Investigation of Mutations in the non-coding regions of the cancer genome

A Thesis submitted to
Indian Institute of Science Education and Research Pune
in partial fulfilment of the requirements for the
BS-MS Dual Degree Programme
by

Arkajyoti Ghoshal

Bs-Ms 5th year

Dept. of biology

Guided by:

Dr. Sudhakaran Prabakaran

University of Cambridge

Dept. of Genetics

IISER PUNE

Certificate

This is to certify that this dissertation entitled ‘Investigation of mutations in the non-coding regions of the cancer genome’ should appear heretowards the partial fulfillment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Arkajyoti Ghoshal at Indian Institute of Science Education and Research under the supervision of Dr. Sudhakaran Prabakaran, University of Cambridge, Department of Genetics , during the academic year 2017-2018.

Student: Arkajyoti Ghoshal

Signature:

Supervisor: Dr. Sudhakaran Prabakaran

Signature:

Thesis Advisory Committee: Dr. Krishanpal Karmodiya

Declaration

I, hereby declare that the matter embodied in the report titled "Investigation of mutations in the non-coding regions of the cancer genome" is the results of the investigations carried out by me at the Indian Institute of Science Education and Research, Pune under the supervision of Dr Sudhakaran Prabakaran and the same has not been submitted elsewhere for any other degree.

Student: Arkajyoti Ghoshal

Signature:

Supervisor: Dr. Sudhakaran Prabakaran

Signature:

Acknowledgement

I would like to thank my family and all the people associated with this project especially my supervisor Dr. Sudhakaran Prabakaran, who guided me through the difficult phases during the course of this project, my lab members including Narendra Meena, Matt Wayland, Felix Jackson and Chaitanya Erady. I am also thankful to my TAC member Dr. Krishanpal Karmodiya for his valuable insights during mid-term evaluation. Lastly, I would like to thank IISER for providing me this opportunity.

Abstract

Encode shows that most of the human genome is biochemically active with respect to transcription, translation, etc. though most of the products and their roles are not yet known. Previous studies in our lab using whole genome sequencing, total RNA seq transcriptomics, and proteomics data, integrated in a compact framework developed by our own group called ‘Systems Proteogenomics’, has shown evidence of translation of protein-like products from the entire genome (Prabakaran, S. et al., 2014.). We hypothesize that most of these ‘protein-like’ products could be actively involved in maintaining our physiology as studies have indicated that 90% of disease-associated mutations are mapped to regions that are currently identified as ‘noncoding’ (St. Laurent et al., 2014). Here, we aim to investigate mutations in these noncoding regions that code for these protein-like products and develop a pipeline to validate them and try to understand their roles in context of development and progression of five different types of cancer, especially.

Contents

List of Figures

Introduction:

After decades of research, curing cancer still remains a challenge and the complexity of the disease plays an important role. At the genetic level, cancer refers to any sets of mutation in the DNA of somatic cells that alter important physiological processes of the cell, including cell cycle progression and apoptosis inhibition, causing them to proliferate unchecked. With increased cell division, there is an accumulation of greater number of mutation (genomic instability), a trademark of cancer (1,2). There are two types of these mutations- driver and passenger. A driver mutation is one that gives a tissue and its clones selective advantage in a particular microenvironment (say tumor microenvironment) through increased survival or reproduction, leading to clonal expansion. (clones are sets of cells descending from a common ancestor). Passenger mutations are those who have no direct effect on the fitness of a clonal population with respect to its survival but often occurs in the same genome with driver mutations. As these are more frequent than driver mutations, separation of the driver mutations (information) from background passenger mutations (noise) is a very important step.

Extensive studies have been carried out to understand the roles and significance of driver mutations in protein-coding regions. They are mostly oncogenic somatic mutations and include SNVs, indels, CNVs, etc. Their expression in a cell is often related to its acquisition of 'hallmarks of cancer' including unconstrained proliferation, replicative immortality, immune evasion (3). Much research has been done to understand the nature of these mutations at a computation and experimental level. However, most of these studied 'driver' mutation occur in the protein-coding regions (4,5). Previously, due to lack of techniques like WGS, most of the datasets used for these studies would mostly contain information of known proteins and markers. The process of understanding the significance of these mutations is further complicated by the fact that ratio of functional driver mutation to biologically insignificant passenger mutations biased on the side of passenger mutations. Thus, separation of the driver mutations from background passenger mutations is a very important step. Filtration must be done carefully, without losing too much information and maintain a high degree of accuracy.

Assimilating a complete list of such genes and their mutations is of utmost importance to understand a disease and design its therapeutics accordingly. This includes cancer too. This started off projects like Cancer Gene Census project (6). With increased technological tools, the nature of human molecular data improved and in-depth databases of cancer patient samples were created. One of them includes The Cancer Genome Atlas (TCGA) which is the most comprehensive database for cancer patient genome, covering both 'coding' and 'non-coding' region data for more than 30 different cancer types (7).

Such databases are important sources of information, especially of the 'non-coding' regions of the human genome. This is mainly because prior to NGS, such detailed information regarding non-coding regions was difficult to capture and only recently have studies been started to understand the roles and significance of mutations in the non-coding regions and how they affect the development and progression of cancer.

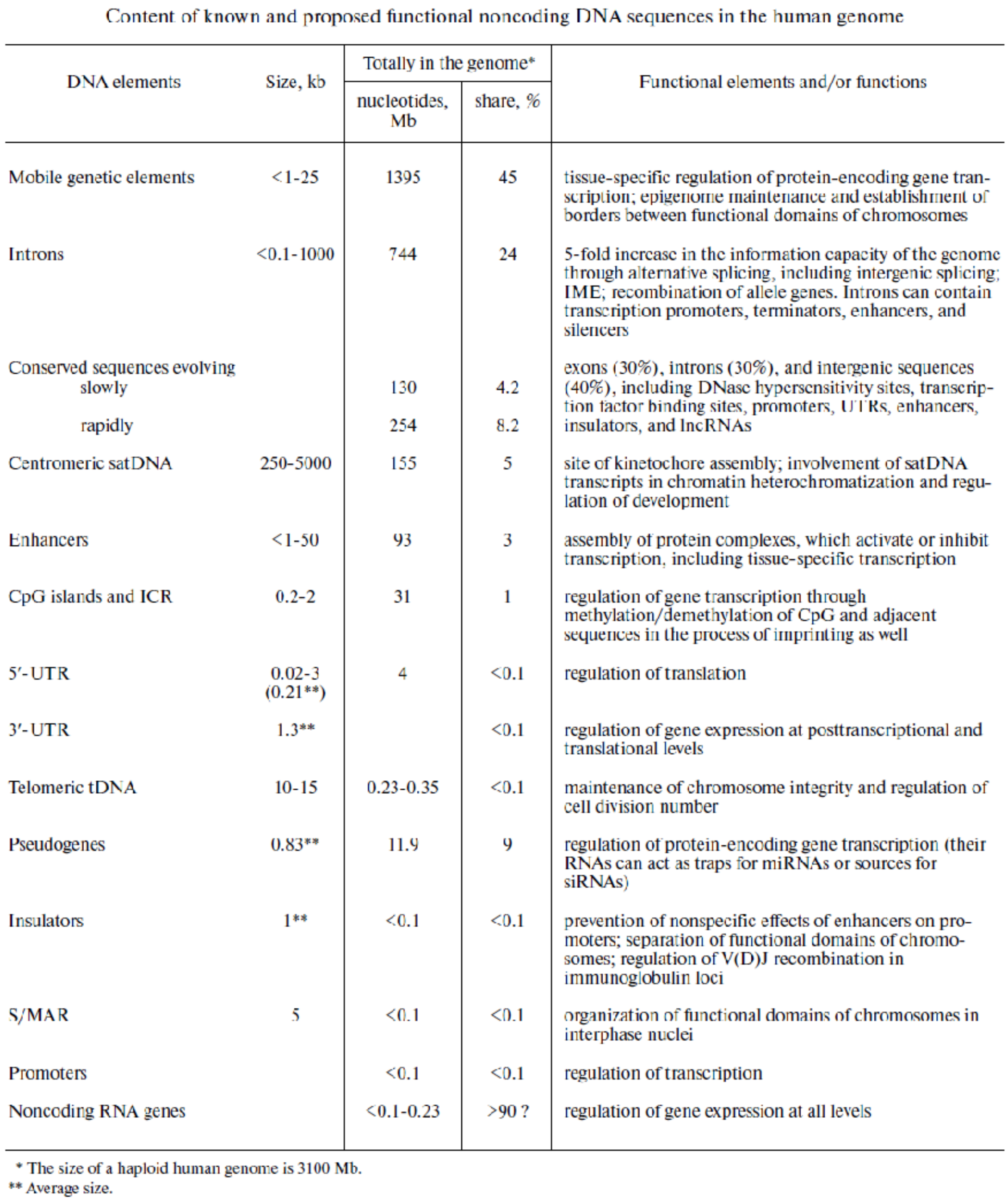
Our lab is one of such labs working to understand the roles of non-coding regions in diseases including cancer. Previous studies in our lab has shown evidence of translation of protein-like products from most of the human genome (8). Such results along with results showing 90% of disease-associated mutations being mapped to regions that are currently identified as noncoding regions (8) makes us speculates us that most of these protein-like products that we and few others have identified could be actively involved in maintaining normal human physiology. We thus hypothesize that mutations in these regions might be responsible for cancer and if so, study the mutations and their roles in oncogenesis arising from different non-coding regions of the human genome.

Our lab has developed a frame-work for studying these genomic regions and their associated mutations in order to understand their functional implication in diseases. Such a frame-work, called 'Systems Proteogenomics' using Whole Genome Sequencing, total RNA-seq transcriptomics and proteomics data to identify and categorize potential 'onco-regions' of the 'non-coding' genome and investigates the roles of mutations in cancer progress so as to devise suitable therapeutics for the disease. Details of the frame-work are provided in Materials and Methods section.

Statistics show that currently in the world, 5 of the top ten most common cancer types include 1) Breast and 2) ovarian cancer (two of the highest affecting cancer types for females), 3) lung (the highest number of cases/mortality world-wide (9)), 4) prostate (similar for males) and 5) skin cancer (one of the highest affecting cancer types in world) (10,11)

Thus, for our project we chose to focus on mutations involved in these 5 cancer types. Acquisition of datasets, processing, etc. are described in the Materials and Methods section in details.

In the late 1960s, when people started finding out non-protein coding DNA sequences all across the genome, they classified them as 'junk' as they had no protein coding potential (12). With advent in technology like WGS, it was seen that protein coding regions or exons cover only 2-3% of the genome. While most of the non-coding genome actually do not code for any peptides, they might not without function. Such hypotheses are supported by observations showing that in 'higher' organisms like human, throughout evolution, at least 10% of non-coding regions (ncrs) are conserved evolutionarily (13) and most of the human genome show pervasive translation(15), as shown in our lab (8). Such regions include repeat elements, 3' and 5' UTRs, non-coding RNAs and others. While, most of their functions are being actively checked for, the table below in Fig. 1 shows the different ncrs and propose their possible functions as found by researchers.



Content of known and proposed functional noncoding DNA sequences in the human genome

DNA elements	Size, kb	Totally in the genome*		Functional elements and/or functions
		nucleotides, Mb	share, %	
Mobile genetic elements	<1-25	1395	45	tissue-specific regulation of protein-encoding gene transcription; epigenome maintenance and establishment of borders between functional domains of chromosomes
Introns	<0.1-1000	744	24	5-fold increase in the information capacity of the genome through alternative splicing, including intergenic splicing; IME; recombination of allele genes. Introns can contain transcription promoters, terminators, enhancers, and silencers
Conserved sequences evolving slowly		130	4.2	exons (30%), introns (30%), and intergenic sequences (40%), including DNase hypersensitivity sites, transcription factor binding sites, promoters, UTRs, enhancers, insulators, and lncRNAs
rapidly		254	8.2	
Centromeric satDNA	250-5000	155	5	site of kinetochore assembly; involvement of satDNA transcripts in chromatin heterochromatization and regulation of development
Enhancers	<1-50	93	3	assembly of protein complexes, which activate or inhibit transcription, including tissue-specific transcription
CpG islands and ICR	0.2-2	31	1	regulation of gene transcription through methylation/demethylation of CpG and adjacent sequences in the process of imprinting as well
5′-UTR	0.02-3 (0.21**)	4	<0.1	regulation of translation
3′-UTR	1.3**		<0.1	regulation of gene expression at posttranscriptional and translational levels
Telomeric tDNA	10-15	0.23-0.35	<0.1	maintenance of chromosome integrity and regulation of cell division number
Pseudogenes	0.83**	11.9	9	regulation of protein-encoding gene transcription (their RNAs can act as traps for miRNAs or sources for siRNAs)
Insulators	1**	<0.1	<0.1	prevention of nonspecific effects of enhancers on promoters; separation of functional domains of chromosomes; regulation of V(D)J recombination in immunoglobulin loci
S/MAR	5	<0.1	<0.1	organization of functional domains of chromosomes in interphase nuclei
Promoters		<0.1	<0.1	regulation of transcription
Noncoding RNA genes		<0.1-0.23	>90 ?	regulation of gene expression at all levels

* The size of a haploid human genome is 3100 Mb.
** Average size.

Figure 1: Function of different non-coding regions. This table lists the possible functions of ncrs (Source: Patrushev et al, 2014)

However, one fascinating observation about ncrs come from GWAS studies and work done by St. Laurent et al (2014) showing that more that 90% of disease-associated mutations are mapped to ncrs (14). One can thus hypothesize that these regions are necessary for maintenance of healthy physiology and are thus important ‘peptide’ coding region of the genome, though annotated outside the known exons.

Our lab have been working on identifying and annotating such non-coding regions that show transcription and translational potential. We have built custom databases for such regions with resources curated from publications, online databases and experimental results. For my project, I have used 3 such regions. They are 1) Small ORFs or sorfs, 2) Denovo genes, 3) Pseudogenes (Translated, Transcribed and non-transcribed). Details regarding these (e.g. number of ncrs, etc are mentioned in Materials and Methods section).

As a brief outline for this project, I mapped mutations for 5 cancer types to these 4 ncrs, selected particular ncrs and their mutations based on certain biological parameters and performed downstream analyses. The end product is a list of each type of ncrs and mutations that must be taken into consideration for these specific cancer type diagnosis.

This framework can be integrated to other models say that takes up such ncrs, tries to predict its mutated protein structures and come up with pathways that can cause oncogenesis. This has a huge potential for therapeutic purposes and when combined with similar studies on known protein-coding genes, can theoretically combat cancer and many such disease types.

Materials and Methods:

A] Collection of Mutation Dataset

- We downloaded the entire database of non-coding mutations from COSMIC v82(Catalogue of Somatic Mutations in Cancer) (16, download link provided). COSMIC is the world’s largest resource of high precision, (manually) expert curated database of somatic mutations pertaining to cancer (about 33 cancer types till v82). Datasets are collected from different sources as indicated in Fig. 2, taken from their website. These datasets provide a deep insight into the cancer genomic landscape from a somatic perspective. New and potentially significant data are continually captured and updated regularly.

