Proposal title: **Multi-omic analysis to identify the mutational landscape, biomarkers and novel drug targets for Pancreatic cancer of Indian cohorts**

Principal Investigator: **Dr. S. Mahalingam,** Professor, Department of Biotechnology, Indian Institute of Technology Madras, Chennai, India

Introduction:

Cancer is one of the fatal health problems faced by humanity today. The limited survival of cancer patients is likely due to a high proportion of patients with advanced disease stages, a lack of suitable markers for early detection, and failure to respond to available therapy. There is a growing need to identify and characterize the mechanism of tumor initiation/progression, as well as the efficacy of different drugs on individual tumors. In addition, due to the heterogeneity of cancers among different populations, population-specific approaches are critical to managing cancer in specific populations. With the rapid growth in knowledge of the molecular basis of cancer progression and evolution, biomedical research and the health-care system in India need to shift toward a vision of "personalized medicine", which may improve the standard of medical care by including an individual's genetic and molecular information in the clinical decision-making process.

Pancreatic cancer, specifically pancreatic ductal adenocarcinoma (PDAC), is an important public health problem and has become the 4th leading cause of cancer-related death worldwide, with little improvement in outcomes despite decades of research. Survival rates are among the worst for any cancer, with the mortality to incidence ratio being 98%. Its occult nature and the lack of non-invasive sensitive biomarkers result in diagnosis often after the cancer has advanced locally to the point of being non-resectable or has metastasized to distant sites. As current therapeutic results are so dismal, it is essential to identify the genetic factors that influence the development of this disease to implement preventive and screening strategies that can reduce the burden of this lethal cancer. Apart from a few exceptions, most clinical trials in PDAC have failed to demonstrate a clinically meaningful survival benefit. This is perhaps not surprising because of the non-availability of genomic information. This hampers clinical trial efficiency, as the responsive phenotype of a therapeutic regimen would fall below the detection threshold of most conventional treatment. Surgery remains the only chance of cure, yet only less than 10% of patients will be alive at five years after pancreatic resection. Few chemotherapeutics provide some improvement in outcome, and even then, for approved therapies, the survival benefits are marginal. In India, like most countries in the world, there is no screening program for the early detection of pancreatic cancer. Furthermore, symptoms related to pancreatic cancer tend to be nonspecific, including weight loss, abdominal pain, nausea, and dyspepsia. This may explain the failure of conventional clinical trial designs to show any meaningful survival benefit, except in small and undefined patient subgroups. Both of the above contribute to the late presentation of the cancer and its notoriously poor outcomes. Consequently, there is an urgent need to identify PDAC specific biomarkers and drug targets to develop novel therapeutic approaches that leverage treatment selection for patients with PDAC in India.

Genomic sequencing studies of pancreatic cancer have revealed a small set of consistent mutations found in pancreatic cancers in the western population, which may not apply to the Indian population due to genetic heterogenicity. With the development of next-generation sequencing technology, genomic sequencing and analysis can be performed to identify therapeutic targets and biomarkers in individual patients and personalize treatment selection for Indian cohorts. Incorporating preclinical discovery and molecularly guided therapy into clinical trial design has the potential to significantly improve outcomes in this lethal malignancy. In this proposal, we plan to carry out genomic sequencing (Whole exome and transcriptome) of Indian pancreatic cancers and identify potential biomarkers for early detection and drug targets for therapeutic development.

State of scientific knowledge:

Pancreatic cancer (PDAC) is the fourth leading cause of cancer death worldwide and is projected to be the second within a decade. Advances in therapy have only achieved incremental improvements in the overall outcome but can provide notable benefit for undefined subgroups of patients. Because symptoms of patients with early-stage PDAC are uncommon and nonspecific, there are many challenges in the early detection of PDAC in clinical practice. Carbohydrate antigen 19–9 (CA 19–9) is the most widely used biomarker for PDAC. However, because of its relatively low sensitivity and specificity (70–90% and 68–91%, respectively, for diagnosis of PDAC), CA19-9 is not an ideal biomarker for screening and early detection of PDAC and its main clinical application is as a marker for monitoring progression and response to therapy. At present, there are no successful detection methods that are effective and non-invasive. Imaging tests, including US, CT, MRI, EUS, ERCP, and FDG-PET, have their respective advantages and are involved in all aspects of clinical management of PDAC. But when lesions are not identified by imaging, molecular approaches are the only solution for improving early diagnosis. The lack of any specific genomic diagnostic markers, the difficulty in establishing a tissue diagnosis, and the aggressive nature of pancreatic adenocarcinomas, which respond poorly to both chemotherapy and radiotherapy, contribute to the exceptionally high mortality associated with this type of cancer. As a consequence, there is an urgent need to better understand the molecular pathology of PDAC and the identification of biomarkers and novel drug targets for early diagnostics and to develop novel therapeutic strategies.

