**MCL Supplement: A Pre-Cancer Atlas (PCA) for Breast, Lung, Pancreas, and Prostate**

## A. Background/Significance

One of the critical barriers to developing new approaches for cancer detection and prevention is the lack of understanding of the key molecular and cellular changes that cause cancer initiation and progression. Unlike the extensive work that has been done profiling advanced stage tumors, few studies have comprehensively profiled the molecular alterations found in precancerous tissues. Premalignant lesions are currently characterized by histologic changes that precede the development of invasive carcinoma[1,2].These lesions can often be identified in regions surrounding an invasive tumor or in biopsies taken from patients undergoing diagnostic evaluation for suspicion of cancer. Currently, limited metrics exist to identify lesions that will likely progress to carcinoma and require intervention from those that will naturally regress or remain stable [3,4]. Characterization of the molecular alterations in premalignant lesions and the corresponding changes in the microenvironment would hasten the development of biomarkers for early detection and risk stratification as well as suggest preventive interventions to reverse or delay the development of cancer.

Our pilot study will establish the feasibility of transcriptomic, genomic and immune profiling of FFPE premalignant lesions from multiple organ sites, collected and profiled with uniform SOPs across multiple institutions within the MCL consortium. We will characterize the molecular alterations in precancerous lesions and the corresponding microenvironment in four major organ sites, in order to uncover the molecular and cellular determinants of premalignancy, and establish standardized sequencing and immunohistochemistry protocols on FFPE precancerous tissue. We will also evaluate the technical feasibility of single nuclei sequencing of small FFPE pre-cancer lesions. Successful completion of the proposed pilot study will set the stage for expansion and development of a comprehensive Pre-Cancer Atlas (PCA) as part of the NCI's moonshot

## B. Specific Aims

**Aim 1: Collect premalignant lesions (PML) and their associated microenvironment via LCM from FFPE tissue across four organ sites (breast, lung, pancreas & prostate).**

**Aim 2: Perform bulk RNA and DNA seq on premalignant FFPE samples (and flash frozen tissue where available) and compare the genomic/transcriptomic alterations within and across organ sites.**

**Aim 3: Determine the feasibility of performing single "nuclei" RNA-seq on FFPE PML tissues and compare with single "nuclei" RNA-seq from fresh frozen and single "cell" RNA-seq of fresh PML tissue**

**Aim 4: Characterize the immune microenvironment via a multiplex IHC panel and determine its relationship with the PML genomic/transcriptomic landscapes**

**C. Approach**

**Aim 1: Collect premalignant lesions (PML) and their associated microenvironment via LCM from FFPE tissue across four organ sites (breast, lung, pancreas & prostate).**

**Methods.**
*I. Patient Population/Sample Collection:* Overview of the sites collecting PML tissue from the respective organs is provided in **Table 1** and a full description of the biospecimens to be obtained is described in detail for each organ type below.

| Organ Site | Breast | | Lung | | Pancreas | Prostate | |
|---|---|---|---|---|---|---|---|
| Type of PML | DCIS | | AAH, Squamous Dysplasia/CIS | | PanIN, IPMNs | PIN | |
| Collection Site | UCSF/UCSD | UVM | BU*/UCLA | Vanderbilt/Moffitt | MDACC* | JHU | Stanford* |
| # of Patients | 20 | 20 | 20 (10 of each type) | 20 (10 of each type) | 20 (10 of each type) | 20 | 20 |
| Total Patients per Organ | 40 | | 40 | | 20 | 40 | |
| Note: single nuclei/cell RNA-Seq will be performed on 4-5 FFPE samples from each of the organ types | | | | | | | |

**In addtion to the Broad Institute, these sites will perfrom single cell RNA-seq on fresh tissue from 5 patients*

**Table 1. Breakdown of cohort by tissue type and collection site.**

***1. DCIS lesions from breast tissue:***
DCIS lesions will be collected from 40 patients (20 from UCSF/UCSD & 20 from UVM) with primary low or high-grade DCIS diagnosed from a breast core biopsy. Subsequent resected lumpectomy or mastectomy tissues will be prospectively sampled in the vicinity of the prior biopsy site using multiple approaches: 1) Live cells (heterogeneous mix) will be obtained as a cell scrape slurry from the lesion surface or by fine needle aspirate (FNA); 2) For a subset of specimens where size is sufficient, a block of breast tissue with DCIS will be fresh-frozen; 3) The remainder of the specimen will be taken for routine formalin-fixation and paraffin-embedding (FFPE). The FFPE sample will be annotated to identify the matched FFPE tissue block adjacent to the fresh-frozen sample and will be sectioned for use in bulk and single nuclei sequencing . We will dissect DCIS, adjacent normal and when available, associated carcinoma. In addition, when possible, normal tissue will be collected from a tissue block lacking lesions as well as collection of blood. A subset of patients (n = 5 | FFPE, flash frozen and fresh) will be sent to the Broad Institute for single nuclei/cell sequencing.

***2. AAH and squamous dysplastic/CIS lesions from airway and lung tissue:***
For squamous cell lung cancer, we will collect endobronchial biopsies from abnormal airway regions identified on autofluroscence bronchoscopy or identify PMLs in the margins of resected lung tissue. We will study 20 patients (5 each from BU/UCLA/Vanderbilt/Moffitt) with pre-invasive squamous lesions (moderate-severe dysplasia or carcinoma *in situ* (CIS)) identified on pathologic examination. LCM of the premalignant region and adjacent normal epithelium will be performed as well as the invasive tumor for those collected from the resection margin (n=5 from UCLA). On a subset of lesions collected at bronchoscopy (n=5), we will collect additional biopsies that will be flash frozen and fresh for single nuclei and cell sequencing, respectively, performed at the Broad Institute. In parallel to the work at the Broad, BU will perform single cell RNA-seq on these freshly cell sorted tissues (n = 5). Blood will be collected on all patients for genomic studies.

For lung adenocarcinoma, we will collect resected FFPE lung tissues from 20 patients (10 from UCLA and 10 from Vanderbilt/Moffitt) with early stage lung adenocarcinoma that harbor atypical adenomatous (AAH) premalignant lesions in the resection margin. We will LCM multiple AAH regions (3-5 per patient) as well as adjacent regions of normal epithelium and invasive adenocarcinoma. In addition, blood will be collected on all patients for genomic studies.

***3. PanINs and IPMNs from pancreatic tissue:***
For pancreatic cancer PML, we will collect low and high grade lesions from 20 patients (all 20 from MDACC) representing microscopic Pancreatic Intraepithelial Neoplasia (PanIN) (n = 10), as well as

macroscopic Intraductal Papillary Mucinous Neoplasms (IPMN) (n=10) from surgically resected specimens along with blood samples. Archival FFPE specimens of microscopic PanIN lesions, occurring multi-focally adjacent to invasive PDAC, and archival IPMN lesions (with or without associated invasive cancer), along with the adjacent normal tissue, will undergo LCM and utilized for bulk DNA and RNA sequencing. If matched frozen tissues are available for a subset of these FFPE samples, we will bank for comparison of profiles. Because IPMNs are macroscopic lesions, they provide an opportunity for obtaining the samples fresh and therefore can be used for single cell sequencing (in contrast to PanINs). Therefore, 5 freshly obtained IPMNs will be used for the single cell RNA sequencing studies performed at both the Broad Institute and MDACC, and the matched FFPE and/or frozen sections from these lesions (obtained from the adjacent PML) will be sent to Broad Institute as a pilot to assess "single nuclei" RNA sequencing. Since PanIN lesions are not visible in fresh tissue, we will not be able to perform single cell RNA-Seq in these PMLs.