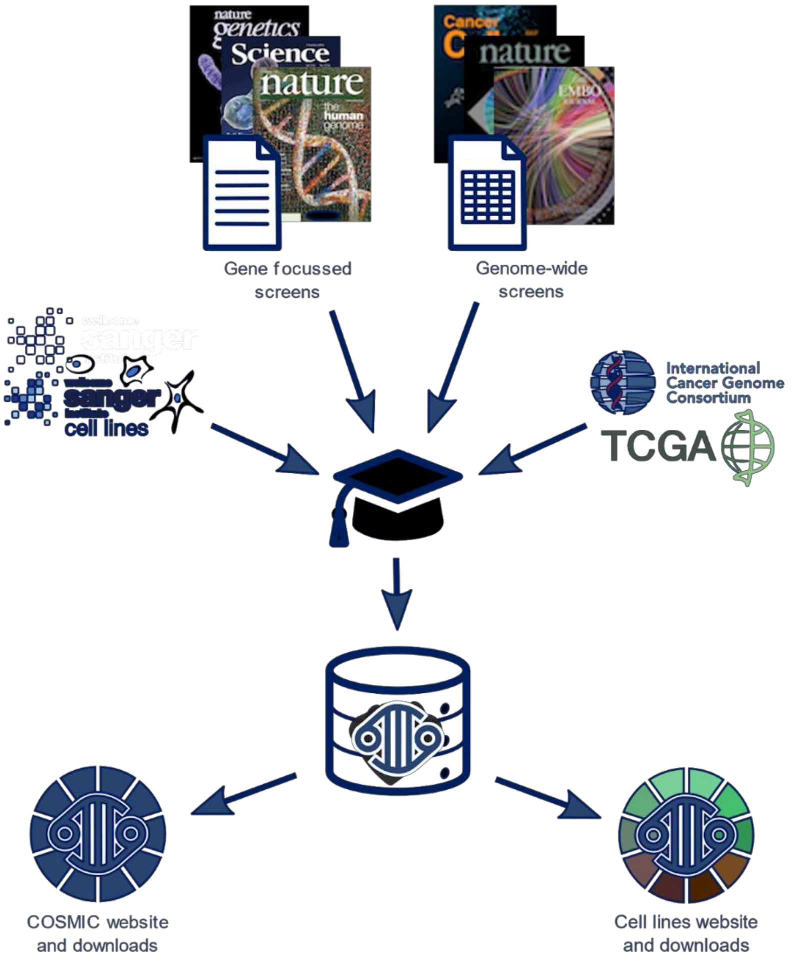

Gene focussed screens
Genome-wide screens
wellcome sanger institute cell lines
International Cancer Genome Consortium
TCGA
COSMIC website and downloads
Cell lines website and downloads

Figure 2: Cosmic Database; overview and sources of data collection

- After that, the file (.txt) was filtered for the 5 different cancer types using scripts written in Python (v3.7), individual cancer type datasets extracted and formatted to generate 5 .csv files.

B] Formation of NCR datasets

- There are 4 types of datasets as mentioned above, each of which has been curated and collected by our lab.

- Sorfs or small open reading frames (Orfs) are orfs of length 100 to 300 nucleotides and produces ‘small proteins’ (avg. size of 100 amino acids (a.a.)) [Note: avg. coding-region protein size in humans is about 500 a.a. (17)]. Initially unannotated as it was thought to be ‘non-coding’, with advent of techniques like ribosome profiling and deep-sequencing, it is now seen that many of such orfs are actively translated. With research, the importance of such peptides in maintain human physiology is being investigated and appreciated, as is the fact that many of such functional sorfs are evolutionarily conserved. (18,19). Two such public repositories containing validated sorfs are Sorfs.org and Smprot (20,21) from which another master's student in our lab have curated and constructed a human sorf dataset containing 574212 entries. We used our own database for studies.

- Denovo genes are active genes arising from previously non-coding regions via different mechanisms. They have high death rates but some of them can get integrated into the genomic network and become integrated part of it. Another student in our lab have built a database of 42 such experimentally validated Denovo genes which was used.

- Pseudogenes are non-coding regions arising from coding genes by processes like a) duplication, during which mutational processes can render a copy/copies of a functional gene 'disabled' or b) retrotransposition of a functional mRNA (without start/stop sites) into the genomic cDNA via reverse transcription or may be coding regions lacking functional start/stop site. c) Sometimes, they can accumulate mutation to their flanking regions providing them with a working start/stop site or both and their functions maybe altered but they can become active again. Many labs are working on them, one of the foremost being Mark Gerstein's lab ate Yale University and they have set up public databases like Pseudogene.org (http://pseudogene.org/index.html). Such databases were curated by our lab to obtain three types of datasets: Translated (34), Transcribed (540) and non-transcribed (1159) pseudogenes. I used all 3 types for my work.

All these files were converted to .csv format and had alteast 4 common annotations including 'Id' (database specific), 'Chromosome number', 'Start' and 'Stop' positions (basically genomic co-ordinates) or formatted accordingly.

Once these were established, several analyses were done as described below.

1. Mapping of mutations to ncrs: For each of the 3 ncrs, genomic (g) co-ordinates were matched to mutational (m) co-ordinates for same chromosome to see if a mutation lies within the region. Python script written for this purpose will be uploaded in our lab github account soon. For large datasets (e.g. sORFs), R Studio was used. (will also be available in github)
2. Once such mapping was complete, resultant .csv files, containing the genomic co-ordinates as well as the mutational co-ordinates were counted (Python script) for the number of mutation in each ncr for all cancer types. This was then plotted for visualization as seen in Results section
3. Next, we were interested in finding out how many of such ncrs were involved in the different cancer types. For this purpose, all the mapped datasets for each ncr was merged together for the different cancer types and a venn diagram was plotted using R studios. Thus ncrs relative to each type of mutation as well as common to different combinations of

the mutations were identified. All such files could be used for separate set of downstream analyses. The plots can be seen in Results section.

4. Next, we decided to classify these mutations based on their impact in terms of ‘pathogenicity’ or disease-causing potential. For that, we used ‘Functional Analysis through Hidden Markov Models’ or FATHMM scoring system. FATHMM’s web-server utilises the algorithm to predict effects of mutations in both coding and non-coding regions. FATHMM non-coding model (FMKL) uses functional annotations from ENCODE in tandem with Hidden Markov Models on nucleotides (HMM). It also utilizes annotations to learn to weigh significance of each such variants (eg SNPs) with high degree of confidence (23). However, the algorithm has been trained on human gene mutation database (The HGMD database http://www.hgmd.cf.ac.uk/ac/index.php) which does contain somatic mutations as well. COSMIC is performing similar studies to create cancer specific FATHMM algorithm. According to current scoring system, nc FMKL are represented as a p value between 0 and 1. Scores <= 0.5 are classified as ‘neutral’, those above 0.5 as ‘deleterious’ and those above 0.7 as ‘pathogenic’. However, we chose to include all mutations with scores >0.5 to ensure incorporation of maximum number of ‘negative’ mutations that might not have been classified correctly by the current algorithm system. The machine learning algorithm currently uses 10 ‘feature groups’ including GC content and ‘100-Way Sequence Conservation’.

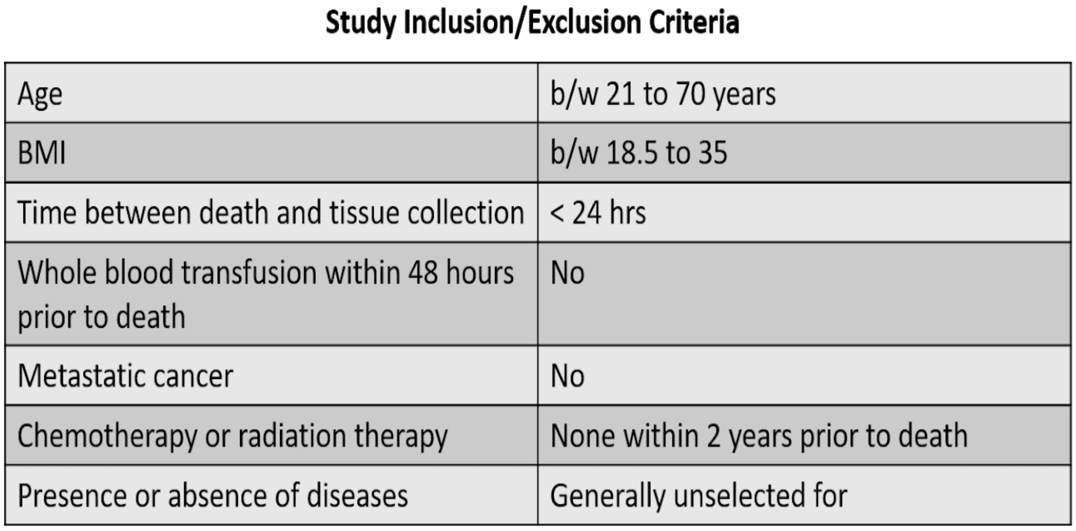
5. Once the criterion was established, each ncr type was parsed for those ncrs with at least 1 pathogenic/deleterious mutation. This was plotted against their total mutation count to have a visualization of distribution of pathogenic mutation. See Results section for details.

6. Pie charts were made using Python scripts to visualize proportions of pathogenic mutation contain ncrs for each cancer type. Also, length distribution of each ncr was plotted as number of nucleotides in x axis v number of ncr in y axis, after dividing the x axis into different bins of increasing length. See Results section for details. This gave us an idea for average size of each ncr and we compared it to known values to filter out extreme outliars. Also, since not much study has been conducted on ncrs, these results are important on their own, giving us an idea on the average length of the 4 ncr types.

7. As a last step of this stage of study, we did Gtex profiling of the ncrs with pathogenic mutations. Gtex or Genotype-Tissue Expression is a project that started in 2010 and tries to better our understanding of genetic changes and their involvement in disease, in order to ‘with the ultimate goal of improving health care for future generations’ (24). It was supported by Director’s office at National Institutes of Health (NIH). Their database helps researchers co-relate between inherited gene expression changes and various diseases. It uses cutting edge research techniques to obtain and store organs and tissues and use them for later

studies. Thus far, it has obtained and analysed maximum number of samples for gene expression and variant identification at a truly unprecedented scale, as opposed to any such similar projects.

Study criteria: The following table, taken from Gtex website (v7), shows the different criteria used during these studies.



Study Inclusion/Exclusion Criteria

Age	b/w 21 to 70 years
BMI	b/w 18.5 to 35
Time between death and tissue collection	< 24 hrs
Whole blood transfusion within 48 hours prior to death	No
Metastatic cancer	No
Chemotherapy or radiation therapy	None within 2 years prior to death
Presence or absence of diseases	Generally unselected for

Adapted from GTex web server (dbGaP Study Accession: phs000424.v2.p1)

Figure 3: Study critera for Gtex Database

Gtex database currently comprises, among other things, expression profile for about 56,200 genes, along with their ensembl ids, for 53 tissue types. These tissue types and the number of samples involved, is depicted in Fig 4

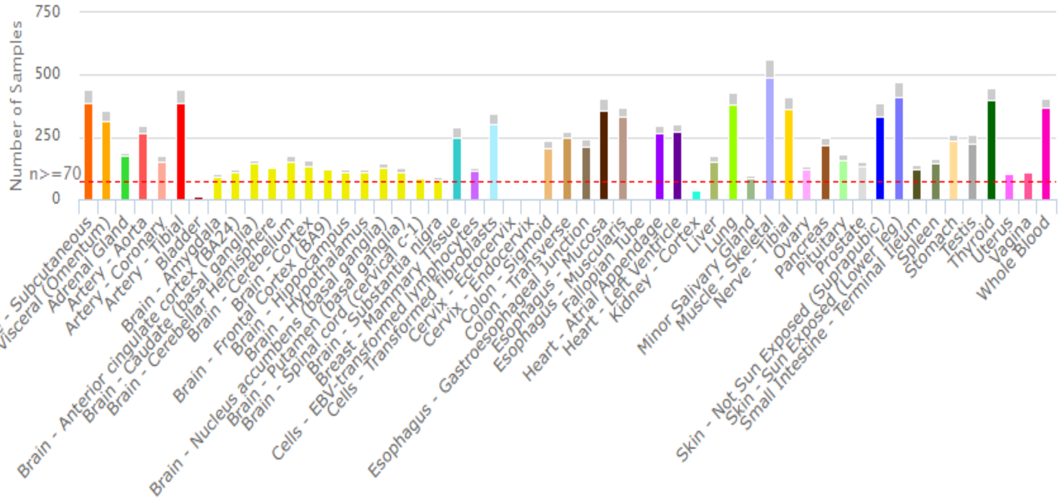


Figure 4: Gtex V7 Sample Counts by Tissues ***(Source: Gtex website)***

Once the dataset was downloaded, Python scripts were written, that enables us to match Ensembl ids of these expressed genes to those of pathogenic mutation containing Denovo and Pseudogenes (note: Ensembl ids for sORFs are not currently accesible). The matched ids were plotted for their expression profiles (in terms of Transcripts per million or TPM) in y axis and the number of tissue types in the x axis. See Results section for details. This allows us to understand which of the ncrs are expressed in healthy/normal tissues and have a visualization of the overall expression profile of each such ncr.

How such profiles can be co-related to diseases is discussed in ‘Future perspective’ section.

Results

a. Mapping of the ncrs to mutations

The proof of concept that mutations related to each caner type might actually lie within the ncrs is shown in the plots below (Fig 5)

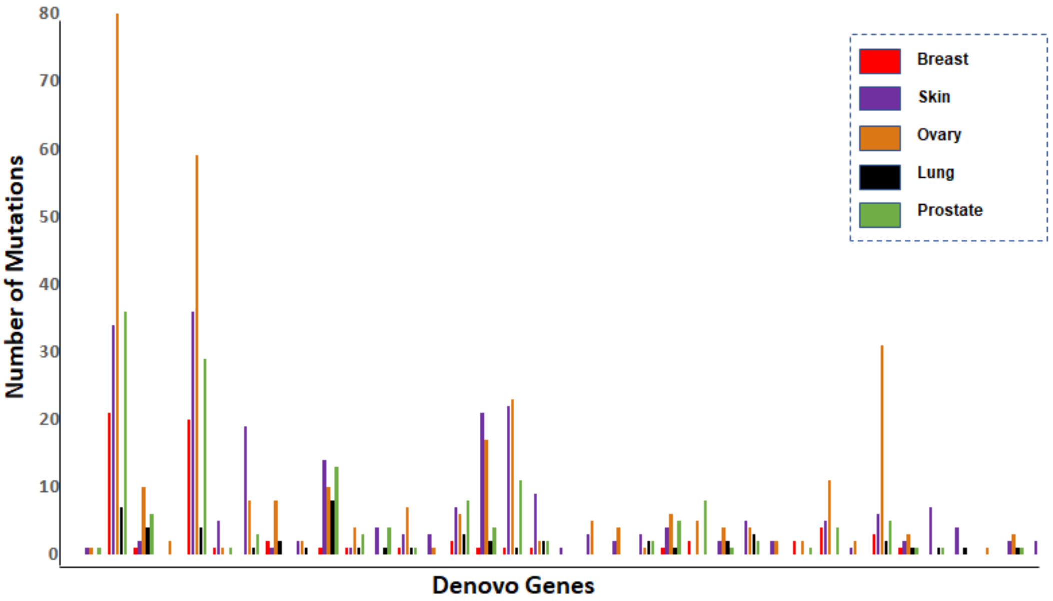


Figure 5: Map of Mutations to Denovo Genes (n=37).