Genomic analyses have improved our understanding of the complex molecular pathology of PDAC. Studies are revealing molecular subsets of patients that may have durable responses to specific therapies, and strategies are being developed to test these assertions. Treatment resistance, however, remains a significant problem even in those that respond initially. Successful translation of large-scale genomic discoveries also requires novel clinical approaches to develop and incorporate personalized medicine into PDAC to improve outcomes in this lethal disease. There is still a long but promising journey between these novel biomarkers and their translation to implementation in clinical routine for patients with PDAC. It remains undetermined how to translate omics techniques and omics information into the prediction and early diagnosis for this deadly malignancy. Despite having reasonable genomic data available for PDAC from Western societies (the US and Australian pancreatic cancer consortia), there are no reliable biomarkers identified for early detection of PDAC. Furthermore, to the best of our knowledge, there is no genomic information available for PDAC from the Indian population. It is well-known that genomic data and the pathways involved in cancer initiation are significantly different between populations. Because of the above reasons, the success rate of PDAC (most cancers) from India is very low compared with Western populations and these data suggest that there is an urgent need to identify more effective and specific biomarkers based on omics information for patients with early-stage PDAC for Indian cohorts to allow timely and curative treatment and monitor recurrence after surgery.

Multiple translational research studies have explored minimally or non-invasive biomarkers in body fluids such as blood, urine, stool, saliva, or pancreatic juice, but their diagnostic performance has not been further validated. Such markers include genetic alterations, epigenetic changes, dysregulated miRNAs, metabolic profiles, ctDNA, and exosomes. These previous studies represent impressive efforts and have provided large numbers of very valuable and promising novel markers that have led to great enlightenment regarding the biology of PDAC. However, due to the limitations of detection techniques and sample quantity and tumor heterogeneity, larger studies and further validation is required for better diagnostics. To date, some biomarkers like miRNA have been superior to CA19-9 in sensitivity and specificity. Due to the lack of symptoms, early PDAC is difficult to detect for a majority of patients. An ideal biomarker should provide a definitive diagnosis of the presence or absence of tumor, provide correct and definitive staging of the disease, and identify the very early stages of a lesion. The majority of published data describe the diagnostic value of biomarkers for distinguishing patients with PDAC from those with non-cancerous diseases, other digestive tumors, and healthy volunteers. Better diagnostic methods are urgently needed. In our opinion, genomic biomarkers will have broad applications once the detection technology is simplified.

We propose to perform a comprehensive, integrated genomic analysis of 50 pancreatic cancers with matched normal tissue samples and their histopathological variants using a combination of whole-genome exome and RNA sequencing, with gene copy number analysis to determine the mutational mechanisms and candidate genomic events important in pancreatic carcinogenesis, and RNA expression analysis to define subtypes and the different transcriptional networks that underpin them. The goal of this integrated analysis of whole exome and transcriptomic data is to extract biological insights and potentially novel diagnostic biomarkers.

Identification of key elements and objectives of the proposal:

Pancreatic cancer [Pancreatic ductal adenocarcinoma (PDAC)] is an important public health problem and is the fourth leading cause of cancer death worldwide with little improvement in outcomes despite decades of research. It is among the cancers with very poor survival rates. Due to the lack of non-invasive sensitive genomic biomarkers, result in diagnosis often after the cancer has advanced locally to the point of being non-resectable or metastasized to distant sites. At present, there are no specific successful non-invasive detection methods for PDAC. We propose to carry out the following:

1. Comprehensive, integrated genomic analysis of pancreatic cancers with matched normal tissue samples from Indian cohorts using a combination of whole-genome exome and RNA sequencing.
2. To identify specific biomarkers for early detection and novel drug targets to design better therapeutics.

Work plan

Aim: 1. Comprehensive, integrated genomic analysis of pancreatic cancers with matched normal tissue samples from Indian cohorts using a combination of whole-genome exome and RNA sequencing.