***4. PINs from <u>prostate</u> tissue:***

For prostate cancer PML, there will be 40 samples of Prostatic Intraepithelial Neoplasia (PIN) collected between the Stanford and JHU sites (20 cases per site). At the Stanford site, 20 prostate specimens detected by PSA screening who have/will undergo surgery (radical prostatectomy) for clinically localized disease will make up the final cohort. The age range of the participants would be 40-75, and we anticipate that 18 will be Caucasian, 1 Asian and 1 Latino or African American based on the practice demographics practice at Stanford. Clinical and MRI data will also be collected for these samples. We will collect low grade (e.g. Gleason score of 6/Grade group 1; n=10) and high grade (Gleason score 4+3=7 or higher/Grade group 3 or higher; n=10) PINs from FFPE samples that have prostate carcinoma. In addition to obtaining LCM archival samples of low and high grade PIN, we will also obtain normal prostatic epithelial from the peripheral, central and transition zones as well as multiple samples of prostate carcinoma in order to obtain the spectrum of Gleason grades in the carcinoma as needed. LCM samples will be used for bulk DNA and RNA sequencing. In addition, single cells will be dissected from FFPE samples to prepare single cell RNA seq libraries using techniques developed at Stanford, and FFPE tissue will be sent to the Broad for single nuclei sequencing. When available, flash frozen and fresh samples from these prostates will be archived and prepared for single nuclei and cell sequencing, respectively, at the Broad Institute and at Stanford (single cell only). JHU will also capture 10 cases (5 grade group 1 and 5 grade group 2) of high grade PIN, normal and invasive adenocarcinoma using frozen sections from fresh frozen tissues. When possible these will be from the same patients as the FFPE samples.

Since frozen sections can be quite challenging to morphologically determine high grade PIN from normal epithelium, for these samples we will perform a number of additional tissue-based characterizations. These will include a multicolor combined basal cells (p63 and CK903) and PIN/carcinoma markers (AMACR) referred to in the cocktail as "PIN4", c-MYC (referred to as MYC) protein[5], by IHC and mRNA by *in situ* hybridization (AM De Marzo, Q Zheng unpublished observations), telomere length by *in situ* hybridization[6] and the 5'ETS/45S rRNA[7]. For these slides, the whole slides will be scanned with a Hammamatsu Nanozoomer with a 40x objective and regions of interest will be annotated as a guide for LCM.

*<u>II. Laser-capture microdissection (LCM)</u>:* FFPE tissue blocks will be sectioned at 7µm thickness and serial sections will be stained with H&E. LCM will be performed utilizing standard LCM systems, such as Leica LMD7000 and ArcturusXT at each site. Regions of premalignancy will be dissected and RNA/DNA will be extracted from microdissected cells using the Qiagen All Prep DNA/RNA FFPE Kit.

*<u>III. Single cell sorting of fresh tissue</u>: Fresh tissue will be dissociated with an optimized digestion protocol for each tissue type with Trypsin/EDTA or dispase and cells will be sorted using a BD FACSAria II. FACS is used to isolate singlet events based on forward scatter height vs. forward scatter area (FSH-H vs. FSH-A). Dead cells (PI+) and red blood cells (GYPA+) are stained and excluded, and live cells (Hoechst 33342+) are sorted.*

**Aim 2: Perform <u>bulk RNA and DNA seq</u> on premalignant FFPE samples and compare the genomic/transcriptomic alterations within and across organ sites.**

**Rationale:** There have been limited studies characterizing the genomic and transcriptomic landscape of premalignant lesions associated with breast, pancreatic, lung or prostate cancers. Characterizing the molecular

determinants of premalignant disease that are unique and shared across multiple organs will enable new candidate biomarkers for early detection and novel therapeutic strategies for early intervention.

**Methods.**

*Bulk RNA-seq of LCM FFPE tissue:* All participating sites will perform bulk RNA-seq in accordance with SOPs developed at BU. In brief, total RNA will be isolated from LCM'd lesion and associated microenvironment tissue using the Qiagen All Prep DNA/RNA FFPE Kit and quality will be assessed with the Agilent Bioanalyzer. Libraries will be generated with the Illumina TruSeq Access kit (for FFPE samples). They will be sequenced on the Illumina HiSeq2500 with 75base-pair paired-end reads. Quality of FASTQ files will be assessed with FastQC. Reads will be aligned to the human genome with STAR and gene-level and isoform-level expression will be quantified with RSEM. Splice junction saturation, transcript integrity, and biotype distributions will be calculated for each sample with RSeQC. DESeq2 and EdgeR will be used to identify associations between gene expression profiles and clinical variables while controlling for confounding covariates.

BU will serve as an RNA-seq Core to assess reproducibility of FFPE RNA-seq methods across sites. We will perform RNA-seq according to the SOP listed above on a subset of samples for each organ type (total n ~ 20).

*Bulk Whole exome-seq of FFPE tissue:* All participating sites will perform bulk whole exome-seq (WES) in accordance with SOPs. In brief, DNA from laser captured material will be isolated using the Qiagen All Prep DNA/RNA FFPE Kit and undergo stringent quality control to ensure high quality input material for genomic profiling. Purified DNA (ideally 100-200 ng) will be used for library preparation and amplification, followed by next generation sequencing using standard protocols distributed by CDMG. Exome-seq methods are considered standardized, thus we will not need a DNA-seq Core to assess reproducibility across sites. We anticipate local centers will use Illumina paired end reads, following the following general approach. **1**) DNA library preparation: Paired-end libraries will be prepared following the manufacturer's protocols (Illumina and Agilent), fragmented to 150-200 bp **2**) Capture of targeted exome: Whole exome capture will be carried out using the protocol for Agilent's SureSelect Human All Exon kit. Purified capture products will be amplified using the SureSelect GA PCR primers (Agilent) for 12 cycles. **3**) Sequencing will be carried out for the captured libraries using at least 100 bp paired-end reads. To achieve high level sensitivity and accuracy for detecting all the mutations in the whole exome, each sample will be sequenced at 200X mean depth. **4**) Read mapping and alignment and variant analysis: Sequence short reads will be aligned to a reference genome (NCBI human genome assembly build 38) using BWA-MEM. Local realignment of aligned reads will be performed using Genome Analysis Toolkit (GATK).

In addition to RNA and DNA-seq, individual sites may choose to perform additional assays (e.g. TCR-seq) on subsets of samples; however these studies will not be detailed in this proposal as they are not uniform across all sites. See individual budget justifications for specifics regarding potential additional assays.