The bars in the y-axis represent the number of mutations for the different denovo gene ids shown in the x-axis. Different cancer types are color-coded as depicted in the legend and each cluster of bars in the x-axis depicts map of mutations for each denovo gene id. There are 37 such clusters.

We initially started with 42 denovo genes among which 37 had atleast 1 type of mutation from the 5 cancer types confirming that oncogenic mutations lie in these regions and such ncrs should be annotated and investigated into.

Fig 5,6 and 7 shows similar results for the three types of pseudogenes (translated, transcribed and non-transcribed respectively). Initially we started working with 34 translated, 539 transcribed and 1159 non-transcribed pseudogenes and mapped mutations to 15, 191 and 353 of them respectively.

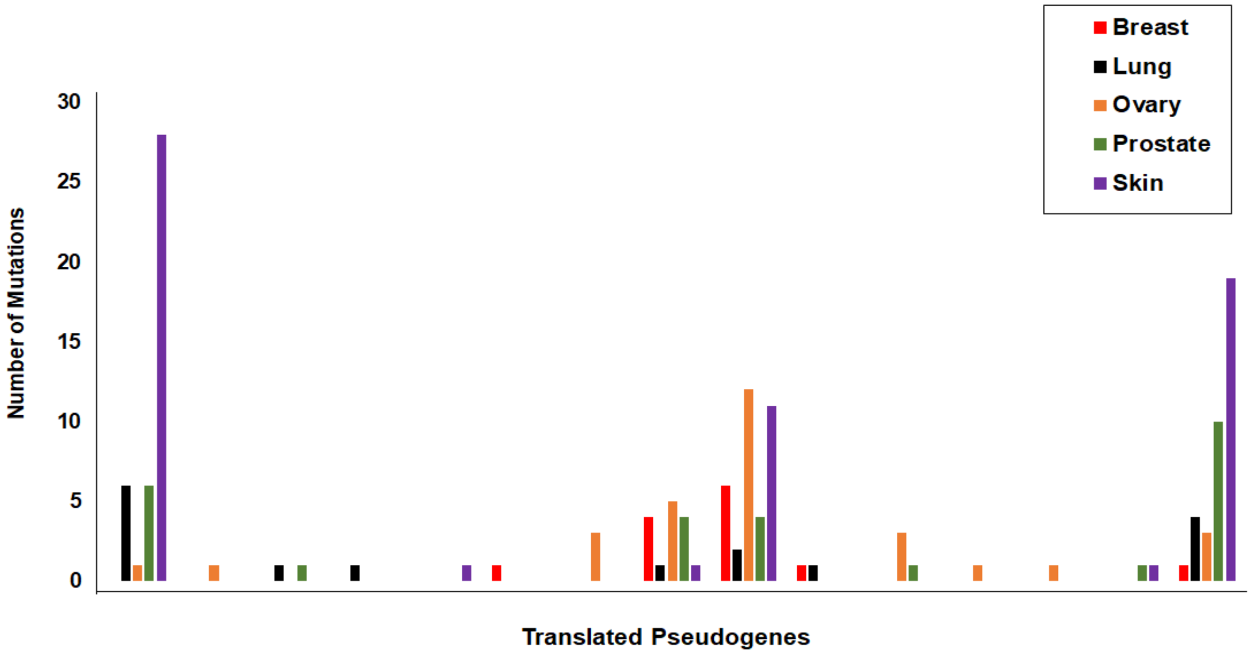


Figure 6: Map of Mutations to Translated Pseudogenes (n=15).

The bars in the y-axis represent the number of mutations for the different pseudogene ids shown in the x-axis. Different cancer types are color-coded as depicted in the legend and each cluster of bars in the x-axis depicts map of mutations for each translated pseudogene id. There are 15 such clusters.

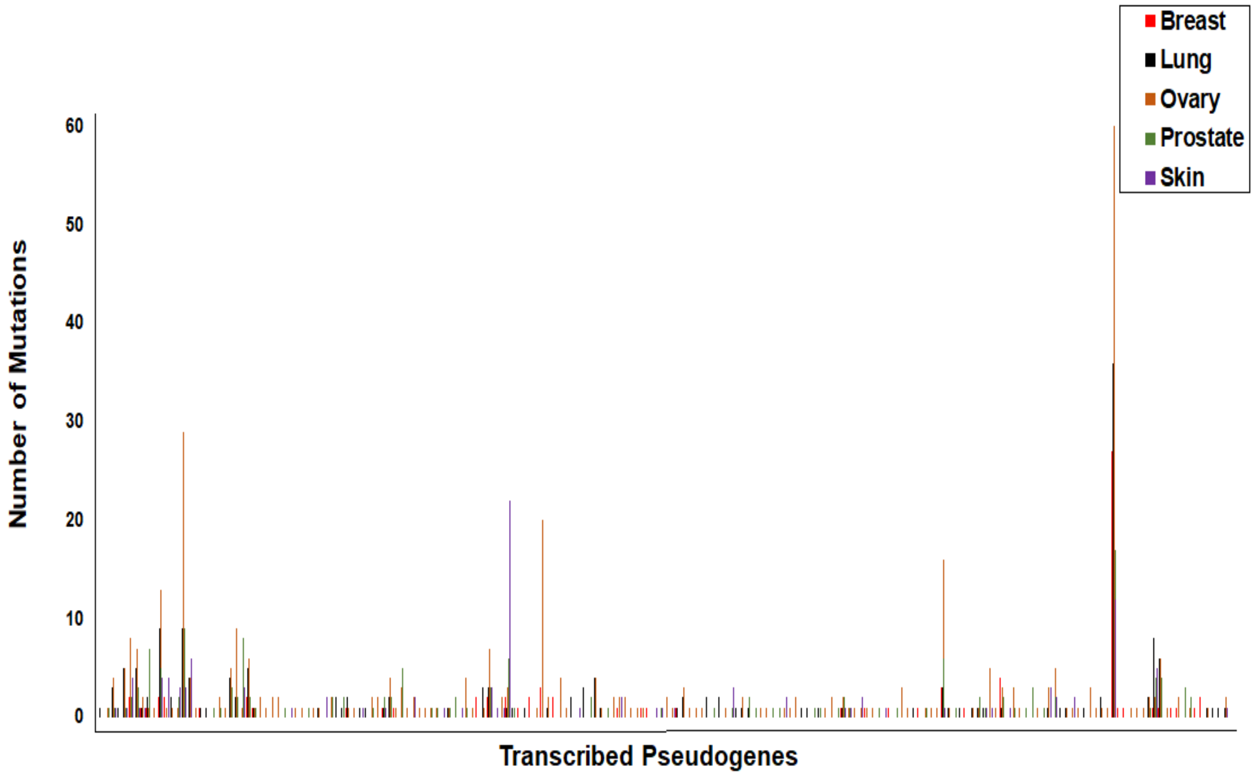


Figure 7: Map of Mutations to Transcribed Pseudogenes (n=191). The bars in the y-axis represent the number of mutations for the different pseudogene ids shown in the x-axis. Different cancer types are color-coded as depicted in the legend and each cluster of bars in the x-axis depicts map of mutations for each transcribed pseudogene id. There are 191 such clusters.

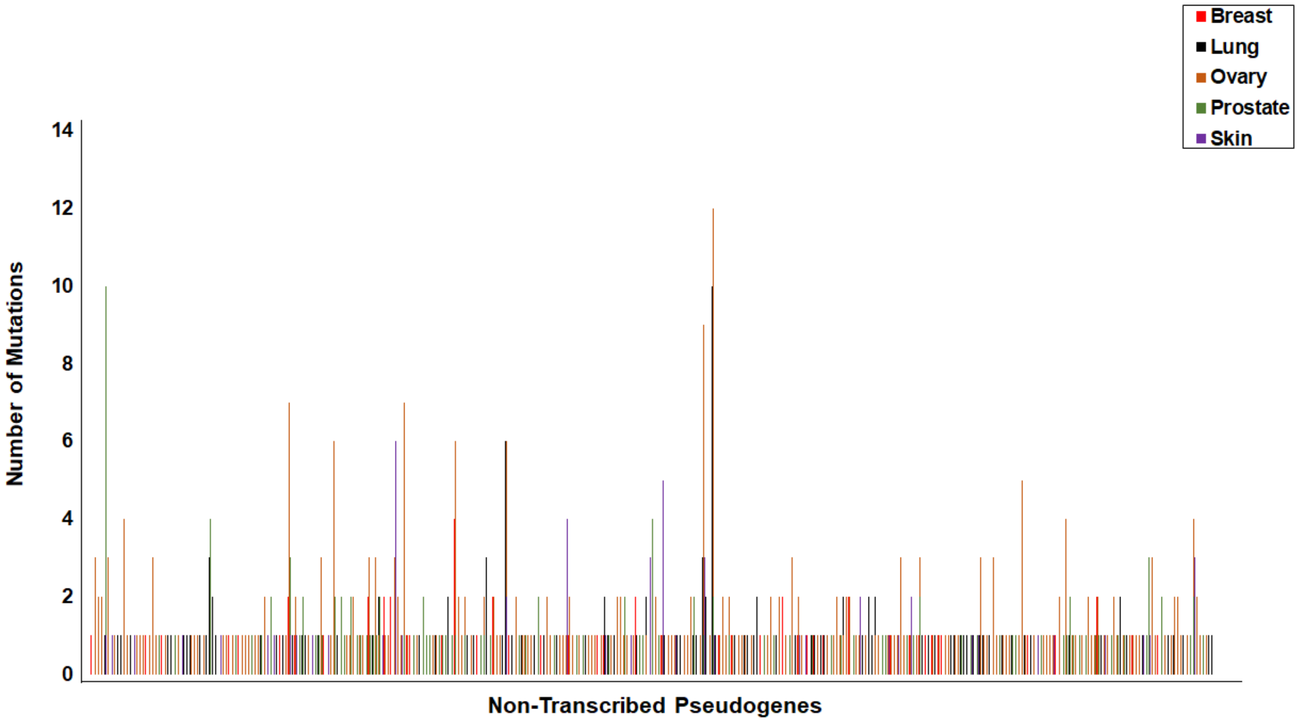


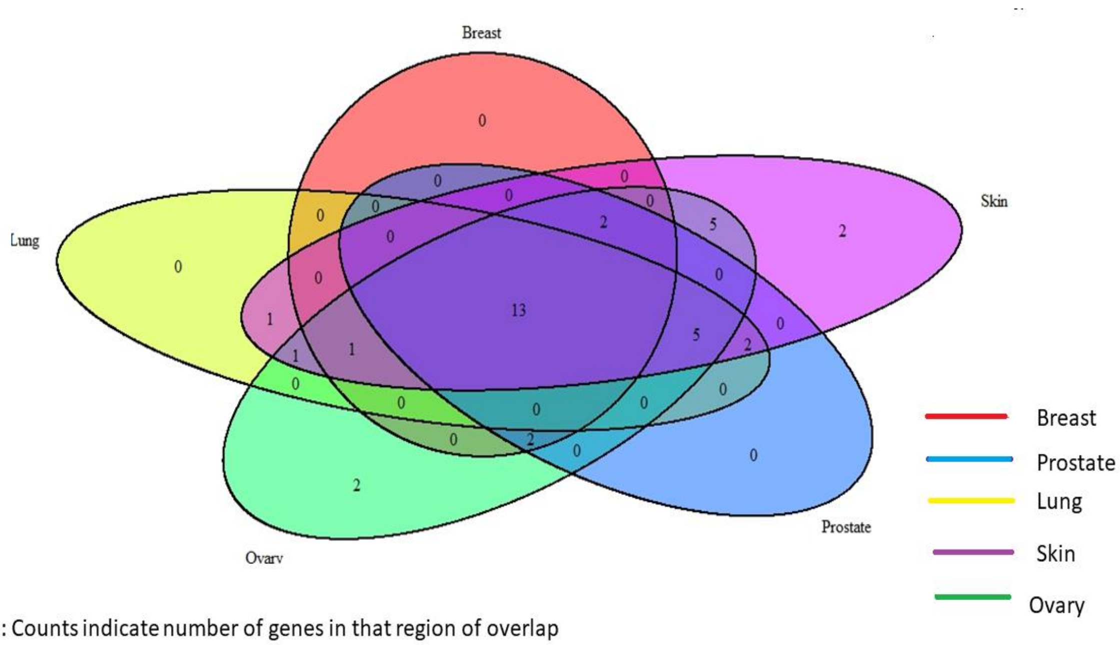
Figure 8: Map of Mutations to Non-Transcribed Pseudogenes (n=353).

The bars in the y-axis represent the number of mutations for the different pseudogene ids shown in the x-axis. Different cancer types are color-coded as depicted in the legend and each cluster of bars in the x-axis depicts map of mutations for each non-transcribed pseudogene id. There are 353 such clusters.

b. Distribution of ncrs across cancer types

Next, we wanted to see how many of each ncr type is involved in different cancers. For this, we used the mapped datasets generated as mentioned above and plotted them as venn diagrams. The plots help us visualize this distribution as shown below.

Fig.9 represents distribution of Denovo genes across cancer types.

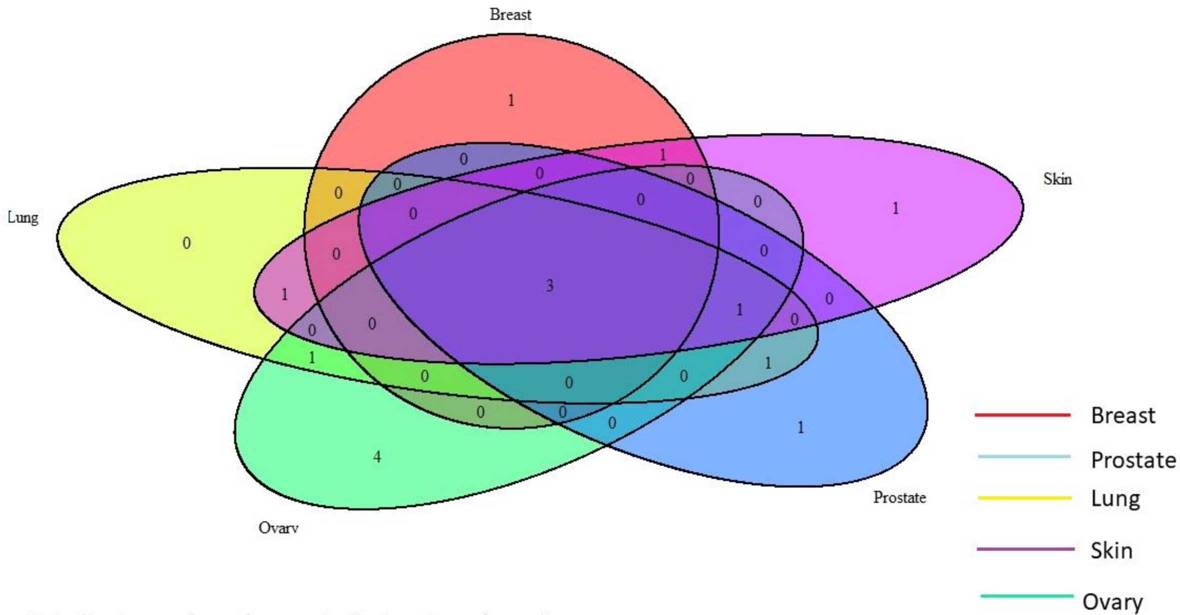


Note: Counts indicate number of genes in that region of overlap

Figure 9: Representation of Denovo Genes Unique and Common to five Cancer Types (n=37).