Rationale: Cancer is one of the fatal health problems faced by mankind today. The limited survival of cancer patients is likely due to a high proportion of patients with advanced disease stages, a lack of suitable markers for early detection, and failure to respond to available therapy. There is a growing need to identify and characterize the mechanism of tumor initiation/progression, as well as the efficacy of different drugs on individual tumors. In addition, due to the heterogeneity of cancers

among different populations, population-specific approaches are critical to managing cancer in specific populations. There is an urgent need to address the lack of cancer genomic data specific to pancreatic cancer from the Indian population. With the rapid growth in knowledge of the molecular basis of cancer progression and evolution, biomedical research and the health-care system in India need to shift toward a vision of "personalized medicine" which may improve the standard of medical care by including an individual's genetic and molecular information in the clinical decision-making process. Towards this, we propose to perform whole-exome and RNA sequencing analysis of pancreatic cancer in India to identify novel biomarkers and drug targets specific to our population for better management.

Current Status: There is no genomic sequencing data available for pancreatic cancer of Indian origin, which is required for the identification of population-specific cancer biomarkers and drug targets. Currently, available cancer biomarkers are developed based on genomic information from the western population, which may not be suitable for the diagnostics and treatment of the Indian population because the genetic background is different. The limited genomic information of the Indian population represented in the publicly available database is the roadblock for understanding cancer pathogenesis, developing the biomarkers and drug targets specific to the Indian population.

Institutional Ethical Clearance: Tissue samples are collected from patients diagnosed with cancer after informed consent and stored at the National Cancer Tissue Biobank (NCTB), Indian Institute of Technology Madras, Chennai, which was approved by the Institutional Ethical Committee of IIT Madras (IITM/IEC 2014043). All samples are coded and stored with patient clinical information by standardized protocols. Tissue samples required for exome and transcriptome sequencing analysis will be obtained from National Cancer Tissue Biobank, IIT Madras. Patients were given consent for sequencing analysis and data sharing for cancer research purposes.

Whole Exome Sequencing (WES) of pancreatic cancer*:* The details of sample processing for exome sequencing were outlined in Figure 1. Genomic DNA from 60 pancreatic cancer samples with respective normal tissue samples will be isolated according to the supplier's protocols. All DNA samples will be quantified and will be used to prepare indexed libraries using the SureSelectXT kit. Library preparation will be performed using a semi-automated 96-well plate method, with washing and clean-up/concentration steps performed on the Beckman Coulter Biomek NXP platform and with ZR-96 DNA Clean & Concentrator-5 plates, respectively. Libraries will be quantified using the Bioanalyzer.

Pooled libraries will be sequenced (2x150 paired-end runs) to achieve a minimum of 100x on target coverage per each sample library. The raw sequence data will be demultiplexed and converted to FASTQ files, adaptor, and low-quality sequences will be trimmed. Whole exome sequencing data will be used for novel somatic and germline mutation detection, microsatellite instability prediction, and somatic copy number alteration (SCNA) analysis and will be used for the development of the database.