**Analytic Approach and Power Calculations.**

There are several questions being addressed in this pilot study. One goal is describing expression and mutational profiles of the premalignant lesions and descriptive statistics including box plots and volcano plots will be used to describe data distributions. The sample sizes support the identification of genes that show altered expression from RNA seq analysis of premalignant lesions versus normal adjacent tissue. We will also perform hierarchical clustering of the samples to evaluate whether the expression data support the existence of multiple subtypes of neoplasia within each type of premalignant lesion. We will also perform hierarchical clustering across premalignant lesions to identify any subgroups that might show similar expression patterns across sites. Mutation and copy number analyses will be compared with existing databases such as COSMIC[8] for somatic mutations and Database of Genomic Variants (http://dgv.tcag.ca/gb2/gbrowse/dgv2_hg19/) for copy number analysis. To integrate different datatypes and evaluate further for similarities among the patient samples we will apply iClusterPlus[9]. The method performs a clustering analysis, while fitting clusters to latent classes of cancers that show similarities. The clustering will be performed using gene expression, splicing, somatic variation, CNA, and immune profiles.

*Alternative splicing. A* majority of human genes with multiple exons undergo alternative splicing and in many cases this process plays a very important role in cancer development[10] including cancer of the lung[11]. Although

this remains to be tested for lung cancer, several studies suggest that splice isoforms can predict progression and survival in other types of cancer[12,13]. To identify alternatively spliced transcripts, we will use a modified version of the DEXSeq bioconductor we have recently published, which improves detection of intron retention events[14].

*Copy Number Alteration (CNA).* Aneuploidy of tumors (here called copy number alteration, which can include amplifications, deletions and rearrangements of selected regions) has recently emerged as an independent predictor of aggressiveness of a lesion. High levels of somatic copy number alteration in tumors are associated with reduced immune cell infiltrates and poor prognosis of patients with cancer[15], but limited data exist for premalignant lesions. CNA will be evaluated from the Exome seq analysis using EXCAVATOR[16].

*Immune Profiling.* Several studies have implicated immune infiltration as a driver of disease progression and patient prognosis in many cancers. Infiltrations of CD4+ and CD8+ T cells have been associated with improved survival[17], while immunosuppressive T-regulatory cells[18] and macrophages have been associated with poor survival for most solid cancers[19]. Deconvolution of whole transcriptome analysis will be used to infer the infiltration by several types of lymphocytes and other white blood cells[20].

*Power computation* PS calculator[21] was used to estimate statistical power for RNA sequencing. Estimation is based on a proposed sample size of 20 premalignant lesions compared with 10 normal adjacent tissues, assuming that only half of samples from each organ site will have enough sample for RNA sequencing to be completed. Type I error for expression studies was conservatively adjusted for the number of tests using Bonferroni correction for 20,000 tests. Our previous study using RNA sequencing data showed average intrasample variance on log(2) scale of expression levels was 0.58 (excluding genes having average read depth of 5 or less). We will have adequate power to detect as statistically significant 2.3 fold increase or decrease in expression level between premalignant and adjacent normal tissue.

*Data QC.* To ensure scientific rigor and consistency among sites in RNA and DNA processing we will include a preliminary analysis of steps in processing and analysis. Protocols for extraction of high quality RNA and DNA from formalin fixed paraffin embedded (FFPE) tissues, which will be used extensively in these studies continue to improve and may have variable implementation among the sites participating in this study. To evaluate consistency of preliminary steps in processing and downstream analyses, we will initially distribute slides from one large FFPE fixed cancer of origin from prostate, breast, lung and pancreatic cancer. Analysis of these samples will allow us to review the DNA and RNA characteristics (yield, purity and strand length) among sites. Downstream analysis of these same samples will also allow us to compare among sites the consistency of variant calls among centers. We will be able to identify if there are some times of calls (such as small insertion deletions) that are more variable among centers versus other types of calls (such as relative gene expression or single base pair substitutions) that we expect to be less variable and to characterize the reliability of findings across sites. We are also including a 5% blind duplicate analysis of RNA sequencing. Samples will be analysed by the participating genomics cores without knowledge of the phenotype. RNA seq and CNA analyses are normalized for batch effects. We will also compare the observed sex to the self-reported sex as based on RNA profiles and exome sequencing of X chromosome genes as another check for processing accuracy and sample management.

**Aim 3: Determine the feasibility of performing single "nuclei" RNA-seq on FFPE PML tissues and compare with single "nuclei" RNA-seq from fresh frozen and single "cell" RNA-seq of fresh PML tissue.**

**Rationale.** Pre-malignant lesions reside within a complex multicellular ecosystem comprised of both pre-malignant and non-malignant cells. Recent advances in single-cell genomics, in particular in single-cell RNA-Seq (scRNA-Seq), many pioneered by the Regev group[22–42] and *in situ* analysis of RNA[43–48], provide extraordinary opportunities for systematic identification of the composition of this ecosystem, to discover new biomarkers, understand tumorigenesis and identify preventive interventions. However, most tissue sections, and in particular pre-malignant samples, are preserved as formalin fixed, paraffin-embedded (FFPE) samples, whereas even the most cutting-edge approaches in single cell transcriptomics, such as massively-parallel single nucleus RNA-Seq[25,49], have only been applied to frozen or lightly fixed tissue. *In situ* RNA hybridization techniques, such as smFISH, are compatible with FFPE, but those relying on sequential hybridization for

multiplex assays have not yet been tested in this way. If this technical challenge is overcome, it will transform our ability to both study cancer and deploy single cell genomics through standard clinical pathology pipelines. Here, we will work together with the MCL consortium and harness our expertise in developing new experimental and computational methods for single nucleus genomics (Regev), tissue processing (Regev) and multiplex spatial analysis of RNA *in situ* (Chen) to develop and apply methods that enable profiling of pre-malignant samples collected as FFPE tissue (MCL consortium).

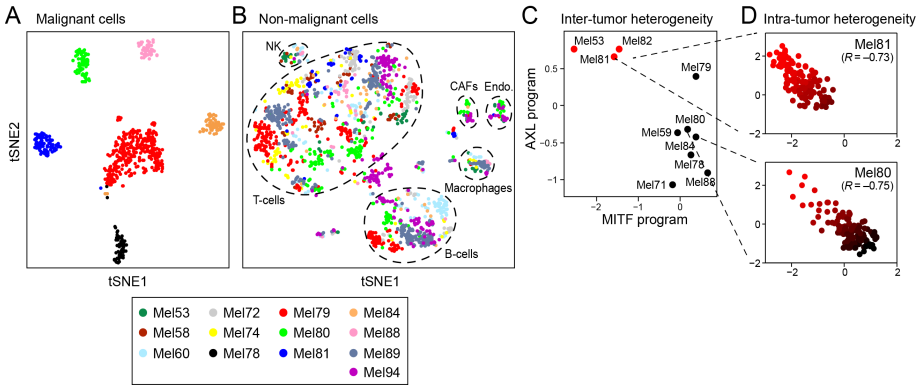



**Fig. 2. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-Seq.** (A, B) Dissection of melanoma using single-cell RNA-Seq. Single-cell expression profiles distinguish malignant and non-malignant cell types. Shown are t-SNE plots of malignant (A) and non-malignant (B) cells. (C, D) MITF- and AXL-associated expression programs vary between tumors, within tumors, and following treatment. (C) Average expression signatures for the AXL program (*y*-axis) or the MITF program (*x*-axis) stratify tumors into 'MITF-high' (black) or 'AXL-high' (red). (D) Single-cell profiles show a negative correlation between the AXL program (*y*-axis) and MITF program (*x*-axis) across individual malignant cells within the same tumor