The different regions of the venn diagram represents the number of denovo genes involved in causing that particular cancer type or its combination. This distribution is for 37 denovo genes.

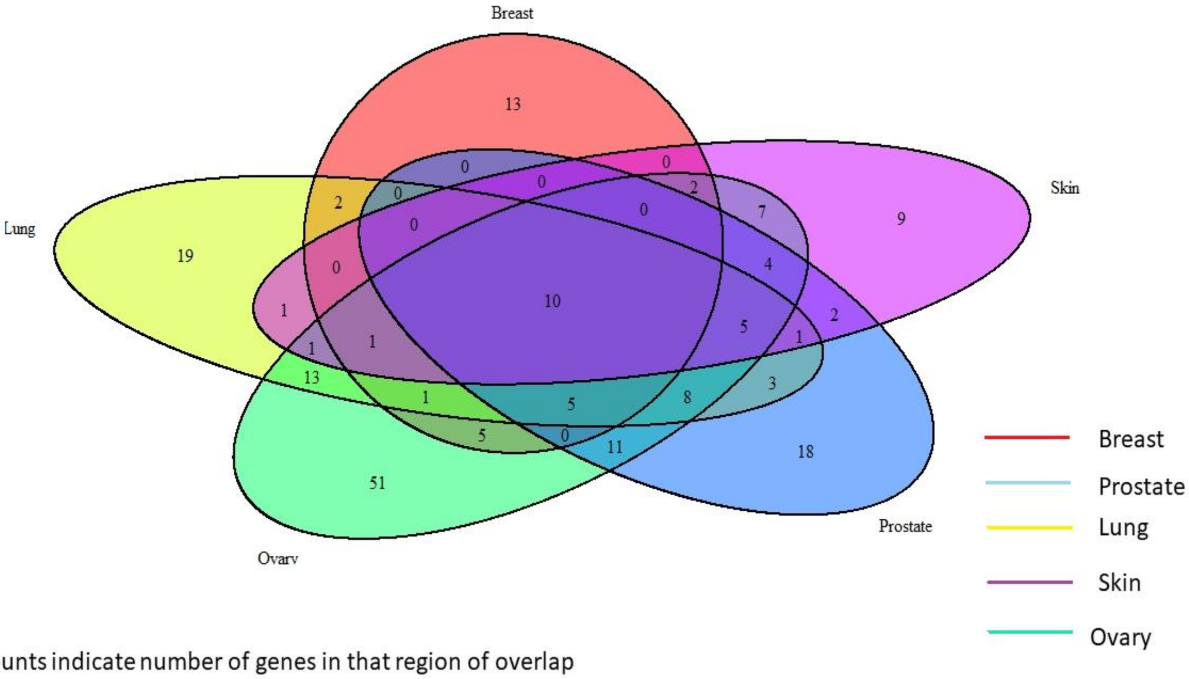
As shown above, out of the 37 denovo genes mapped to mutation, 13 are responsible in all 5, 2 exclusive to ovarian cancer and so on. Similarly, fig 10 to 12 shows venn diagrams for each of the 3 pseudogene types.



Note: Counts indicate number of genes in that region of overlap

Figure 10: Representation of Translated pseudogenes Unique and Common to five Cancer Types (n=34).

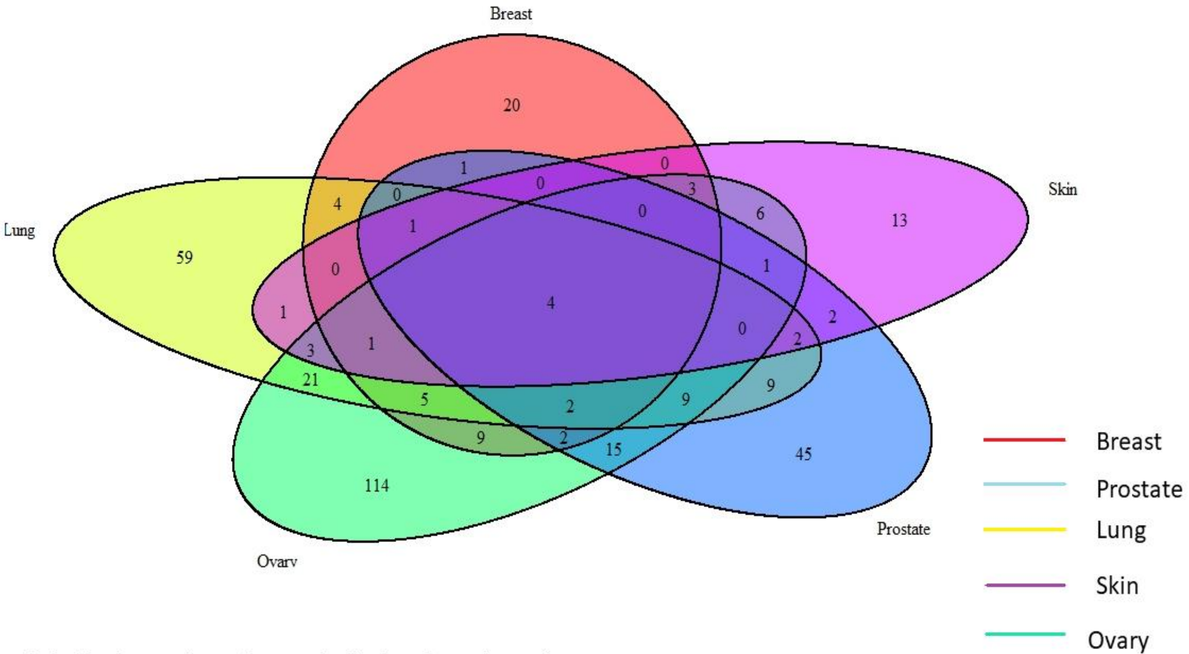
The different regions of the venn diagram represents the number of translated genes involved in causing that particular cancer type or its combination. This distribution is for 34 translated pseudogenes.



Note: Counts indicate number of genes in that region of overlap

Figure 11: Representation of Transcribed pseudogenes Unique and Common to five Cancer Types (n=539).

The different regions of the venn diagram represents the number of transcribed pseudogenes involved in causing that particular cancer type or its combination. This distribution is for 539 transcribed pseudogenes.



Note: Counts indicate number of genes in that region of overlap

Figure 12: Representation of non-transcribed pseudogenes Unique and Common to five Cancer Types (n=1159).

The different regions of the venn diagram represents the number of non-transcribed pseudogenes involved in causing that particular cancer type or its combination. This distribution is for 1159 non-transcribed pseudogenes.

While we are currently working with all the mutated ncrs, we can always choose to work with any particular combination of ncrs e.g only those involved in individual cancer types or in 2/3/4/all cancer types. The results for all such combinations have been generated and will be used for analysis later on.

c. Assigning 'Pathogenicity' to mutations

Now, we were interested to know which of these mutations were actually significant and which are not. This was done using 'pathogenicity scores' assigned to each of the mutations, called 'FATHMM (Functional Analysis through Hidden Markov Models) scores'. Details about this system is mentioned in the 'Methods and Materials' section. Such scores range between 0 to 1 and mutations above 0.5 are considered deleterious while those above 0.7 are considered outright pathogenic or capable of causing diseases. As mentioned above, we decided to work with scores >0.5 and used this filter to parse out ncrs meeting this condition. Such ncrs were then plotted for their counts of pathogenic mutation against their total number of mutations from all cancer types. While we worked with all ncrs that have atleast 1 pathogenic mutation, one can essentially use this step to filter out ncrs based on number of minimum number of pathogenic mutation (say, atleast 5 pathogenic mutations present) and cancer types (say any combination of cancer as shown in the venn diagrams above).

Fig. 13 shows such a plot for denovo genes. The x-axis indicates the denovo gene ids which are not shown (due to confidentiality) and y axis the number of mutations.

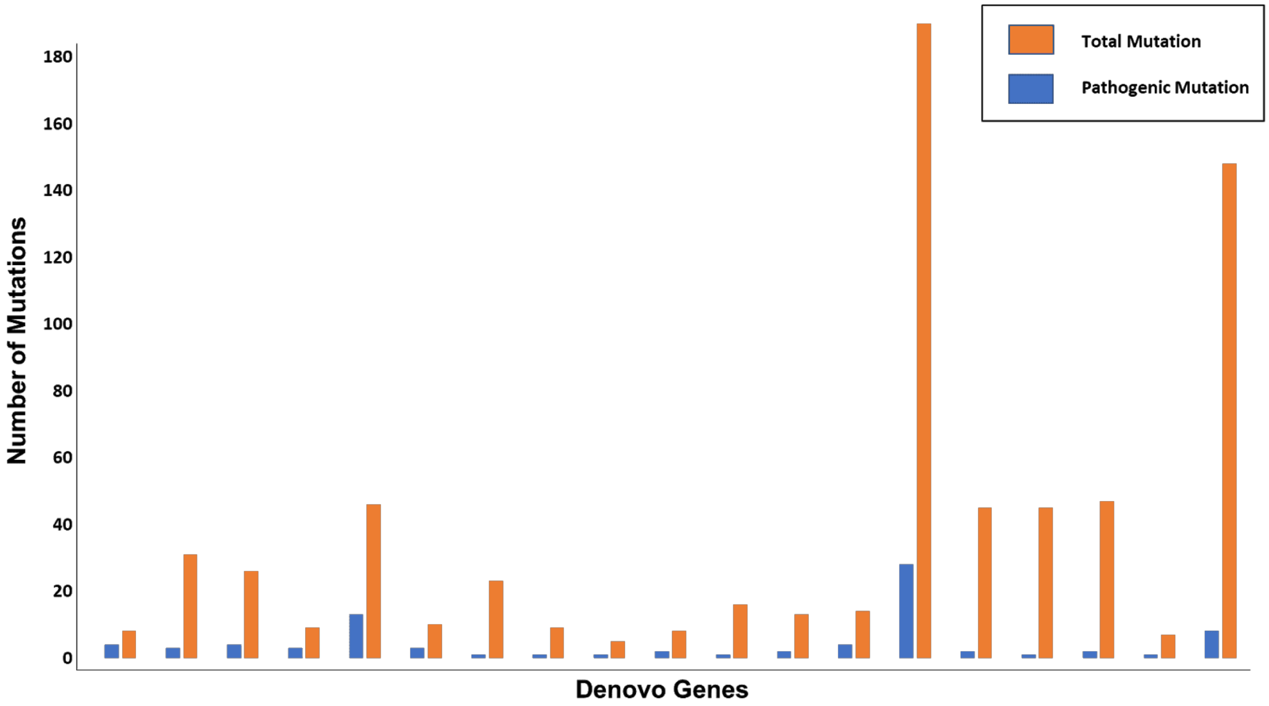


Figure 13: Representation of Pathogenic Mutations vs all Mutations in Denovo Genes (n=19).

The x-axis represents denovo gene ids and y-axis the number of mutations for each denovo gene. Each cluster (blue and saffron bar) is for one denovo gene and shows its corresponding count of pathogenic and total mutation respectively.

Thus, as seen in Fig. 13, out of the 37 mutated denovo genes, 19 has atleast 1 pathogenic mutation to them. Fig. 17 shows how many of these are from which cancer types and is represented as a pie chart.

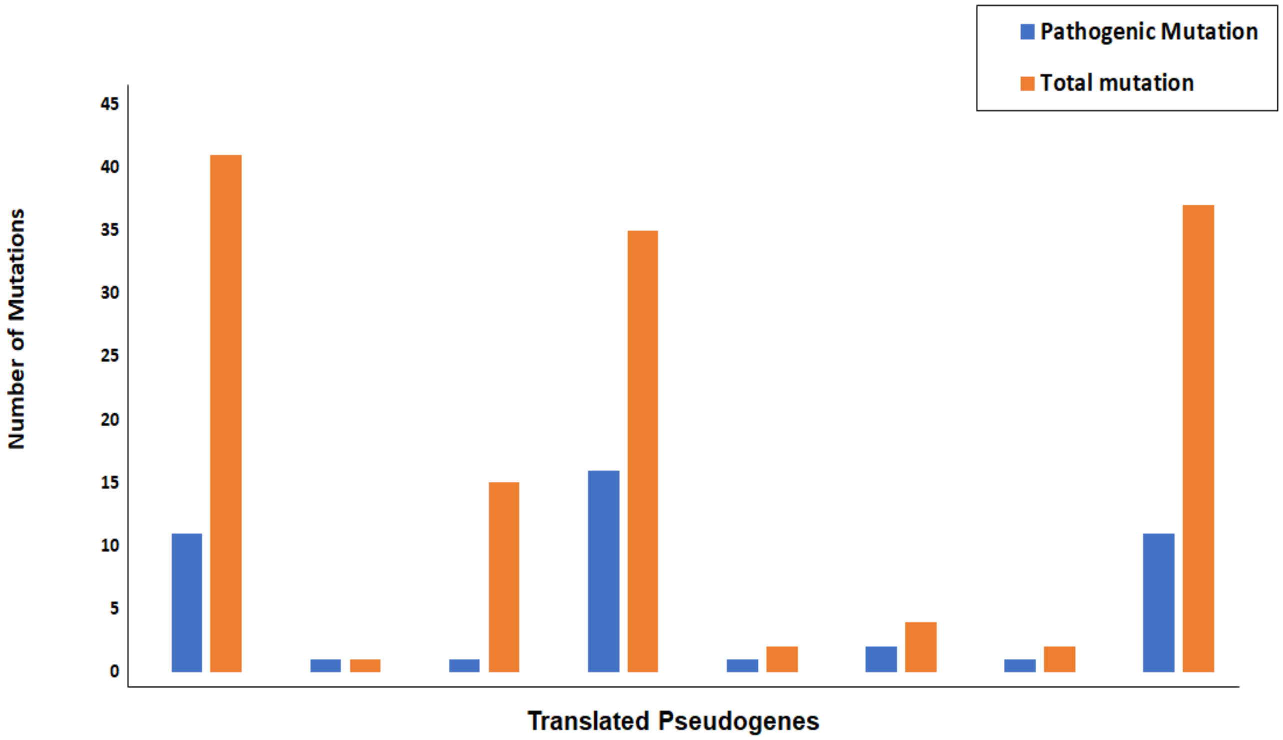


Figure 14: Representation of Pathogenic Mutations vs all Mutations in Translated pseudogenes (n=8).

