector machines using sequential minimal optimization.Advances in kernel methods: Support vector learning. Advances in Kernel Methods: Support Vector Learning. Cambridge, Massachusetts: MIT Press 1998; pp. 185-208. http://dx.doi.org/10.3390/cimb43020061 PMID: 34449545

[21] Diogo D, Tian C, Franklin CS, *et al.* Phenome-wide association studies across large population cohorts support drug target validation. Nat Commun 2018; 9(1): 4285. http://dx.doi.org/10.1038/s41467-018-06540-3 PMID: 30327483

[22] Nelson MR, Tipney H, Painter JL, *et al.* The support of human genetic evidence for approved drug indications. Nat Genet 2015; 47(8): 856-60. http://dx.doi.org/10.1038/ng.3314 PMID: 26121088

[23] Sirota M, Dudley JT, Kim J, *et al.* Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med 2011; 3(96)96ra77 http://dx.doi.org/10.1126/scitranslmed.3001318 PMID: 21849665

[24] Dudley JT, Sirota M, Shenoy M, *et al.* Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. Sci Transl Med 2011; 3(96)96ra76 http://dx.doi.org/10.1126/scitranslmed.3002648 PMID: 21849664

[25] He X, He C, Hong W, Zhang K, Wei X. The challenges of COVID-19 Delta variant: Prevention and vaccine development. MedComm 2021; 2(4): 846-54. http://dx.doi.org/10.1002/mco2.95 PMID: 34909755

[26] Karim SSA, Karim QA. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. Lancet 2021; 398(10317): 2126-8. http://dx.doi.org/10.1016/S0140-6736(21)02758-6 PMID: 34871545

[27] Nagy Á, Pongor S, Győrffy B. Different mutations in SARS-CoV-2 associate with severe and mild outcome. Int J Antimicrob Agents 2021; 57(2): 106272. http://dx.doi.org/10.1016/j.ijantimicag.2020.106272 PMID: 33347989

[28] Chen J, Gao K, Wang R, Wei GW. Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies. Chem Sci (Camb) 2021; 12(20): 6929-48. http://dx.doi.org/10.1039/D1SC01203G PMID: 34123321

[29] Zawbaa HM, Osama H, El-Gendy A, *et al.* Effect of mutation and vaccination on spread, severity, and mortality of COVID-19 disease. J Med Virol 2022; 94(1): 197-204. http://dx.doi.org/10.1002/jmv.27293 PMID: 34427922

[30] Wu H, Xing N, Meng K, *et al.* Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. Cell Host Microbe 2021; 29(12): 1788-1801.e6. http://dx.doi.org/10.1016/j.chom.2021.11.005 PMID: 34822776

[31] Martinot M, Jary A, Fafi-Kremer S, *et al.* Emerging RNA-dependent RNA polymerase mutation in a remdesivir-treated B-cell immunodeficient patient with protracted coronavirus disease 2019. Clin Infect Dis 2021; 73(7): e1762-5. http://dx.doi.org/10.1093/cid/ciaa1474 PMID: 32986807

[32] Focosi D, Novazzi F, Genoni A, *et al.* Emergence of SARS-COV-2 spike protein escape mutation Q493R after treatment for COVID-19. Emerg Infect Dis 2021; 27(10): 2728-31. http://dx.doi.org/10.3201/eid2710.211538 PMID: 34314668

[33] Hassan SS, Choudhury PP, Roy B. SARS-CoV2 envelope protein: non-synonymous mutations and its consequences. Genomics 2020; 112(6): 3890-2.