**Preliminary Data.** The Regev lab has demonstrated the feasibility and potential of using scRNA-seq to understand tumors, their microenvironment and the specific role played by the immune system. Specifically,

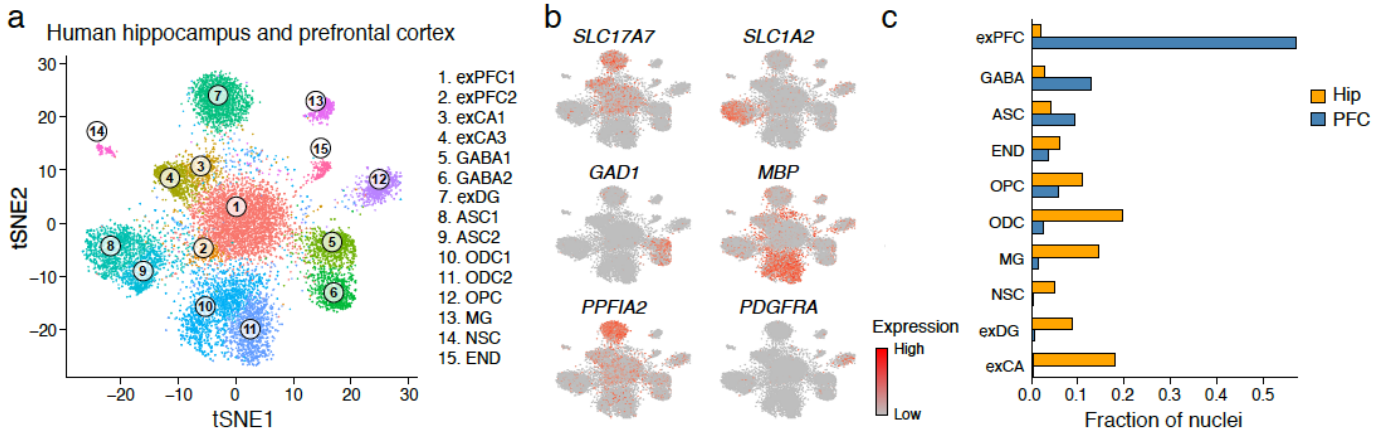



**Fig. 1. DroNc-seq distinguishes cell types and signatures in adult post-mortem archived frozen human brain tissue.** (**a**) Cell-type clusters. tSNE embedding of 14,963 DroNc-seq nuclei profiles from adult frozen human hippocampus (Hip, n=4) and prefrontal cortex (PFC, n=3) from five donors. Nuclei are color-coded by cluster membership and clusters are labeled *post-hoc*. (**b**) Marker genes. Shown is the same plot as in (a) but with nuclei colored by the expression level of known cell-type marker genes. (**c**) Fraction of nuclei from each brain region associated with each cell type. Cell types are sorted by known enrichment in PFC *vs*. Hip.

we have shown that scRNA-Seq of tumors provides a powerful lens into the tumor ecosystem[23,28–30,41], by: (**1**) Analyzing thousands and tens of thousands of cells per tumor – recovering its complete ecosystem of malignant and non-malignant cells[23,29]; (**2**) Creating single nucleus RNA-Seq[49] (sNuc-Seq) to analyze preserved samples – either lightly fixed or fresh-frozen. (**3**) Scaling sNuc-Seq to massively parallel scale, by inventing DroNc-Seq[25] (**Fig. 1**); (**4**) Showing – by molecular histology of proposed markers and signatures from scRNA-seq in both matched and independent samples – that our predictions and results are robust and generalizable[23,28–30,41]; (**5**) Developing and validating extensive analytics and computational pipelines for the distinction of malignant and non-malignant cells by inferred CNVs (validated rigorously)[23,23,29,50], for clustering of cells into types and states and for identifying their signatures[30,31,35,37]; and (**6**) Demonstrating how to deconvolve bulk (legacy) samples based on single cell profiles, redefining tumor classification by the microenvironment, and discovering cell-cell interactions in the tumors[23,29,41].

In this context, we have already made key discoveries on the ecosystem of diverse tumors (Fig. 2). For example, in the first single cell RNA-Seq study of a tumor ecosystem, we focused on metastatic melanoma[29]. Here, we distinguished malignant from non-malignant cells by gene expression clustering and by estimation of

copy number variations (**Fig. 2a,b**); we discovered drug-resistant cells (**Fig. 2d**; red dots) in tumors believed to be drug-sensitive based on population RNAseq data (**Fig. 2c**; blue dots); we determined the subsets of immune cells present in the tumors and TME and performed further T-cell stratification based on expression of key transcripts.

**Aims and Methods.** We will develop, optimize and test new methods for snRNA-seq on FFPE samples in order to characterize the complex ecosystem of PML lesions. To mitigate the risk posed by FFPE, and to complement with critical spatial information, we will also optimize and apply multiplex *in situ* RNA analysis of FFPE using signatures derived from the snRNA-Seq data and/or bulk RNA-Seq data (**Aim 3.1**). Once successful, we will apply it to PML tissues across the MCL consortium including samples from breast, lung, pancreas, and prostate (**Aim 3.2**).

### Aim 3.1 Develop a method for profiling PML FFPE samples using snRNA-Seq

*Aim 3.1.1 Testing and Optimization of Single nucleus RNA-seq of FFPE samples.*

While FFPE is the most common method of archiving solid tissue pathology specimens, both cellular structures and RNA are damaged in this process. In recent years, technical innovations (*e.g.*, [51]) have allowed bulk RNA-Seq from FFPE samples, enabling clinical studies, but, to date, the minute quantities of RNA in single cells and the damage to structures have left FFPE out of reach of single cell genomics. To test the feasibility of single-nucleus RNA-Seq on FFPE samples we will first perform a technical pilot. Together with the Spira lab (BU) we will compare single cell and single nucleus RNA-Seq analysis of matched fresh, fresh-frozen and FFPE lung tissue from the same specimen. At the early pilot phases, we will not use precious PML tissue, but other lung tissue of similar technical characteristics, that can be made readily available. We will use our standard droplet-based technologies to process the fresh (single cells) and frozen tissues (single nucleus) and compare them to the FFPE sample. We will use FFPE lung tissue to optimize the method and we will profile and compare 7,000 nuclei from 5 fresh, 5 frozen and 5 FFPE lung tissues.