The x-axis represents denovo gene ids and y-axis the number of mutations for each translated pseudogene. Each cluster (blue and saffron bar) is for one pseudogene and shows its corresponding count of pathogenic and total mutation respectively.

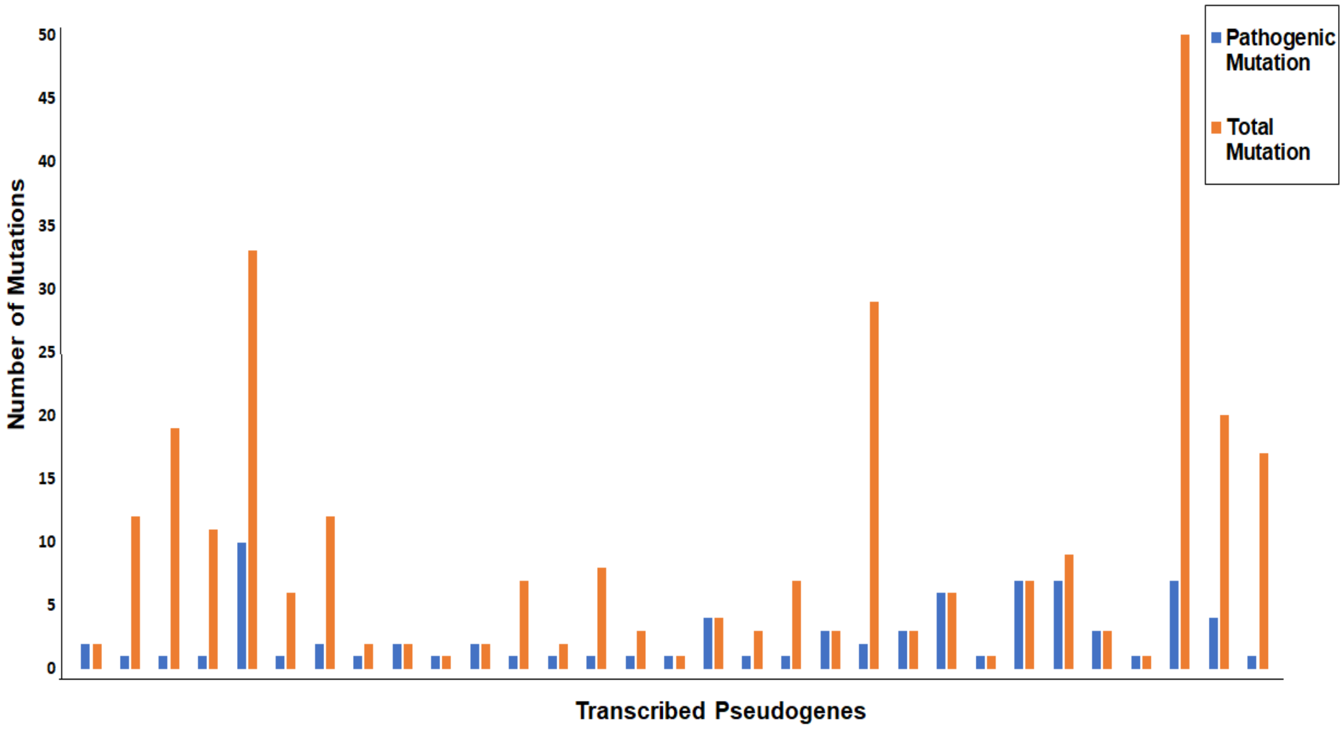


Figure 15: Representation of Pathogenic Mutations vs all Mutations in Transcribed pseudogenes (n=31).

The x-axis represents denovo gene ids and y-axis the number of mutations for each transcribed pseudogene. Each cluster (blue and saffron bar) is for one pseudogene and shows its corresponding count of pathogenic and total mutation respectively.

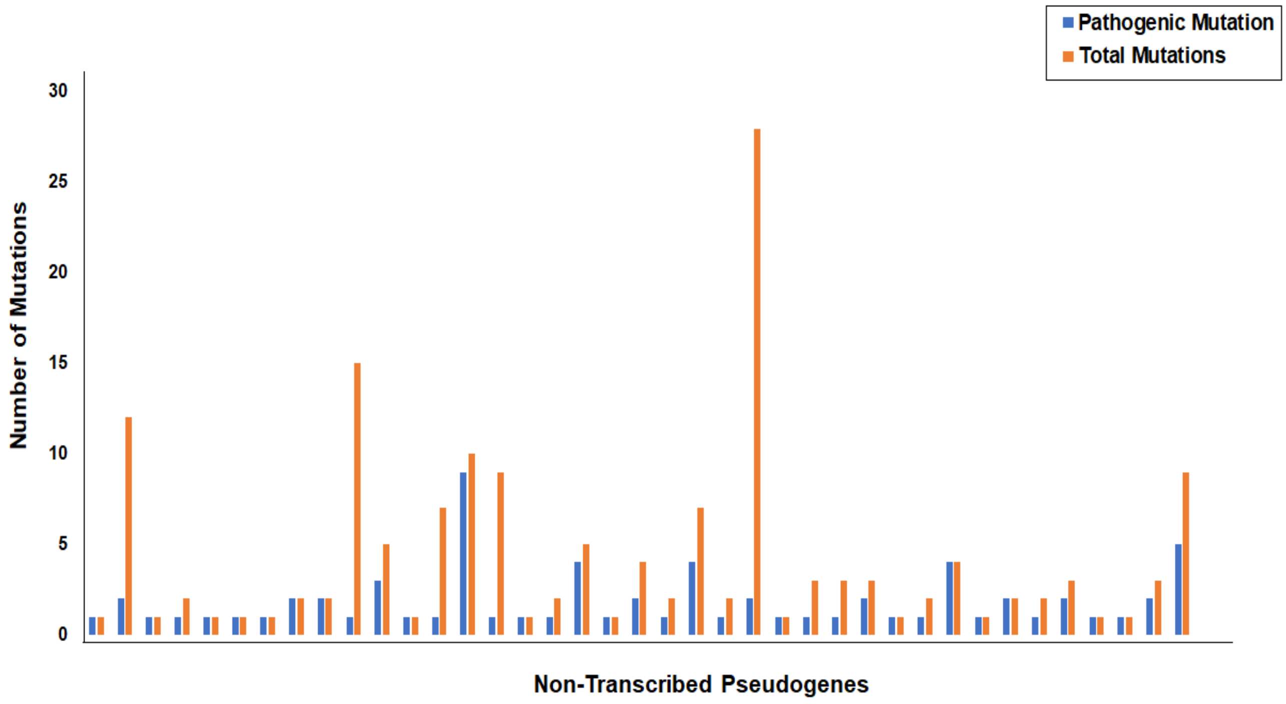
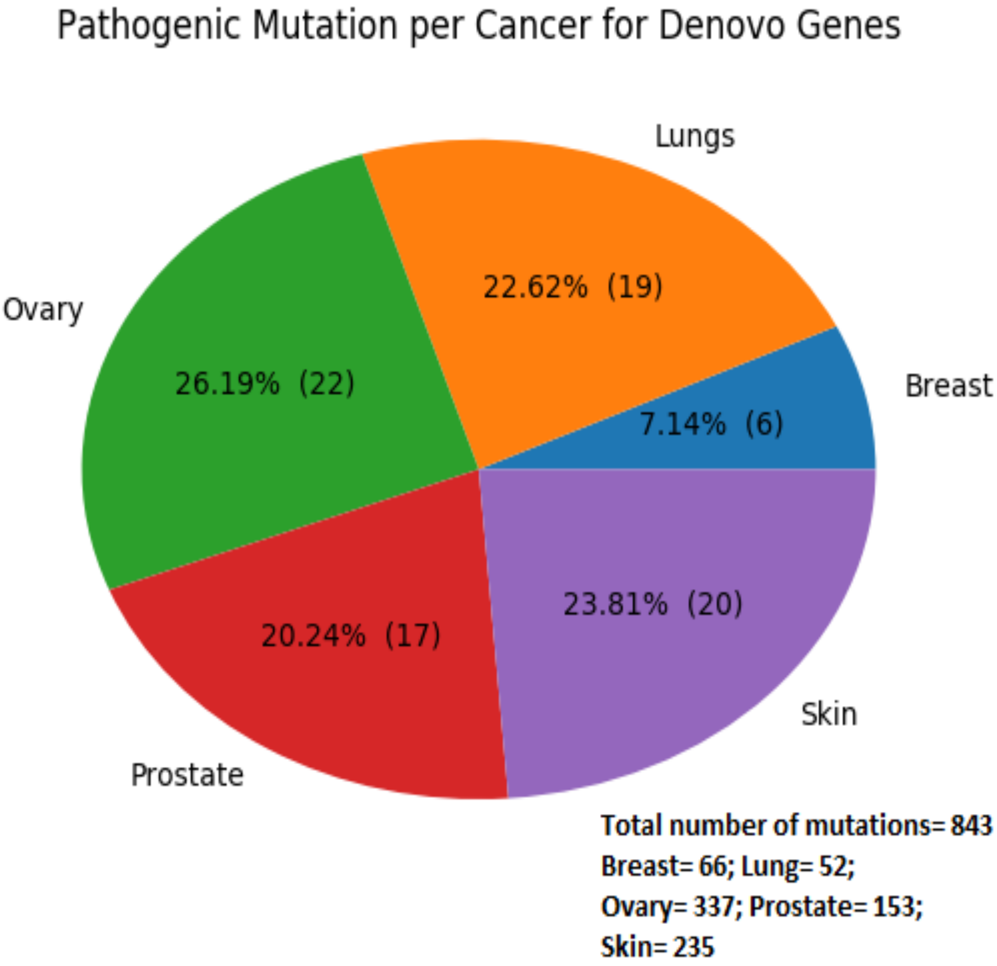


Figure 16: Representation of Pathogenic Mutations vs all Mutations in Non-Transcribed pseudogenes (n=39).

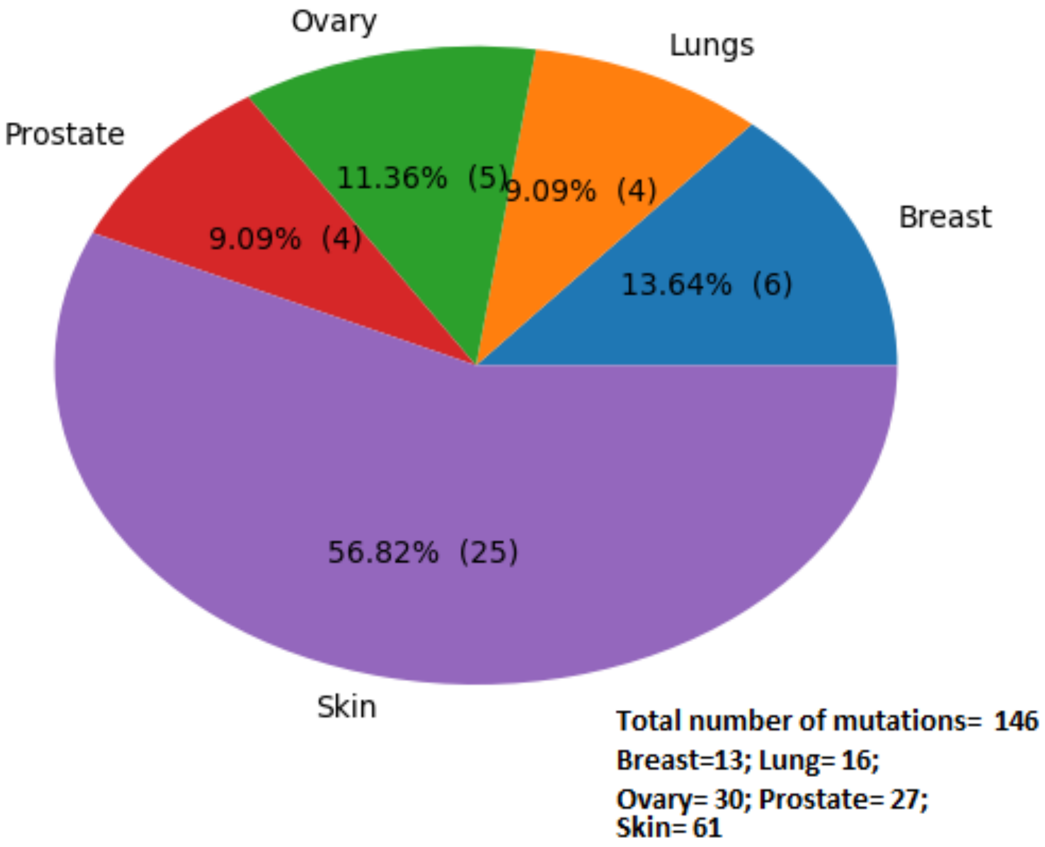
The x-axis represents denovo gene ids and y-axis the number of mutations for each non-transcribed pseudogene. Each cluster (blue and saffron bar) is for one pseudogene and shows its corresponding count of pathogenic and total mutation respectively.



Total number of mutations= 843
Breast= 66; Lung= 52;
Ovary= 337; Prostate= 153;
Skin= 235

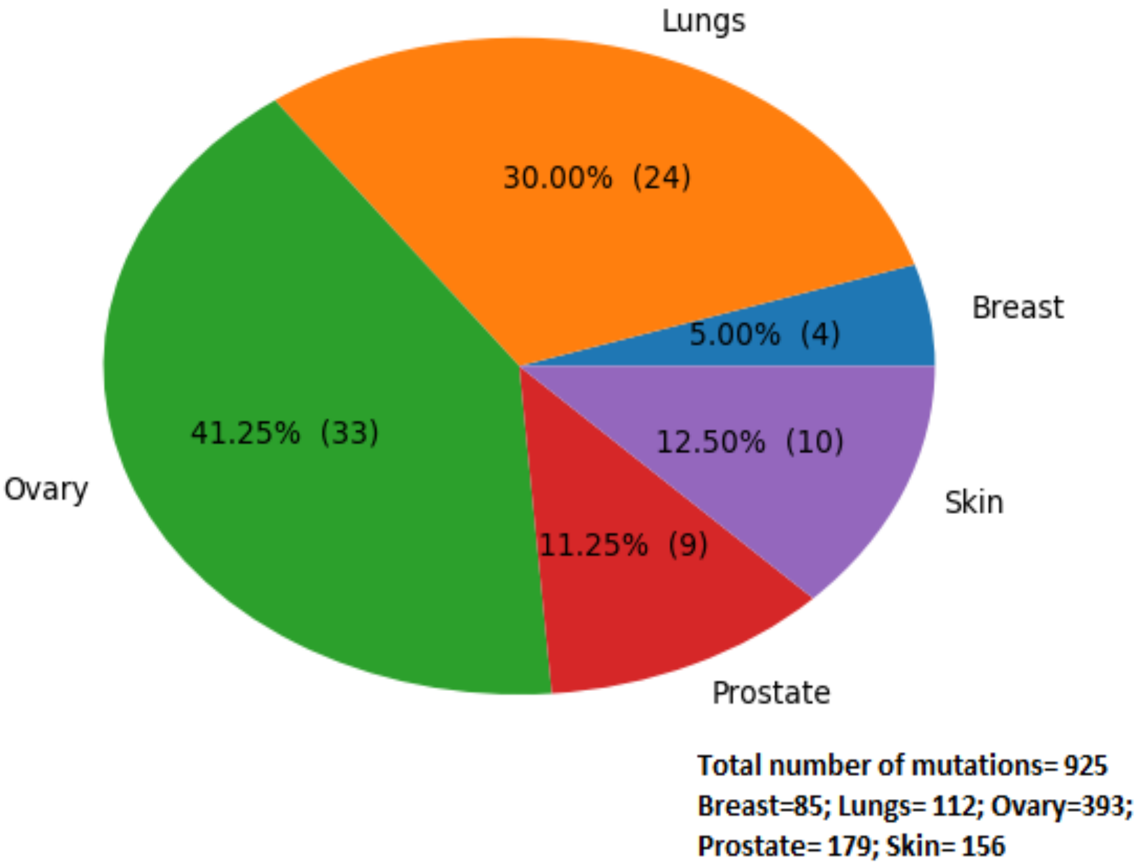
Figure 17: Representation of Pathogenic Mutation per Cancer for Denovo Genes

The values within braces indicate the counts of mutations and is preceded by the percentage value with respect to total number of mutation. For example, the red section indicates that of all mutations (153) involved in prostate cancer, 17 of them (or 20.24%) are pathogenic.