http://dx.doi.org/10.1016/j.ygeno.2020.07.001 PMID: 32640274
[34] Rahman MS, Hoque MN, Islam MR, *et al.* Mutational insights into the envelope protein of SARS-CoV-2. Gene Rep 2021; 22100997 http://dx.doi.org/10.1016/j.genrep.2020.100997 PMID: 33319124
[35] Barrett T, Wilhite SE, Ledoux P, *et al.* NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res 2013; 41(Database issue): D991-5. PMID: 23193258
[36] Desai N, Neyaz A, Szabolcs A, *et al.* Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection. Nat Commun 2020; 11(1): 6319. http://dx.doi.org/10.1038/s41467-020-20139-7 PMID: 33298930
[37] Pierce CA, Sy S, Galen B, *et al.* Natural mucosal barriers and COVID-19 in children. JCI Insight 2021; 6(9)e148694 http://dx.doi.org/10.1172/jci.insight.148694 PMID: 33822777
[38] Chan YH, Fong SW, Poh CM, *et al.* Asymptomatic COVID-19: disease tolerance with efficient anti-viral immunity against SARS-CoV-2. EMBO Mol Med 2021; 13(6): e14045. http://dx.doi.org/10.15252/emmm.202114045 PMID: 33961735
[39] Rother N, Yanginlar C, Lindeboom RGH, *et al.* Hydroxychloroquine inhibits the trained innate immune response to interferons. Cell Rep Med 2020; 1(9): 100146. http://dx.doi.org/10.1016/j.xcrm.2020.100146 PMID: 33377122
[40] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. embnet j 2011; 17(1): 1-3. http://dx.doi.org/10.14806/ej.17.1.200
[41] Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 2016; 11(9): 1650-67. http://dx.doi.org/10.1038/nprot.2016.095 PMID: 27560171
[42] Li H, Handsaker B, Wysoker A, *et al.* The sequence alignment/map format and SAM tools. Bioinformatics 2009; 25(16): 2078-9. http://dx.doi.org/10.1093/bioinformatics/btp352 PMID: 19505943
[43] Freshour SL, Kiwala S, Cotto KC, *et al.* Integration of the drug–gene interaction database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Res 2021; 49(D1): D1144-51. http://dx.doi.org/10.1093/nar/gkaa1084 PMID: 33237278
[44] Wishart DS, Feunang YD, Guo AC, *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018; 46(D1): D1074-82. http://dx.doi.org/10.1093/nar/gkx1037 PMID: 29126136
[45] Nurgali K, Jagoe RT, Abalo R. Editorial: Adverse effects of cancer chemotherapy: Anything new to improve tolerance and reduce sequelae? Front Pharmacol 2018; 9: 245. http://dx.doi.org/10.3389/fphar.2018.00245 PMID: 29623040
[46] Stroup TS, Gray N. Management of common adverse effects of antipsychotic medications. World Psychiatry 2018; 17(3): 341-56. http://dx.doi.org/10.1002/wps.20567 PMID: 30192094
[47] Bujotzek A, Fuchs A, Qu C, *et al.* MoFvAb: Modeling the Fv region of antibodies. MAbs 2015; 7(5): 838-52. http://dx.doi.org/10.1080/19420862.2015.1068492 PMID: 26176812
[48] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 2010; 31(2): 455-61. PMID: 19499576
[49] Justo Arevalo S, Zapata Sifuentes D, J Huallpa C, *et al.* Dynamics of SARS-CoV-2 mutations reveals regional-specificity and similar trends of N501 and high-frequency mutation N501Y in different levels of control measures. Sci Rep 2021; 11(1): 17755. http://dx.doi.org/10.1038/s41598-021-97267-7 PMID: 34493762
[50] Singh H, Kakkar AK, Chauhan P. Repurposing minocycline for COVID-19 management: mechanisms, opportunities, and challenges. Expert Rev Anti Infect Ther 2020; 18(10): 997-1003. http://dx.doi.org/10.1080/14787210.2020.1782190 PMID: 32552044
[51] Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. Virol J 2019; 16(1): 69. http://dx.doi.org/10.1186/s12985-019-1182-0 PMID: 31133031
[52] Bwire GM. Coronavirus: Why men are more vulnerable to Covid-19 than women? SN Compr Clin Med 2020; 2(7): 874-6. http://dx.doi.org/10.1007/s42399-020-00341-w PMID: 32838138
[53] Al-kuraishy HM, Al-Gareeb AI, Alzahrani KJ, Alexiou A, Batiha GES. Niclosamide for Covid-19: bridging the gap. Mol Biol Rep 2021; 48(12): 8195-202. http://dx.doi.org/10.1007/s11033-021-06770-7 PMID: 34664162