For FFPE samples, we will optimize the depolymerization of paraffin and reverse formalin fixation during this step by carrying out the incubation with Xylene or FRISCR[52] at a higher temperature. The temperature and duration of depolymeration will need to be tested. Also, agents other than xylene will need to be tested to optimize subsequent nuclei extraction (possibly even RNAlater). It is possible that xylene may work as proteinase K is used in the subsequent bulk step. This suggests that the RNA-protein interaction is intact, possibly preserving the nuclear structure. Once tissue is extracted from paraffin and formalin cross-linking is reversed then a variety of commercial, detergent, and mechanical nuclei isolation methods will need to be tested to optimize getting intact nuclei and good RNA content. Initially, we will process the FFPE sample using our standard droplet based snRNA-Seq method (DroNC-Seq[25] or similar protocols). Since these protocols aim to capture the 3' end of transcripts it could be that the RNA degradation may be circumvented although, it is expected that the template switching activity may be hindered to some degree. If these methods do not work we will also test modified protocols to overcome the limitations of FFPE and will need to optimize methods for working with short cDNAs such as using the: CEL-seq[53,54], random prime and removal of rRNA reads (by including normalization with DSN and CRISPR), a targeted approach, similar to Illumina's TruSeq Access kit using an oligo set to capture genes of interest, not-so-random priming and ligation based approach, where every transcript is tagged through ligation then transcripts are pooled and the rRNA is depleted with riboZero. We will rely on expert personnel and best practices we have applied in numerous previous projects[25,35,39–42,49,51,55] for RNA-seq, scRNA-seq and snRNA-Seq optimizations.

We will assess the performance of single-nucleus RNA-Seq on FFPE tissue using our experimental and computational QCs. A positive result would be determined by passing specific analysis criteria, including: Passing QC measures for recovery of sufficient nuclei/sample with sufficient number of genes per nucleus (*e.g.*, >500); Ability to determine the genetic status of the malignant cells by copy number variations, if appropriate (CNVs); Ability to group nuclei profiles by relevant cell types (malignant and non-malignant (with relevant subtypes as present)), within and across patients; Ability to detect co-expressed cell state gene signatures (*e.g.*, cell cycle signatures and signatures of immune cell states); and Ability to identify differentially expressed genes associated with cell type clusters. We will assess each sample as it is processed for each of these criteria, to obtain a continuous estimate of progress. Establishing these methods would determine embarking on the full project plan. FFPE samples will be compared to the sequencing output using the paired

frozen and fresh samples. We do not expect FFPE samples to reach the same performance as fresh or fresh-frozen samples, even at the end of optimization.

*Aim 3.1.2. Analyze multiple PML FFPE tissues using multiplex in situ spatial methods*

Spatial analysis of RNA in situ provides both important complementary information about the organization of cells and their respective states, as well as a potential technical alterative, because smFISH has been applied in the past to FFPE sections. Notably, analysis of sets of RNA probes (multiplex FISH) or targeted sequencing *in situ* all require pre-defined signatures. Thus, they greatly benefit from the cell specific signatures identified by single cell transcriptomics, even if those were obtained on separate samples. Finally, computational analysis[34,56] can allow us to infer the spatial expression of the remaining genes. We will test several methods that are compatible with FFPE preparation including: Multi-color, single-molecule fluorescence in situ hybridization (smFISH), which enables analysis of several different transcripts simultaneously (up to 130 from recent published reports[43,44]). We will also test *in situ* sequencing using a recently published method[57–59] that utilizes rolling circle amplification and padlock probes to quantify up to 100 genes per tissue at single cell resolution. We will also subject samples to *in situ* sequencing using FISSEQ (in collaboration with Ed Boyden, Fei Chen and using Readcoor technology). For all signature assays, will use either signatures derived from bulk analysis or from snRNA-Seq on FFPE samples (depending on success of Aim 3.1.1). We will profile ~5 fresh, ~5 frozen and ~5 FFPE lung tissues, subjecting adjacent sections to both smFISH, in situ sequencing and FISSEQ.

**Aim 3.2. Profile pre-cancerous FFPE samples from the MCL consortium**

Once a positive outcome is obtained from *either* **Aim 3.1.1** or **Aim 3.1.2** we will proceed to profile PMLs across MCL labs from lung (covered in **Aim 3.1**), breast, prostate and pancreas for the full-scale project. If performing snRNA-Seq from FFPE samples is not successful (**Aim 3.1.1**), we will pursue two alternatives. (1) We will use signatures from bulk analysis followed by multiplex RNA signature-based assays on FFPE PMLs (**Aim 3.1.2**); and (**2**) We will prospectively collect several PML tissues, flash freeze them, perform DroNc-Seq and use the snRNA-Seq data to deconvolute the bulk samples.

**Analytic Approach.** Overall, we will profile 7,000 nuclei from ~4-5 samples of each PML type. To determine the number of nuclei to profile we rely on power analysis**.** We estimate how many nuclei we should sequence, if the tissue is expected to have *N* cell types, the rarest of which is in proportion *P*, and we would like to recover at least *n* cells of each type, at a confidence level *C*. Under a few conservative assumptions (one cell type dominate and all others are equally rare; Central Limit Theorem holds; samples are independent), the number of cells from each of the rare types ($T_i$) will distribute as: $E[T_i] = N^* p_{min}$, $SD[T_i] = sqrt(N^* p_{min} *(1- p_{min}))$. For example, to detect at least 20 cells of each cell type as low as 0.5% out of 15 types would at 95% confidence would require ~6,800 cells. Note that we can use this strategy adaptively to iteratively determine *N*, *P*, and *n* in a data-driven way, as we collect an initial number of cells based on an "educated guess" of these parameters, and reassess them once the initial data is collected (e.g. if it reveals additional groups).
We will perform computational analyses on PML samples with the following goals and approaches:

- Distinguish malignant from non-malignant cells. We will combine (1) gene expression clustering, which separates cells into groups based on their expression profiles, and (2) estimation of copy number variations (CNVs) from the average expression of genes in large chromosomal regions within each cell[41]

- Determine cell type specific signatures. We extend feature selection strategies to select markers from single-cell transcriptomes that define discrete subtypes (classes) or continuous spectra, either learned directly from the profiles (*e.g.*, using PCA**,** clustering; or trajectory finding) or incorporating prior knowledge (*e.g.,* known markers). For each feature (gene), we consider both how informative it is for classifying cells into each class *and* how reliably it is measured[27,37,40].

- Assess and discover cell states. We will use a unified and efficient pipeline that: (1) determines the lineage identity of immune cells in the TME by comparison to known compendia (*e.g.*, GTEx[60], and immune cell atlases[61]); (2) identifies co-varying transcriptional modules; and, (3) computes the activity of modules with respect to individual cells, cell clusters and cell spectra. To score signatures, we use a weighted formulation that accounts for both the 'information content' of each signature gene and the 'per-gene per-cell quality' of each measurement[29,30,37,40,49,50].

- Use scRNA-seq profiles to deconvolve bulk RNA-Seq profiles. We will use the cell type and state specific signatures to decompose bulk RNA-Seq profiles obtained from all bulk profiled PML samples. In one

application, we perform a single ("flat") deconvolution: this approach is ideal to distinguish the key compositions of a bulk tumor sample (malignant cells, CD8+ T cells, B cells, CAFs, etc). In a second application, we perform a step wise ("hierarchical") deconvolution to distinguish certain states only present in cells of a specific type.