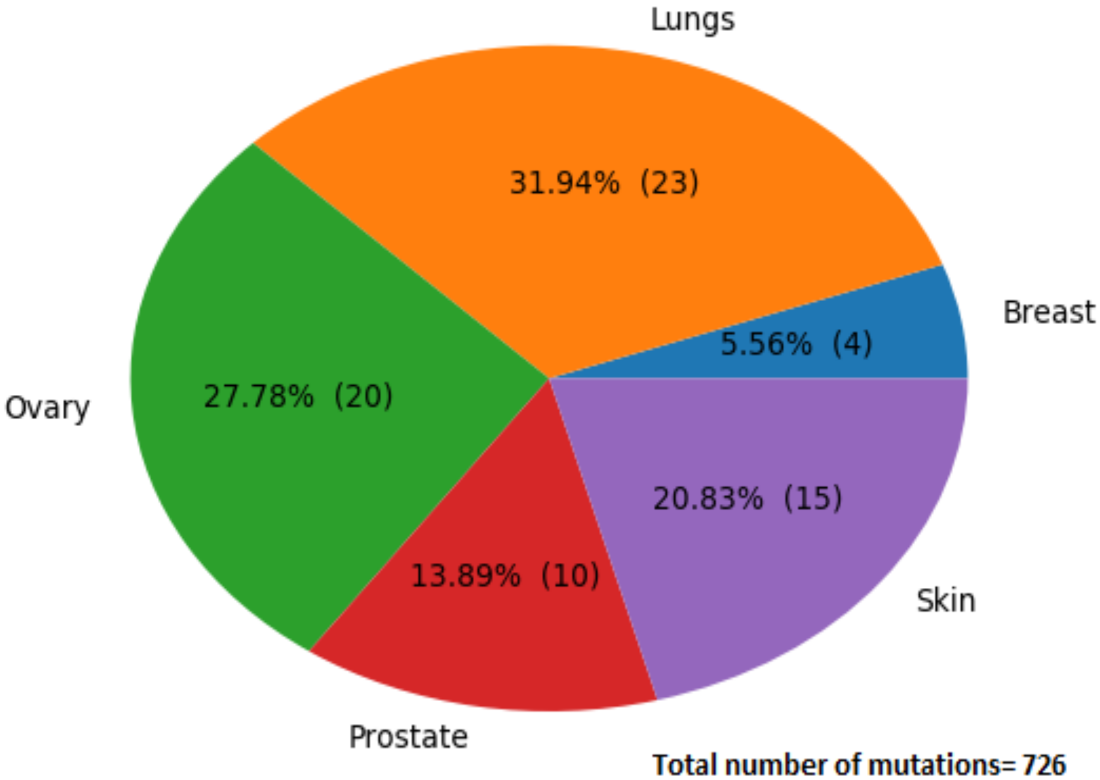
Total number of mutations= 146
Breast=13; Lung= 16;
Ovary= 30; Prostate= 27;
Skin= 61

Figure 18: Representation of Pathogenic Mutation per Cancer for Translated Pseudogenes



Total number of mutations= 925
Breast=85; Lungs= 112; Ovary=393;
Prostate= 179; Skin= 156

Figure 19: Representation of Pathogenic Mutation per Cancer for Transcribed pseudogenes



Total number of mutations= 726
Breast= 70; Lungs= 56; Ovary= 316;
Prostate= 154; Skin= 128

Figure 20: Representation of Pathogenic Mutation per Cancer for Non-Transcribed pseudogenes

d. Length Distribution of ncrs

Since not much work has been done with ncrs (especially those shown here) before, average length of these are not clear. We decided to see their length distribution in order to have an estimation of the average length and use it as another filtration step to select for only those ncrs falling within their average length spectrum and eliminating the rest. Fig.21 shows such a distribution for denovo genes.

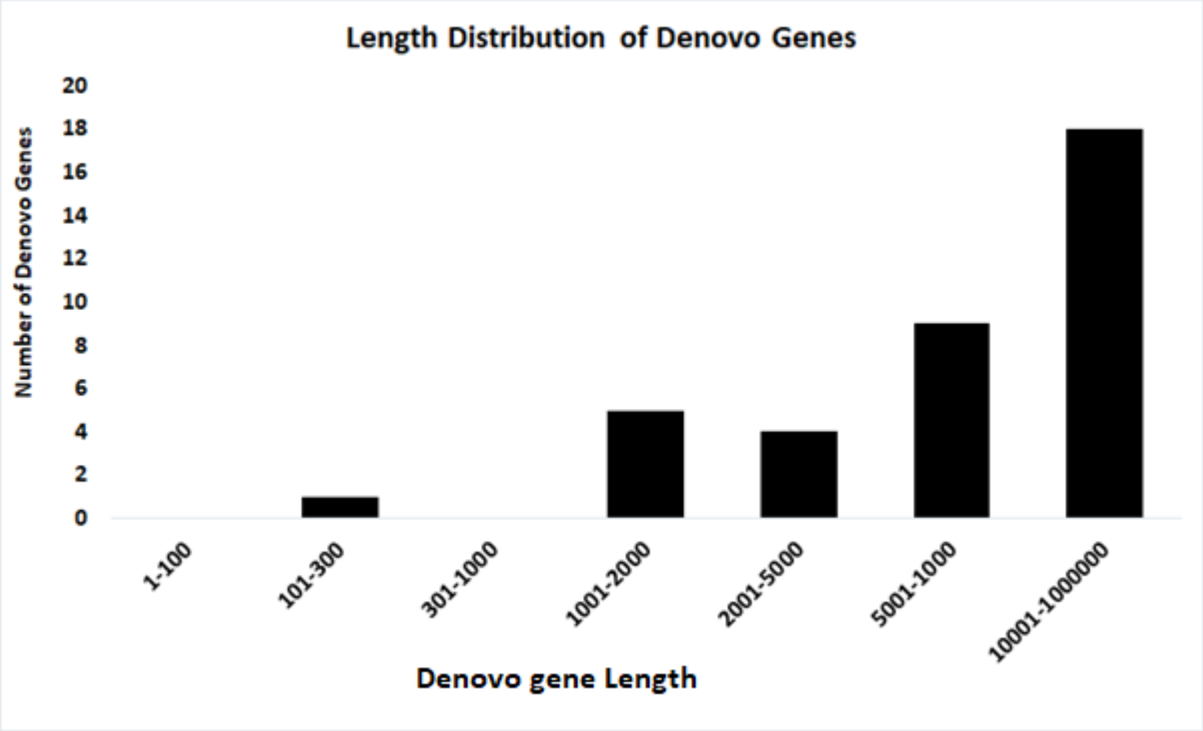
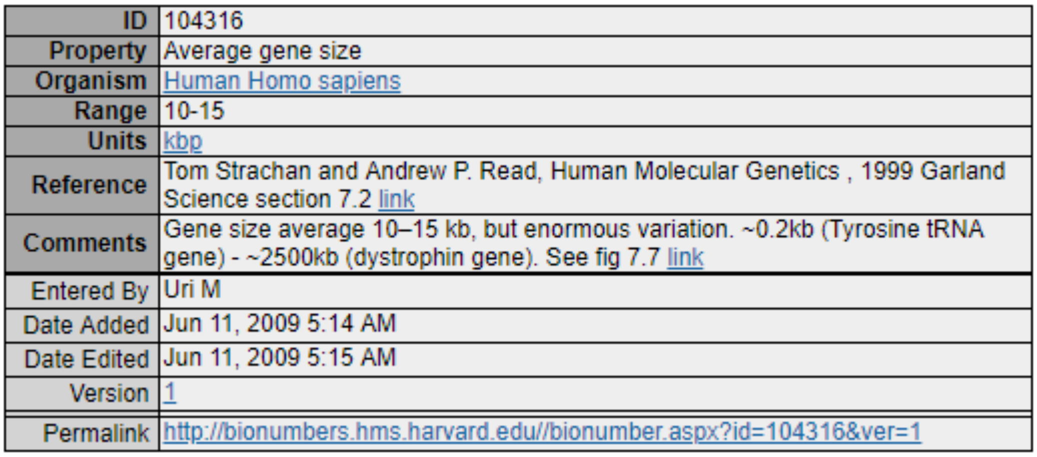


Figure 21:: Representation of length distribution for denovo genes; this is calculated based on all the 42 denovo genes, with an average length of 30746.74 nucleotides. This average was calculated by taking the arithmetic mean of lengths of all 42 denovo genes. Heavy bias towards the last bin indicates technical error in dataset generation.

As seen in the figure above, sizes were broken into discrete bins of continuous intervals and while there is a heavy bias towards the 10k-1000k bin size, that can mostly be due to technical errors occurring during the dataset generation. For a perspective, the average size of a human gene is around 10-15k nucleotides, as shown in Fig.22 below



ID	104316
Property	Average gene size
Organism	Human Homo sapiens
Range	10-15
Units	kbp
Reference	Tom Strachan and Andrew P. Read, Human Molecular Genetics , 1999 Garland Science section 7.2 link
Comments	Gene size average 10–15 kb, but enormous variation. ~0.2kb (Tyrosine tRNA gene) - ~2500kb (dystrophin gene). See fig 7.7 link
Entered By	Uri M
Date Added	Jun 11, 2009 5:14 AM
Date Edited	Jun 11, 2009 5:15 AM
Version	1
Permalink	http://bionumbers.hms.harvard.edu//bionumber.aspx?id=104316&ver=1

Figure 22: Average gene size in humans (Adapted from Strachan and Read)

Ideally, ncrs are of smaller size; sORFS, for example, are defined as orfs with sizes between 100 to 300 nucleotides. Thus, size exclusion was not used as filtration. However, from our dataset, average denovo gene length, eliminating those above 10k range is

Similarly, Figs. 23, 24 and 25 show length distribution for the 3 pseudogene types; average length for each case is 4137 nucleotides for 34 translated, 2848 nucleotides for 539 transcribed and 1160 nucleotides for 1159 non-transcribed pseudogenes.

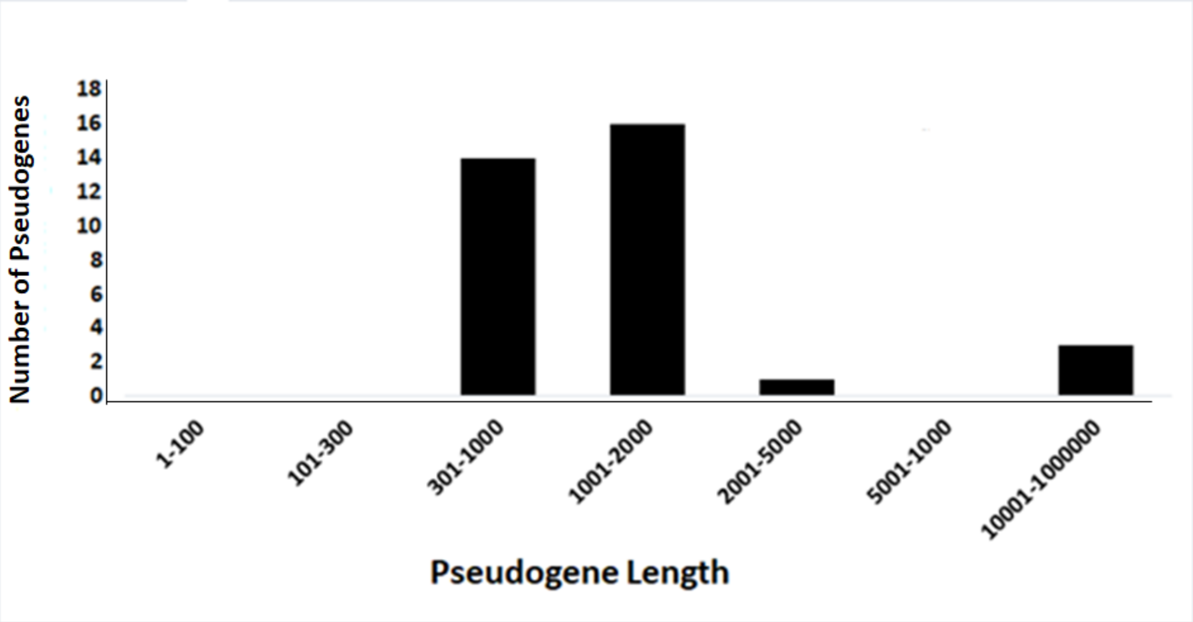


Figure 23: Representation of length distribution for translated pseudogenes; average length= 4137 nucleotides (n=34).