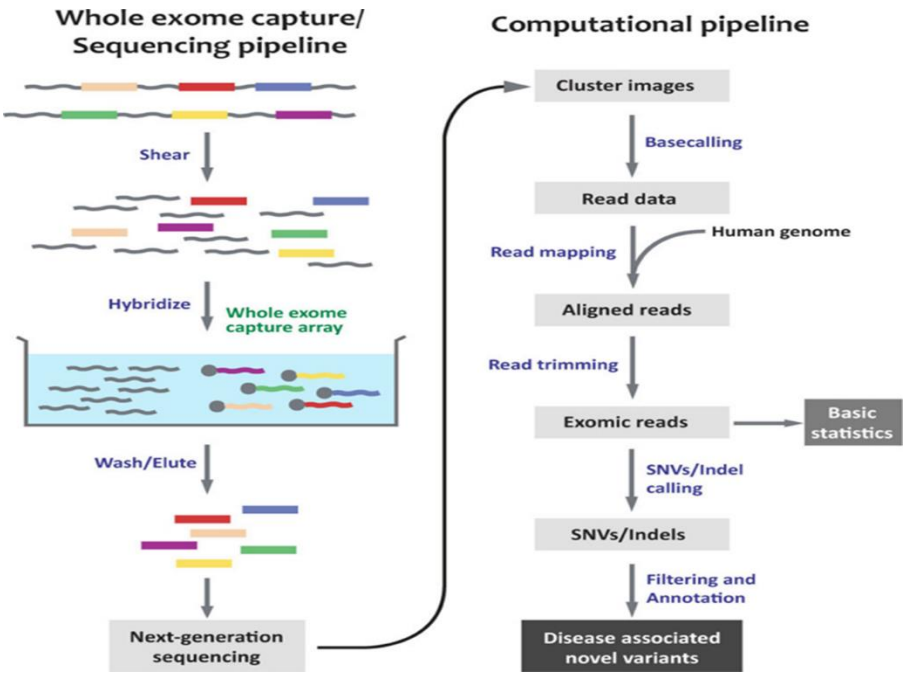

Figure 1. Whole Exome Sequencing and analysis pipeline

Transcriptome Sequencing (RNA Sequencing) of pancreatic cancer: Total RNA will be extracted from 60 pancreatic cancer with respective normal tissue samples using RNeasy Mini kit (Qiagen) as per the manufacturer's instructions. The processing of samples for transcriptome analysis is outlined in figure 2. The quality and quantity of RNA will be measured using Qubit™ RNA BR assay (Thermo Fisher Scientific, USA) on the Qubit™ fluorometer (Thermo Fisher Scientific). The RNA Integrity (RIN value) will be determined using RNA6000 Nano kit (Agilent) on Bioanalyzer 2100 (Agilent).

Truseq mRNA stranded kit (Illumina) will be used to construct the cDNA library for each sample, as per the manufacturer's instructions. cDNA libraries will be quantified and then subjected to paired-end sequencing (2x150bp). The *.bcl2* raw files will be converted into *.fastq* files, and the raw data quality will be checked using FastQC (v0.11.5) and MultiQC (version 1.0) for various parameters such as the percentage of bases above Q20/Q30 respectively, sequencing adaptor contamination rate, etc. The raw reads will then be processed to remove the adapter sequences and the low-quality reads and then will be used for gene fusion, copy number alteration, and differential expression analysis, and the development of the database.

Data Analysis: Linux platform will be used to perform all further bioinformatics analyses, using Perl, Python, bash, and R scripts. The *.fastq* files will be aligned to the reference genome (build hg38) obtained from RefSeq browser (Release 82) using Star Aligner (version 020201). On completion of the 2-pass alignment and indexing, read groups will be added to the files, and the duplicates will be removed (Picard tools version 2.9.2). Qualimap (version 2.2.1) will be used to evaluate the *bam* file characteristics. Stringtie tool (v1.2.2) will be used to identify the possible novel transcripts. The expression counts for each gene will then be calculated using the feature Counts tool. DeSeq2 will be used to estimate the level of expressed transcripts and calculate the level of differential expression in genes between samples. The expression level for each gene will be normalized using housekeeping genes in terms of reads per kilobase of the transcript (million mapped reads – RPKM). The reference GTF file will be obtained from RefSeq. Heatmaps and

volcano plots will be generated using RStudio (version 1.1.453). UpSet plot generation will be performed using Intervene.

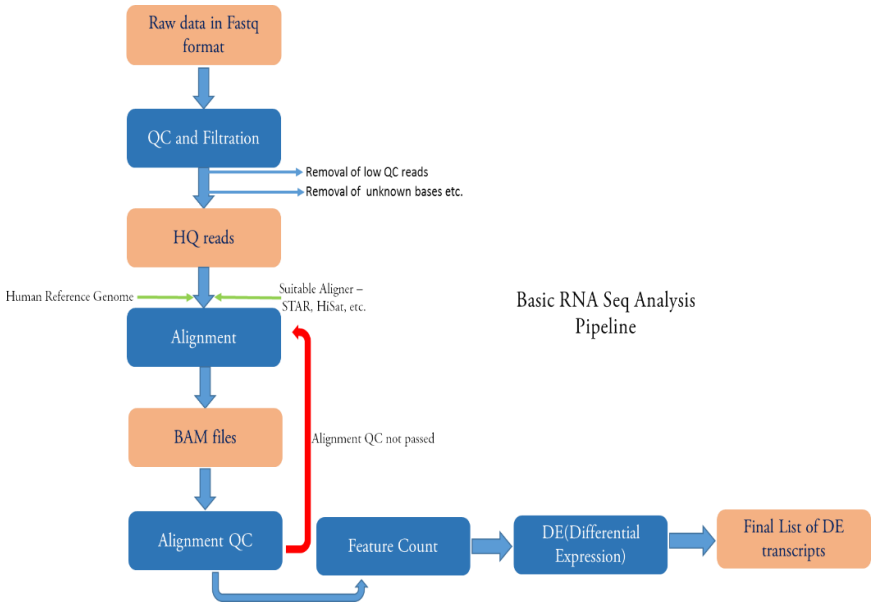

Figure 2. Transcriptome Sequencing and analysis pipeline

Gene Ontology and Pathway Enrichment Analysis: Based on the results from the gene expression analysis, Pathway and Gene Ontology (GO) enrichment analysis will be carried out to identify the functional role of the differentially expressed genes. Pathway (Wiki Pathways & Kyoto Encyclopedia of Genes and Genomes – KEGG) enrichment and GO (Molecular Function, Biological Process and Cellular Component) analysis will be carried out using WebGestaltR.