- Combine single cell signatures and cohorts analyzed in bulk. We combine the single cell signatures (for cell type and state) and a cohort of bulk profiles: We (1) determine any changes in cell proportions by decomposing the bulk profile with all cell type signatures; (2) recover the malignant profile as a residual; (3) score for the proportion of cells in specific states; (4) associate each such feature with key clinical parameters, and (5) identify genes expressed by one cell type that affect another.

*In addition to single cell sequencing at the Broad Institute (described above), we will perform single cell sequencing of fresh tissue from 5 patients in parallel at BU, MDACC and Stanford. Single cell methods vary across sites, providing an opportunity for us to assess variability of these new technologies and potentially help guide platform selection for future PCA studies. See budget justification for details on each site's single cell methods.*

**Aim 4: Characterize the precancer phenotype and immune microenvironment and determine its relationship with the PML genomic/transcriptomic landscapes**

**Rationale.** Precancer and in situ neoplasia comprises an array of diverse phenotypes within each of the organ sites. These phenotypes are a function of factors including the cell of origin, genome changes, and interplay with the microenvironment. Immuno-editing, in particular, is likely to occur at the precancer stage, as early invasive lesions would be rapidly eliminated without active immune evasion and without the relative immune protection of the in situ niche. Our objective, is to characterize, capture, archive, and share (disseminate) the morphologic and molecular phenotypes of the precancer lesions including the precancer immune microenvironment, for all of the lesions sampled in the proposal.

In order to achieve this objective, three subaims are proposed.

**4.1: Precancer phenotype analysis.** The morphologic subtypes and immunophenotypes will be reviewed and categorized by study pathologists. Immunophenotype will be determined by tissue conservative multiplex IHC methodology.

**4.2: Precancer tumor immune microenvironment analysis.** Leveraging the consensus immune cell characterization panels developed in the MCL pathology working group, we will evaluate the immune cell repertoire in and around the precancers.

**4.3: Multidimension histology and IHC image archive.** Where possible (as in laser capture methods) pre and post laser capture images will document precise populations sampled for the atlas. In addition, serial sections will be used in aims 4.1 and 4.2 above and these data digitized and arrived. Metadata will include all clinical, radiologic report, and pathology report data abstracted to utilize defined data elements from the MCL/EDRN consensus vocabulary server. Note: radiologic images can also be included. We will use the JPL LabCAS system as the initial repository with the expectation that this will be linked to NIH databases or cBioportal for dissemination.

**Methods.**

***4.1* Precancer Phenotype Analysis**

True serial sections will be cut fresh from (frozen OCT or FFPE) tissue blocks for a. Pilot H&E, b. LCM (when needed), c. additional phenotype (see **Table 2),** and d. Immumo-profile (see 4.2). In the case of frozen section blocks, whenever possible the residual frozen section blocks will be converted into FFPE control blocks using standard methods. For each of the organ sites there are well characterized patterns and phenotypes of precancer and *in situ* neoplasia. In some cases, the precancer lesions will be associated with an adjacent or nearby invasive carcinoma, for which specific phenotyping is also planned for this study.

**Breast:** This project focuses on ductal carcinoma in situ (DCIS) which comprises multiple morphologic subtypes, a range of cytologic nuclear grade, presence and expent of necrosis[62,63], variable proliferative rate, as well as phenotypic diversity with respect to ER, PR[64] and Her2[65] and other hormone receptors[66], and

prognostically validated IHC biomarkers[67]. More recently a gene expression panel combining these marker (measured by rtPCR) applies an algorithm to score recurrence risk (OncotypeDCIS), though similar algorithmic analysis can be generated from IHC analysis of the protein level expression[68]. This project could reasonably be extended to include a broader landscape of breast cancer risk marker lesions, and less well documented potential precursor lesions, such as lobular carcinoma in situ (LCIS) and atypical papillary lesions, respectively. Here we will focus on DCIS lesions, and try to include both lower grade, smaller lesions as well as high grade and more extensive lesions. Methods for gene expression applied to DCIS in the past have required a bias toward the larger high grade lesions in order to provide sufficient material for analysis, but here the technology should be capable of sequence analysis on even very small lesions. The effort to correctly and specifically define these input low grade DCIS lesions will employ digital imaging to facilitate expert consensus diagnosis across the consortium.

**Prostate:** In the prostate, the project focuses on high grade PIN lesions, and extends the current work on comparing lower and higher grade invasive lesions. Standard phenotypic features of the PIN lesions will be cataloged, including the morphologic and cytologic subtypes. Markers of basal cells aid in the distinction between in situ and invasive lesions, and AMACR will be employed as above using the multiplex chromogenic PIN4 stain, as well as increased Ki67 correlate with neoplasia. In addition, a small panel of markers to stratify PIN lesions based on the current state-of-the-science understanding will be applied including c-MYC (referred to as MYC) protein by IHC[5], telomere length by in situ hybridization[6] and the 5'ETS/45S rRNA chromogenic in situ hybridization[7].

**Lung:** Lung cancer appears to develop from specific phenotypes of precancer with predictable relationships to the invasive cancer phenotypes[69]. In 2015 the WHO revised the classification of lung lesions to include adenocarcinoma in situ (AIS) and minimally invasive adenocarcinoma (MIA) in addition to atypical adenomatous hyperplasia (AAH)[70]. In addition, the bronchial lining may undergo squamous metaplasia with subsequent dysplastic changes to the point of squamous cell carcinoma in situ. Markers to establish an alveolar origin (TTF1) versus squamous differentiation (p63) are sometimes helpful to distinguish, although often the pattern and location are sufficient. Neuroendocrine differentiation (CD56, Synaptophysin) may be under-recognized in well differentiated lesions. Most of the relevant ancillary lung markers are best derived from the sequence analyses proposed above (e.g. EGFR mutation). However, some markers may be important to add as IHC detection of protein. ALK, for example has been reported in early lesions including CMPT[71]. p53 expression is a surrogate for mutations which often stabilize the protein. Her2 amplification is accompanied by overexpression and seen in a subset of early adenocarcinomas[69], following from the breast data this may be present even more often in the precancer lesion as well.