This average was calculated by taking the arithmetic mean of lengths of all 34 translated pseudogenes

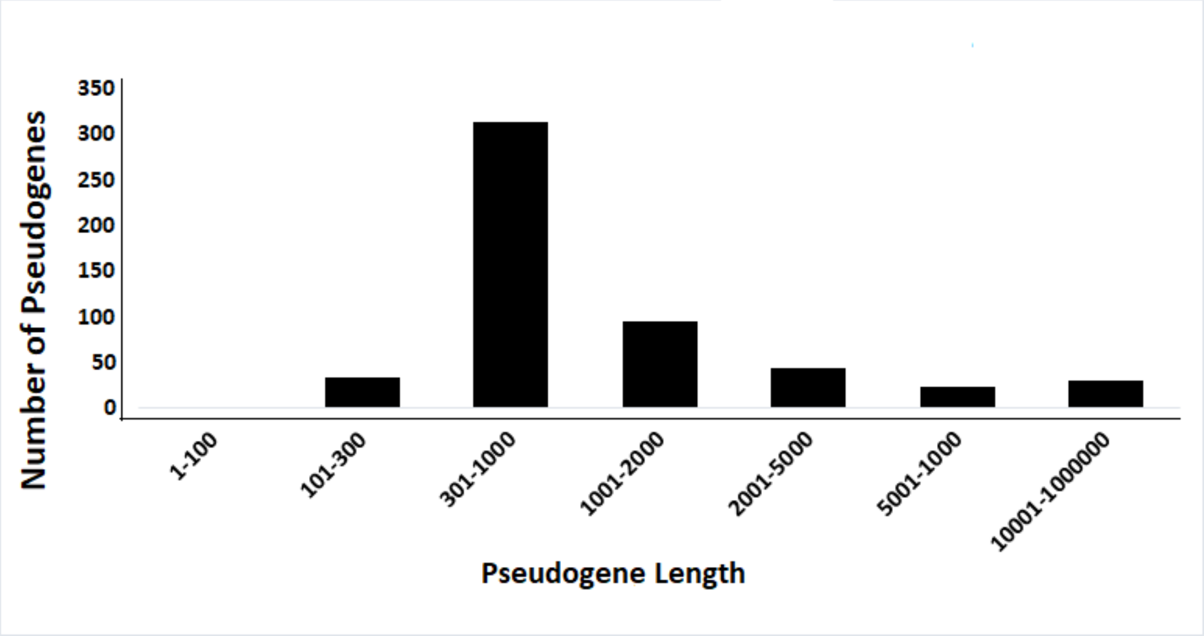


Figure 24: Representation of length distribution for transcribed pseudogenes. average length= 2848 nucleotides (n=539). This average was calculated by taking the arithmetic mean of lengths of all 539 transcribed pseudogenes

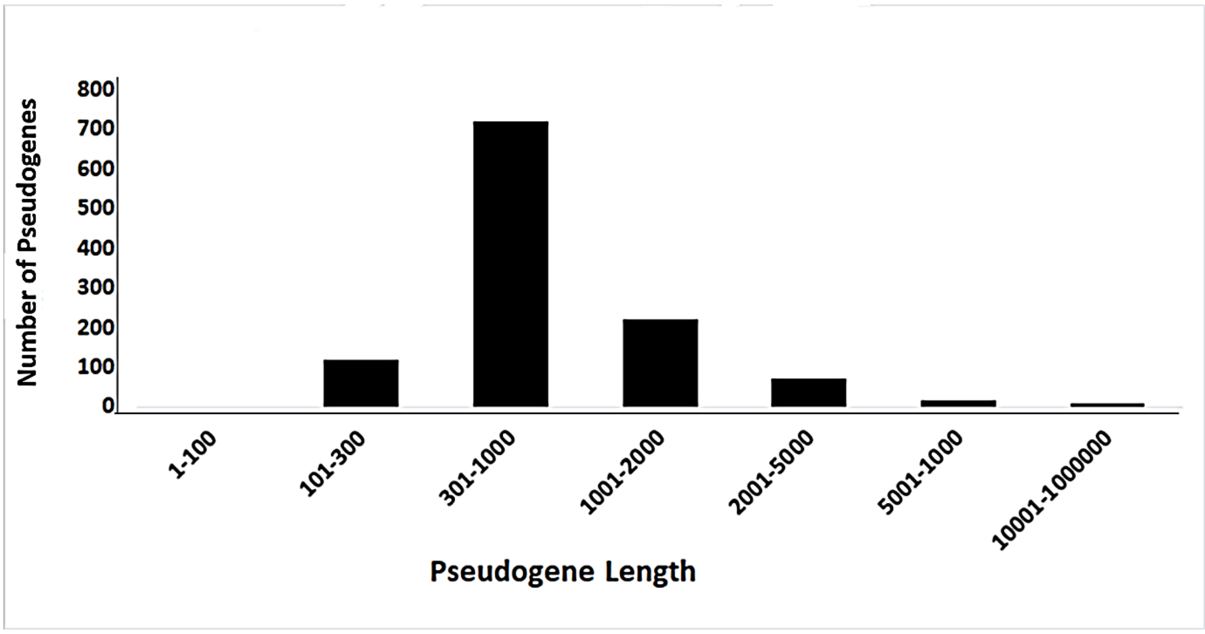


Figure 25: Representation of length distribution for non- transcribed pseudogenes; average length is 1160 nucleotides (n=1159).

This average was calculated by taking the arithmetic mean of lengths of all 1159 non-transcribed pseudogenes

Thus, it is seen that, for pseudogenes, average length can be thought to be (4137+2848+1160)/3 or 2715 nucleotides. This can be used later on as an estimate of average pseudogene length and to select for those whose length is identical to this value; making the filtration process more precise.

e. Gtex Profiles of ncrs

The last step of this project was to look into the Gtex expression profiles of the ncrs that have pathogenic mutations mapped to them. As mentioned in the 'Materials and Methods' section, Gtex is a database providing a list of genotypes collected across tissue samples of different individuals and their expression profiles shown in terms of Transcripts per million or TPM.

Fig. 26 shows the expression profile for denovo genes

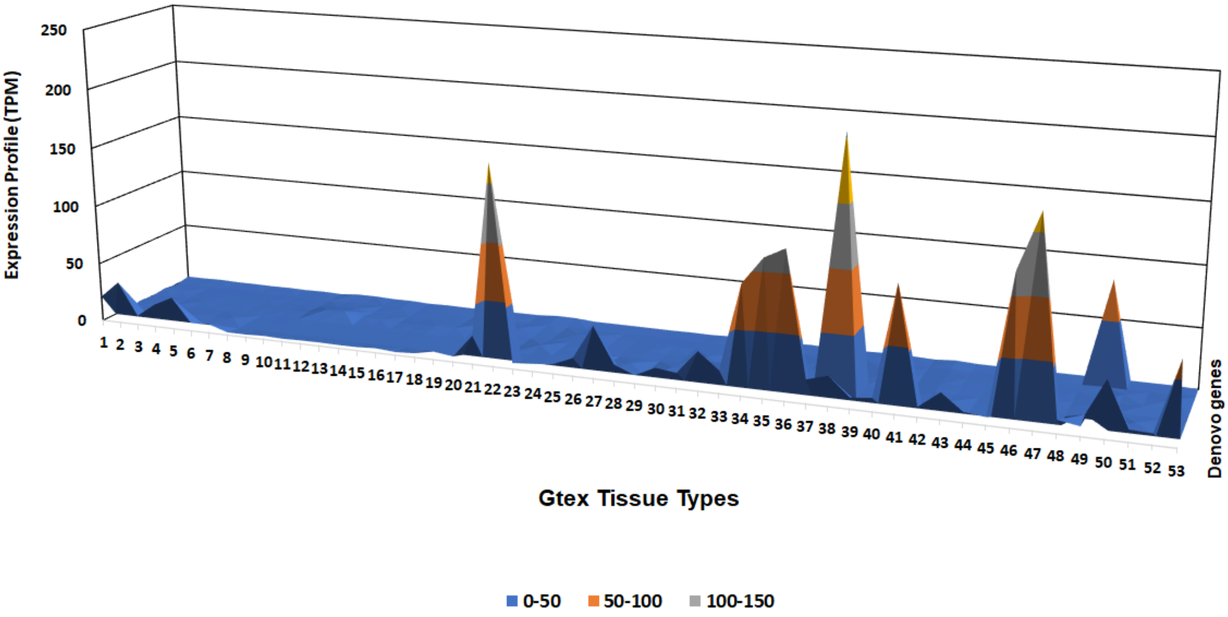


Figure 26: Gtex profile for denovo genes (n=8).

This is a surface plot with the 8 denovo gene ids on 1 axis, gtex tissue ids on another and the corresponding expression profiles of each denovo gene for each tissue type is represented on the third axis

As mentioned above, there were 17 denovo genes with pathogenic mutations. When their ENSEMBL ids were mapped to the Gtex database, we found a positive hit for 8 of them. Their expression levels were plotted.

Similarly, Fig. 27-29 shows expression profiles for the 3 types of pseudogenes.

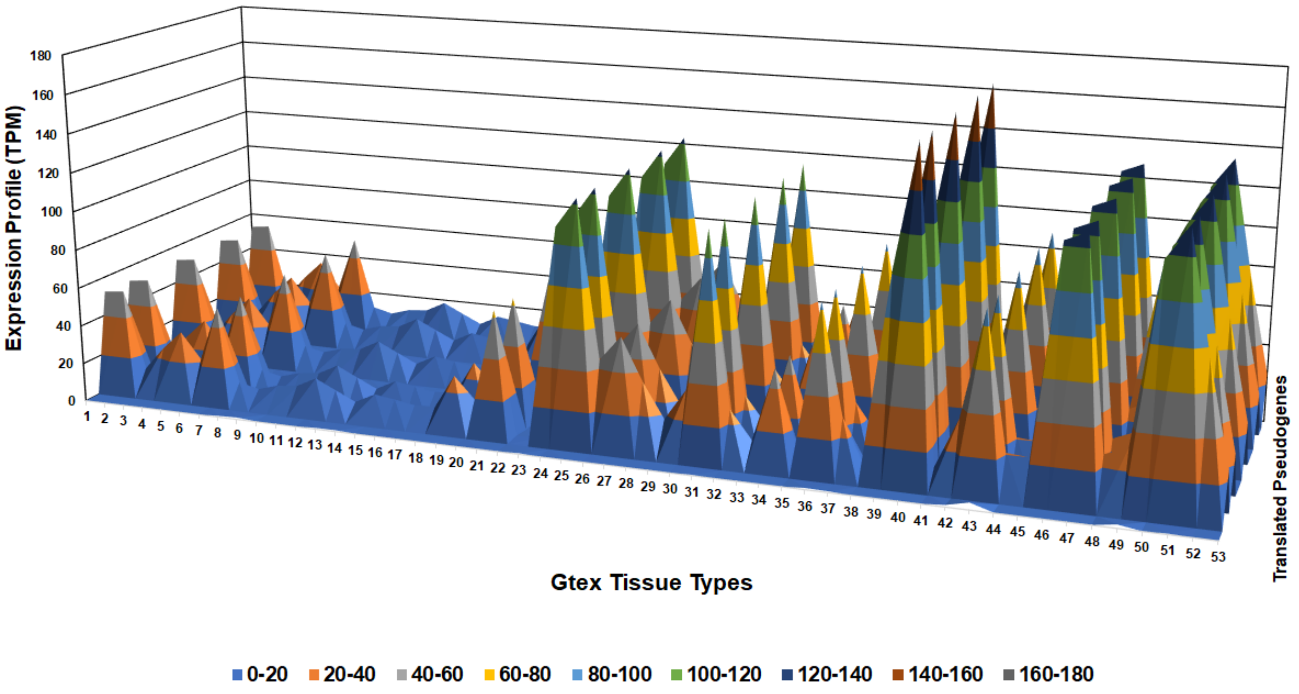


Figure 27: Gtex profile for Translated pseudogenes (n=16).

This is a surface plot with the 16 transcribec pseudogene ids on 1 axis, gtex tissue ids on another and the corresponding expression profiles of each denovo gene for each tissue type is represented on the third axis.

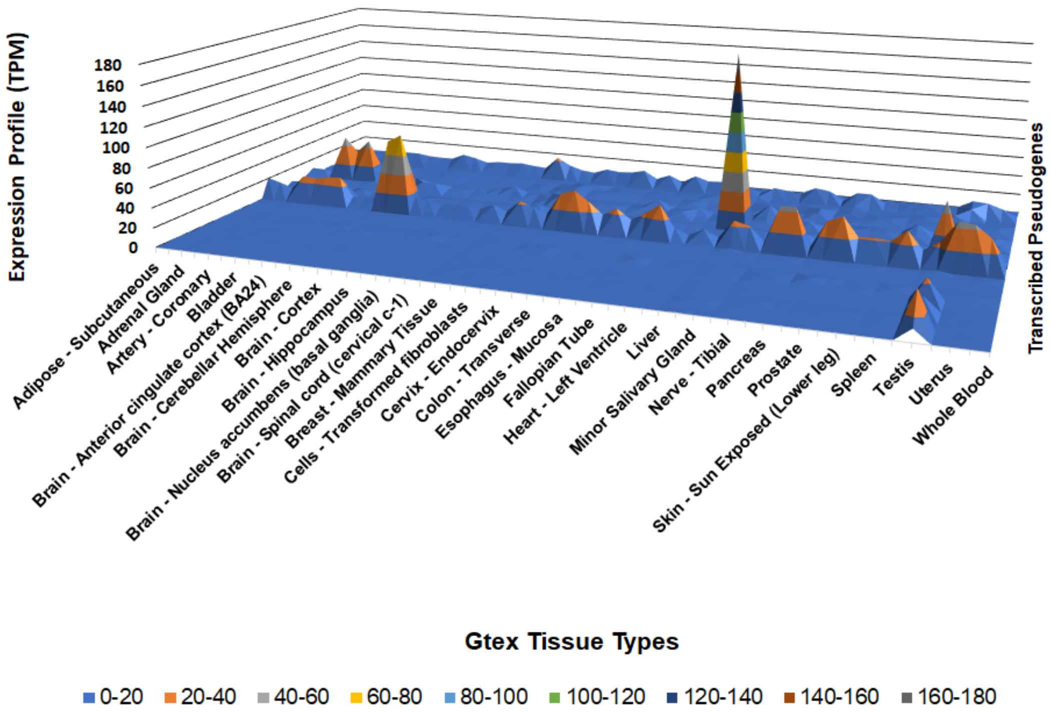


Figure 28: Gtex profile for Transcribed pseudogenes (n=28).

This is a surface plot with the 28 transcribed pseudogene ids on 1 axis, gtex tissue ids on another and the corresponding expression profiles of each pseudogene for each tissue type is represented on the third axis.