Aim. 2. To identify specific biomarkers for early detection and novel drug targets to design better therapeutics.

Based on the differential expression profile and mutational landscape confirmed using RT-qPCR and Sanger sequencing, the significantly deregulated genes will be selected for developing NGS panel for early detection. Targeted NGS panels will have more clinical advantage over other existing diagnostic methods, due to faster turnaround times and higher sequencing depths resulting in higher analytical sensitivity and specificity. The panel will be designed to interrogate SNVs, indels, CNVs/LOH, and both known and novel fusions associated with major fusion partners. One of the major benefits of using an NGS-based genetic testing strategy is the ability to detect multiple types of aberrations in a single assay. Also, as an alternate, we will develop antibody-based detection assay based on the availability of antibodies. We will test the detection limit of both assays before being validated with a large number of tumor samples.

Summary of novelty and intellectual merit of the proposal:

Cancer is one of the fatal health problems faced by mankind today. The limited survival of cancer patients is likely due to a high proportion of patients with advanced disease stages, a lack of suitable markers for early detection, and failure to respond to available therapy. This is due to the non-availability of population-specific or individual genomic sequences. For example, diagnostic markers and anti-cancer drugs developed against western populations are currently used in India. Due to the genome variation and tumor heterogeneity between different people even within the

population, the survival rate is very poor in India compared with the western population. There is an urgent need to identify Indian population-specific genomic information, which is very critical to develop efficient diagnostic markers for early detection and drug targets for novel anti-cancer therapeutics. In addition, due to genomic heterogeneity in cancers among different populations, population-specific approaches are critical to managing cancer in specific populations. We planned to perform comprehensive genomic sequencing analyses of pancreatic cancer of Indian origin to identify diagnostic markers for early detection and drug targets for novel anti-cancer therapeutics. The genomic information generated from the proposed work will help to improve our ability to detect early, resectable pancreatic carcinomas, and provide a mortality benefit to patients at significantly elevated risk of PDAC.

It is important to note that there is no representation of Indian pancreatic cancer genome sequence data in the international cancer genome consortium or country-specific cancer genome databases from the USA (like TCGA), UK, and Australia. Very little or no genome sequence information is available for cancers of Indian origin (except some for oral cancers). This is a unique data set and will be critical to developing biomarkers for early detection and novel drug targets for better therapeutics for Indian cohorts. It will also provide a further impetus for international collaborations, as already evidenced by existing collaborations.

Budget (INR in lakhs)

S. No	Consumable Materials	Total
	Whole Exome Sequencing (100 pancreatic samples (*50 cancer tissue+50 matching normal tissue*): DNA Extraction kit, Quantitative & qualitative analysis of DNA by Nanodrop spectrophotometry method followed by Qubit fluorometric method (Qubit DNA BR assay), Next-generation sequencing library construction reagents (DNA Fragmentation by mechanical shearing, Adenylation, Adapter ligation, Hybridization and Enrichment PCR), Quantitative analysis of Sequencing library by Qubit Fluorometric method (Qubit DNA HS assay), Quantitative & qualitative analysis of Agilent Bioanalyzer (DNA 1000 kit) followed by Real-time PCR, Next-generation sequencing by Illumina sequencing by synthesis (SBS) technology (Paired end sequencing and 100X coverage).	**25.00**
	Transcriptome Sequencing (100 pancreatic samples (50 *cancer tissue + 50 matching normal tissue*): RNA Extraction kit, Quantitative & qualitative analysis of RNA by Nanodrop spectrophotometry method followed by Qubit fluorometric method (Qubit RNA BR kit) and RNA integrity (RIN) analysis using Agilent Bioanalyzer (RNA 6000 nano kit), Next-generation sequencing library construction reagents (mRNA enrichment, cDNA synthesis, fragmentation, Adenylation, Adapter ligation and Enrichment PCR), Quantitative analysis of Sequencing library by Qubit Fluorometric method (Qubit DNA HS assay), Quantitative & qualitative analysis of Agilent Bioanalyzer (DNA 1000 kit) followed by Real-time PCR, Next-generation sequencing by Illumina sequencing by synthesis (SBS) technology (Paired end sequencing and 40-60 million reads/sample).	25.00
	Sub-Total	**50.00**