**Pancreas:** There are three precancer lesions in the pancreas, PanIN, IPMN, and MCN each with variation in grade and expression of protein markers associated with progression. MUC1 is expressed in the normal pancreatic ducts and in invasive pancreatic ductal adenocarcinomas[72] and higher grade PanIN, but not in lower grade. Muc4 and Muc5 may be abberantly expressed as well[73]. CDX2 is a marker of the intestinal subtype differentiaton[74]. Proliferative rate (Ki67) correlates with grade, but may also be independently important. p16/Ink4a loss and p53 nuclear expression are common in higher grade lesions, and may reflect breaching of a senescence barrier to invasive neoplasia[75,76]. COX-2 is expressed in PanIN lesions following the trend of increase from normal pancreatic ducts to PanIN to adenocarcinoma. In addition, changes in the stroma (in addition to the immune environment in aim 4.2) may be critical. Here we propose to add IHC for hyaluronic acid (HA) and smooth muscle actin (SMA) to the marker panel[77].

| | ***Breast*** | ***Prostate*** | ***Lung*** | ***Pancreas*** |
|---|---|---|---|---|
| ***Categories and Subtypes*** | *DCIS, solid, cribriform, micropapillary, etg* | *PIN, papillary, cribriform, flat, foamy* | *AAH, SCCIS Mucinous, non-mucinous, other* | *PanIN, IPMN Gastric, intestinal, papillary, other* |
| ***Grade*** | *Nuclear, +/- necrosis* | *High only? Low excluded?* | *Cytologic grade* | *Low, High* |
| ***Standard Markers*** | *ER, PR, Her2* | *HMWK; AMACR; PTEN; ERG* | *Ttf1, p63/p40, CEA, ALK?; CD56, Synp* | *CDX2, Muc1, 14-3-3sigma, others?* |
| ***Additional IHC/chrgISH*** | *P53, AR, VDR, Bag1,Ki67, p16,* | *TelomereFISH, cMyc, Ki67;* | *p53, p16, Her2* | *MUC4, MUC5, p16, p53, Ki67* |

| | | | | |
|---|---|---|---|---|
| ***Markers*** | *cox2* | *ETS/45s ISH* | | |
| ***Immune Microenv.*** | *MCL consensus* | *MCL consensus* | *MCL consensus* | *MCL consensus* |
| ***Metadata*** | *Size/extent, margins, treatment (lump/mast; xrt, etc.); Staging (if assoc with cancer); Screen detected, incidental, other. Addl health and family history* | *Size/extent, margins, treatment (bx v. excision); Staging (if assoc with cancer); Screen detected, incidental, other. Addl health and family history* | *Size/extent, margins, treatment (bx v. excision); Staging (if assoc with cancer); Screen detected, incidental, other. Addl health and family history* | *Size/extent, margins, treatment (bx v. excision); Staging (if assoc with cancer); Screen detected, incidental, other. Addl health and family history* |
| | | | | |

**Table 2.**

4.2 Tumor Immune Microenvironment

The Pathology Working group has developed a consensus immune cell multiplex IHC assay (2x 8plex markers) appropriate for use on FFPE sections and useful in comparisons of the tumor immune microenvironment (TIME) across all organ sites. This panel, developed in collaboration with all of the MCL sites, and with strong industry support from Perkin Elmer and Roche/Ventana is currently moving from the development and optimization stage to validation, as a part of the trans-MCL pathology working group set-aside funds project. In this precancer project, we will apply the consensus panels to the serial sections obtained for the project. At minimum two serial sections are needed for this part of the project, to accommodate the two 8plex panels. Staining will be performed at the UCSF/UCD/UCSD site, with analysis and image archive also central. The data, images, and any reanalysis will accessible via the image archive system (4.3 below) and we will plan to have recurrent meetings to discuss the data and analysis with the site groups, the TIME core (UCSF/UCD/UCSD) and methods experts from the coordinating center.

*4.3* Multidimension histology and IHC image archive

Slide imaging (WSI) and archiving will be initially performed using the UCD CGP Spectrum database, which is the current repository for cohort image archives for the UCSF site group. The CGP Spectrum database is a highly modified version of the Aperio Spectrum Database (Leica Biosystems, IL) which is designed to be used in clinical medical practice. As such, it features HIPAA compliant authentication with permissions control, and audit trail recording for all data elements. It offers multi-platform web access to the database. There is no separate “seat-license” for users, only the cost of the system and system maintenance, a major benefit to the community of users, also enabling us to use the system for publically accessible archives and educational materials. The data is organized in hierarchical tiers defined by the overall Study, Experiments in the Study, Cohorts in the Experiment, Specimens in the Cohorts, Assays /Slides/Images of the Specimens. The Assays/Slides/Images are appended with analyses and annotations. Some analysis tools are built into the system, including FDA approved IHC quantitation algorithms. (NOTE: In collaboration with JPL and the MCL coordinating center we are currently working to migrate this data and the full functionality of this system in an open source architecture based on the JPL LabCAS system.) We aim to utilize the WSI database functionality to enable high quality interaction of expert pathologists for use in the laser capture methods proposed. Pre-capture images can be annotated by the experts marking specific areas for capture, and recording the consensus metadata related to these areas. Laser capture technicians can use this annotation to guide the procedure and produce post laser capture images to document precise populations sampled for the atlas. In addition, serial sections will be used in aims 4.1 and 4.2 above and these data digitized and arrived. Metadata will include all clinical, radiologic report, and pathology report data abstracted to utilize defined data elements from the MCL/EDRN consensus vocabulary server. Note: radiologic images can also be included.

This image and data repository will enable data sharing, with the expectation that this will be linked to NIH databases or cBioportal for dissemination.

**Analytic Approach**

This is a pilot study, with little pre-knowledge as far as expectations for final evaluation of the types and subtypes of the precancer lesions under examination. Nevertheless, the consortium is acutely interested in statistically rigorous analytic methodologies and striving to define the precise objectives of the study, hypotheses (and null-hypotheses) and validity of the conclusions. We anticipate that the precancers, despite

their often small size (and corresponding technical difficulty of this project) have genomic/phenotypic landscape complexity equal to that of invasive cancers. If true, we expect that for the pilot to have reasonable sample density we would need at least 3 examples of any subphenotype expected. For breast, this might be divided into ER+/HER+; ER+/HER-; and ER-/HER-, but this is only the simplest possible stratification. If we superimpose the major phenotypic patterns of solid, cribriform and micropapillary, or alternatively if we expect to find most of the "intrinsic subtypes" of invasive cancer at the precancer stage, it will be important to have 3x3(ER/HER categori)x3(morphology) *or* 3x8or9(intrinsic subtypes) so approximately 18 to 27 samples for this pilot should provide a good starting point. With respect to the tumor microenvironment, we have pre-knowledge that suggests that a. there is some immune reaction detectable around most precancers; b. most have a relatively sparse reaction, perhaps reflecting the in situ niche as partially immune-privileged; c. a minority have a dense inflammatory response; d. inflammatory responses vary in composition, and specific infiltrates are associated with recurrence risk. In the consensus immune panel we will analyze T cell subsets, B cells, NK cells, macrophages, activation level, and proximity mapping in both peripheral/adjacent stroma as well as intraepithelial compartments. PDL1 expression on precancer epithelial cells may be seen, and will be detectable by the assay as well.

Integrated studies. A goal of these studies is to characterize premalignant lesions arising in breast, lung, prostate and pancreatic cancers, and further characterized by distinguishing characteristics such as hormonal status and histological findings. The analyses of data derived from this project becomes a high-dimensional study in which we have a large number of features (gene expressions, somatic mutations and immunoprofiles) and a limited number of samples. After normalization so that these different types of features are measured on similar scales, we will apply k-means hierarchical cluster analysis to identify subsets of samples that show common features, irrespective of the reported histological findings[78]. We will then evaluate if histological findings correlate with the derived hierarchical clusters. We will also apply a supervised approach in which we identify subsets of variables that predict histological subtypes using an L-1 penalization to limit the number of variable that are retained in modeling.