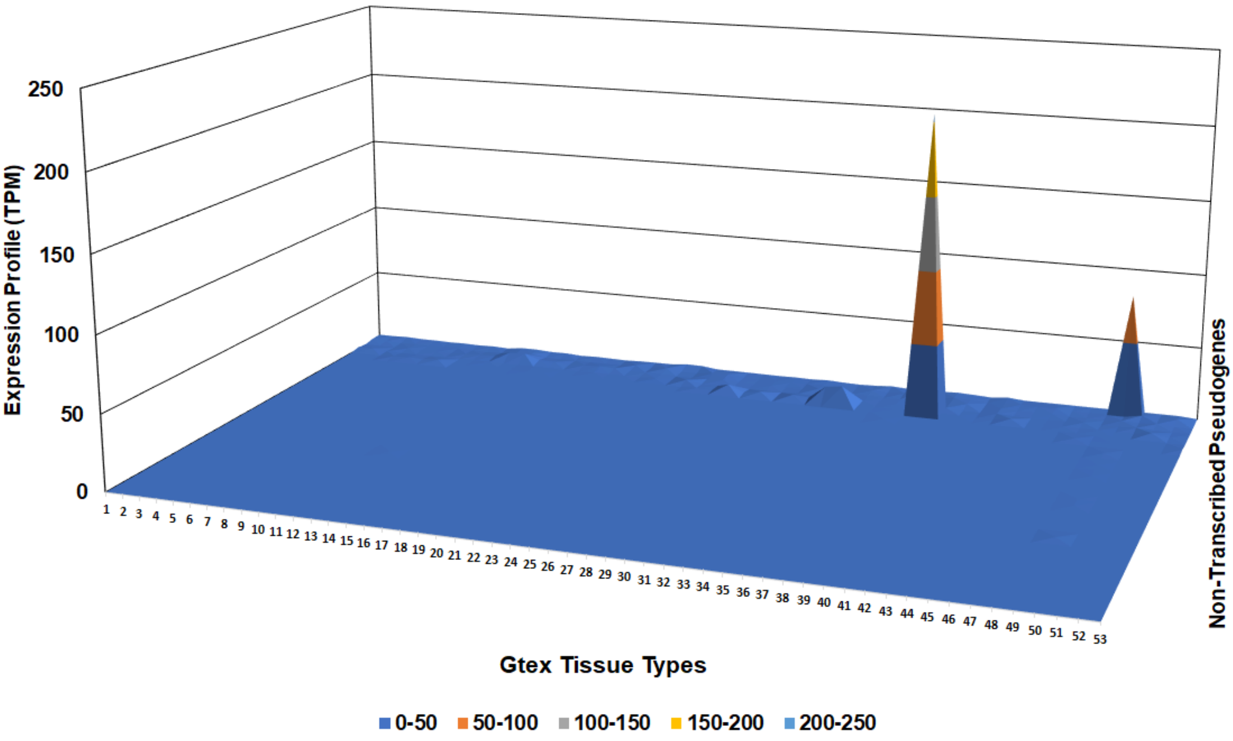


Figure 29: Gtex profile for non-transcribed pseudogenes (n=33).

This is a surface plot with the 33 non-transcribed pseudogene ids on 1 axis, gtex tissue ids on another and the corresponding expression profiles of each pseudogene for each tissue type is represented on the third axis.

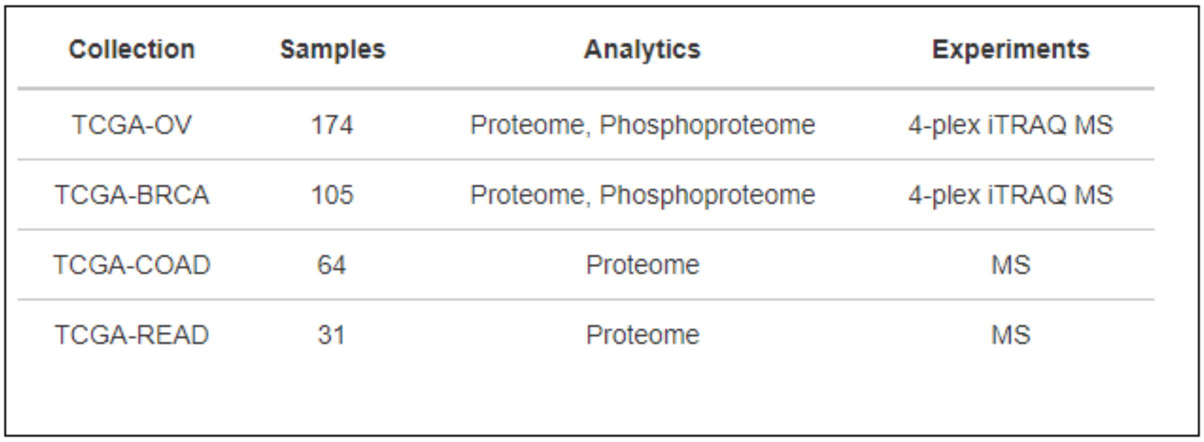
Discussion and Future Perspective

Our study shows that there is an immediate need to start investigating at great depths into the non-coding regions in context of not only cancer but most diseases. The very definition of ‘genes’ and ‘coding’ peptides must be readdressed and ‘micro-peptides’ produced from these ncrs must be investigated with respect to their mutation, functionality, etc.

This thesis mainly tries to focus on the possible correlation between mutation in ncrs and their relevance in cancer, all the while trying to pinpoint ncrs that maybe important for this pathology. While many of these targets show expression in healthy tissues (as seen from Gtex profiling and discussed in section e of ‘Results’), it is unknown if these ncrs and their corresponding peptides are also expressed in their corresponding cancer tissues. This can be achieved using mass-spectrometry(MS) data of peptides from cancer tissue samples.

MS data allows not only identification of proteins and peptides but also quantify their relative abundance and localization of post-translationally modified residues (eg phosphorylated) across biological samples. Recently, The Clinical Proteomic Tumor Analysis Consortium (CPTAC) launched an initiative to understand the molecular basis of cancer (with respect to proteome), using samples from projects like TCGA and analyse their proteomic landscape. The CPTAC dataset includes 4 cancer types- ovarian(TCGA-OV), breast(TCGA-BRCA), rectal(TCGA-READ) and colon (TCGA-COAD). The pipeline used is Common Data Analysis Pipeline (CDAP) [26]

The table below represents a statistical overview of the data for each cancer type and this will be used for our work later on.



Collection	Samples	Analytics	Experiments
TCGA-OV	174	Proteome, Phosphoproteome	4-plex iTRAQ MS
TCGA-BRCA	105	Proteome, Phosphoproteome	4-plex iTRAQ MS
TCGA-COAD	64	Proteome	MS
TCGA-READ	31	Proteome	MS

Figure 30: Statistics of CPTAC data

[Note: the pipeline was validated using mass spectrometry dataset derived from yeast whole cell lysates]

These data will be used to match cases genomic and transcriptomic data of atleast 2 cancer types (breast and ovary) which were used in this study and to derive a possible correlation between target ncrs with respect to particular tissue types, their Gtex profiles and their corresponding CPTAC proteome profile, if any. This will help us understand how many of our target ncrs are actually present in cancer tissue samples.

Once this pipeline is established, our lab will be collecting its own samples in collaboration of bio-banks across India and generate our own genomic, transcriptomic and proteomic data using actual experimental techniques (both novel and standardized) and analyse them as outlined above. This ‘Systems Proteogenomic’ approach will help identify some key ncrs and their corresponding mutated peptides and their functions that may give rise to a particular cancer or cancers.

India is a country with high burden of both communicable and non-communicable diseases and large diversity of patients with respect to (disease) genotype and phenotype [25]. As such when it comes to complex disease like cancer, designing general therapeutics against a diverse population does not guarantee cure in all cases. Thus, a necessity to analyse this diverse germplasm and design 'personalized'/customized therapeutics based on individual genomic and/or environmental background is more important than ever and this workflow is a stepping stone for such an initiative.

A brief overview of such approaches is summarized in Fig.31 below.

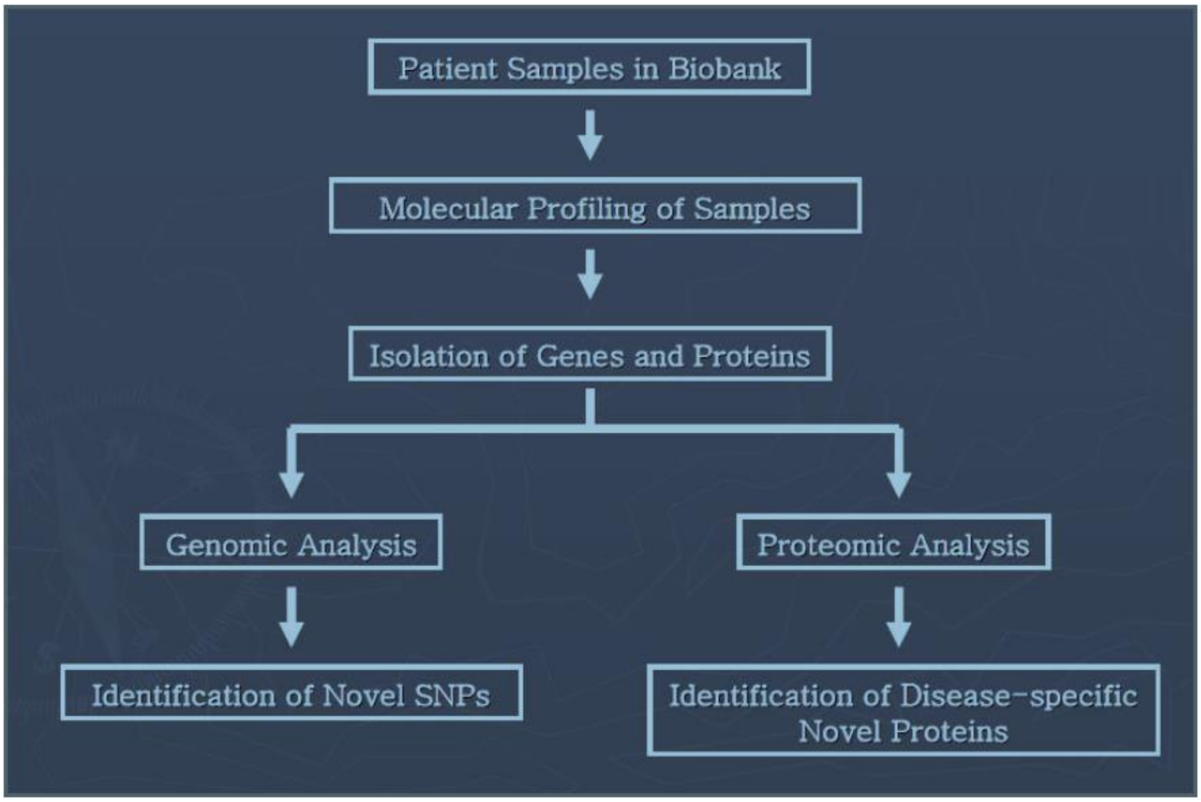

Patient Samples in Biobank
Molecular Profiling of Samples
Isolation of Genes and Proteins
Genomic Analysis
Proteomic Analysis
Identification of Novel SNPs
Identification of Disease-specific Novel Proteins

Figure 31: A bridge between research and therapeutics [Adapted from N.K. Ganguly].

How such cross-platform work can be done has been described in details in the texts above

In developing countries like India, where there is a huge diversity of disease genotypes and phenotypes, success of this kind of approach will allow personalized treatment in a cost-efficient manner that will simultaneously increase the state of healthcare in the country and decrease the high cost of obtaining the same.

One of the last steps of this project will be to predict the mutated structures of the peptides in silico and to identify the same in vivo. Bioinformaticians in our lab are already working on the in silico predictor algorithms for the different ncrs mentioned here. The last step will be to predict the change in behaviour of these peptides and which pathways they are involved in and how their change can possibly affect their biological interaction that might give rise to cancer, though how this will be achieved is yet to be pipelined. This will then be applied to target peptides obtained invivo and the complete nature of mutations in ncrs and their influence on cancer initiation and progression can be understood. Such data, in collaboration with those already present on mutation in known protein coding regions will help provide a complete picture of the cancer genome and its diversity and help us eradicate the disease in future.

It will be interesting to identify similar candidates for ncrs that do not produce peptides but work at transcript level (say rna-mediated silencing) and understand their contributions as well.

Also, another necessary step will be to write new algorithms for pathogenicity classifier, using machine learning and training them on non-coding pathogenic mutations. All classifiers till date mostly train the algorithms on coding pathogenic mutations and this situation needs to be addressed in order to make the pipeline more precise.

Overall, non-coding region biology is a rapidly developing field and understanding it at different levels is essential to address effective therapeutics against several diseases. Hopefully, this pipeline will be one of the first of its kind and can help the research community and medical community alike, thus improving the lifestyle of the common people.

Reference:

1.Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. Nature medicine 10, 789-799 (2004).

2. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. Nature 458, 719-724 (2009).

3. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. Cell 144, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

4.
Santarius, T., Shipley, J., Brewer, D., Stratton, M. R. & Cooper, C. S. A census of amplified and overexpressed
human cancer genes. Nat Rev Cancer 10, 59-64, doi:10.1038/nrc2771 (2010).

5. Futreal, P. A. et al. A census of human cancer genes. Nat Rev Cancer 4, 177-183, doi:10.1038/nrc1299 (2004)

6. Pleasance, E. D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463, 191-196, doi:10.1038/nature08658 (2010)

7. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary oncology 19, A68 (2015).

8. Prabakaran, S. et al., 2014. Quantitative profiling of peptides from RNAs classified as noncoding. Nature communications , 5, p.5429.

9. Parkin, D. M., Bray, F., Ferlay, J., & Pisani, P. (2005). Global Cancer Statistics, 2002. CA: A Cancer Journal for Clinicians, 55(2), 74–108. https://doi.org/10.3322/canjclin.55.2.74

10. https://www.everydayhealth.com/g00/cancer/know-the-most-common-types-of-cancer.aspx?i10c.encReferrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvLmluLw%3D%3D&i10c.ua=1&i10c.dv=13

11. https://www.cancer.gov/types/common-cancers

12. Britten, R. J. & Kohne, D. E. Repeated sequences in DNA. Science 161, 529–540 (1968)

13. Patrushev, L. I., & Kovalenko, T. F. (2014). Functions of noncoding sequences in mammalian genomes. Biochemistry (Moscow), 79(13), 1442–1469.

14. St Laurent, G., Vyatkin, Y. & Kapranov, P., 2014. Dark matter RNA illuminates the puzzle of
genome-wide association studies. BMC medicine , 12, p.97

15. Clark, M. B., Amaral, P. P., Schlesinger, F. J., Dinger, M. E., Taft, R. J., Rinn, J. L., Ponting, C. P., Stadler, P. F., Morris, K. V., Morillon, A., et al. (2011) The reality of pervasive transcription, PLoS Biol., 9, e1000625.

16. http://cancer.sanger.ac.uk/cosmic/download

17. http://book.bionumbers.org/how-big-is-the-average-protein/

18. Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., … Kageyama, Y. (2010). Small Peptides Switch the Transcriptional Activity of Shavenbaby During Drosophila Embryogenesis. Science, 329(5989), 336–339. https://doi.org/10.1126/science.1188158

19. Magny, E. G., Pueyo, J. I., Pearl, F. M. G., Cespedes, M. A., Niven, J. E., Bishop, S. A., & Couso, J. P. (2013). Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames. Science, 341(6150), 1116–1120. https://doi.org/10.1126/science.1238802

20. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling Volodimir Olexiouk; Wim Van Criekinge and Gerben Menschaert Nucleic Acids Research 2017; doi: 10.1093/nar/gkx1130
21. Hao Y., et al. 2017.SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci.Brief Bioinform bbx005
22. Vanderperre, B., Lucier, J.-F., & Roucou, X. (2012). HAltORF: a database of predicted out-of-frame alternative open reading frames in human. Database, 2012(0), bas025-bas025. https://doi.org/10.1093/database/bas025
23. Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., … Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics, 31(10), 1536–1543. https://doi.org/10.1093/bioinformatics/btv009
24.https://www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id=phs000424.v2.p1
25. World Health Organization (WHO), India: WHO Statistical Profile 2015
26. https://cptac-data portal.georgetown.edu/cptac/aboutData/show?scope=dataLevels