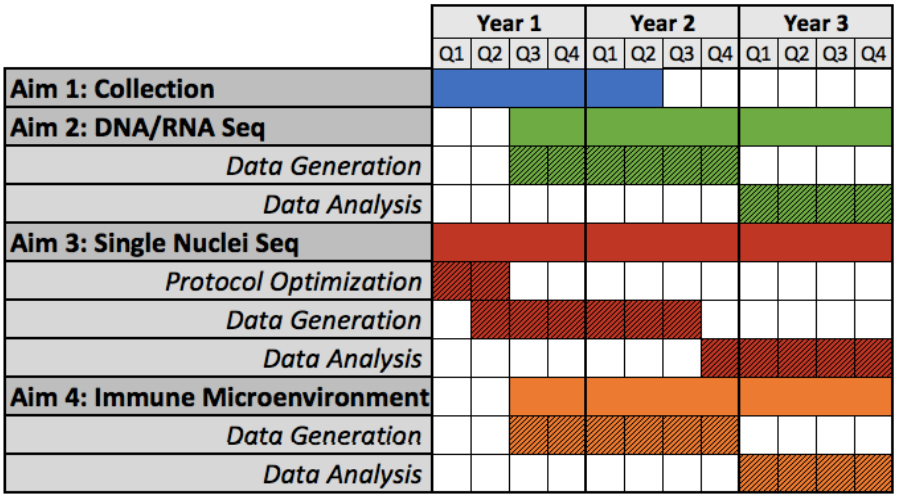
**D. Data Sharing and Dissemination**

**Individual Level Data Sharing:**
Copies of all genomic data, including the original uploaded FASTQ or raw data files, will be uploaded to investigator-restricted areas in the labCAS system that is maintained by the Jet Propulsion Laboratory. Participating groups may also process exome and RNA sequencing files to create local BAM and/or VCF files which may also be uploaded to labCAS. Derived .bam or .vcf files along with specific, relevant clinical data such as the type of cancer, histology, sex, and age in years will also be uploaded to the labCAS system. Finally, data derived from pathology studies in aim 4 will also be uploaded to the labCAS system. No PHI will be uploaded to the labCAS system but identifiers will be assigned by each center that tie the samples assayed by different technologies (such as immune profiling, RNA seq and exome seq). Data will be uploaded to dbGAP 6 months after all quality control steps have been completed on all of the studies performed on the samples. The CDMG will work with the contributing centers to obtain institutional certifications to identify data use restrictions for analysis of the data that are uploaded and to assist JPL in the transfer of data from the labCAS system to dbGAP. If an alternative platform is serving as the data repository for NIH funded research, CDMG and JPL will work with that archiving center. The Dartmouth Office of Sponsored Programs and IRB have extensive experience in interacting with NCBI for uploading data to dbGAP and will provide assistance through the administrative core, who will then assist investigators at the collaborating institutions in this process.

All data base management, tools and web resources have been developed using open source code that will be made available to qualified investigators. All database structures are developed as Health Insurance Portability and Accountability Act ("HIPAA") compliant. Dartmouth will track access and generate reports of use and retrievals by Participating Institutions. While the databases in place are HIPAA compliant, all participants have signed local IRB-approved consent documents at their participating institutions and data stored at the central repository at Dartmouth are all deidentified, so that no breaches of personal information are possible from the Dartmouth site.

**Intellectual Property**:
Inventions developed under this funding, will be reported to the participating institutions for evaluation and pursuit of intellectual property protection strategies. Participating Institutions may evaluate inventions developed under this funding to decide and implement the most appropriate intellectual property protection strategy for achieving the goals of this project. In cases where participating institutions and/or the researchers determine that patent protection is the best strategy for introducing and distributing such inventions it will pursue such strategies. All participating institutions will not knowingly engage in agreements with commercial sources that would hamper access by the academic research community to unique research resources developed under this funding. All participating researchers agree to adhere to the principles and guidelines set forth by the federal government with respect to making such unique research resources available to the qualified investigators within the scientific community.

## E. Timeline




## F. References


1. Wacholder, S. Precursors in Cancer Epidemiology: Aligning Definition and Function. *Cancer Epidemiol. Prev. Biomark.* **22,** 521–527 (2013).
2. Berman, J. J. *Precancer: The Beginning and the End of Cancer*. (Jones & Bartlett Learning, 2011).
3. Nasiell, K., Nasiell, M. & Vaćlavinková, V. Behavior of moderate cervical dysplasia during long-term follow-up. *Obstet. Gynecol.* **61,** 609–614 (1983).
4. Merrick, D. T. *et al.* Persistence of Bronchial Dysplasia Is Associated with Development of Invasive Squamous Cell Carcinoma. *Cancer Prev. Res. (Phila. Pa.)* **9,** 96–104 (2016).
5. Gurel, B. *et al.* Nuclear MYC protein overexpression is an early alteration in human prostate carcinogenesis. *Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc* **21,** 1156–1167 (2008).
6. Meeker, A. K. *et al.* Telomere shortening is an early somatic DNA alteration in human prostate tumorigenesis. *Cancer Res.* **62,** 6405–6409 (2002).
7. Guner, G. *et al.* Novel Assay to Detect RNA Polymerase I Activity In Vivo. *Mol. Cancer Res. MCR* **15,** 577–584 (2017).
8. Forbes, S. *et al.* COSMIC 2005. *Br. J. Cancer* **94,** 318–322 (2006).
9. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinforma. Oxf. Engl.* **25,** 2906–2912 (2009).
10. Le, K., Prabhakar, B. S., Hong, W. & Li, L. Alternative splicing as a biomarker and potential target for drug discovery. *Acta Pharmacol. Sin.* **36,** 1212–1218 (2015).
11. Yc, F., L, M., H, C. & Yl, L. Alternative splicing isoform of T cell factor 4K suppresses the proliferation and metastasis of non-small cell lung cancer cells. *Genet. Mol. Res. GMR* **14,** 14009–14018 (2015).
12. Inoue, K. & Fry, E. A. Aberrant Splicing of Estrogen Receptor, HER2, and CD44 Genes in Breast Cancer. *Genet. Epigenetics* **7,** 19–32 (2015).

13. Gimba, E. R. & Tilli, T. M. Human osteopontin splicing isoforms: known roles, potential clinical applications and activated signaling pathways. *Cancer Lett.* **331,** 11–17 (2013).
14. Li, Y., Rao, X., Mattox, W. W., Amos, C. I. & Liu, B. RNA-Seq Analysis of Differential Splice Junction Usage and Intron Retentions by DEXSeq. *PloS One* **10,** e0136653 (2015).
15. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355,** (2017).
16. Magi, A. *et al.* EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol.* **14,** R120 (2013).
17. Hiraoka, K. *et al.* Concurrent infiltration by CD8+ T cells and CD4+ T cells is a favourable prognostic factor in non-small-cell lung carcinoma. *Br. J. Cancer* **94,** 275–280 (2006).
18. Shimizu, K. *et al.* Tumor-infiltrating Foxp3+ regulatory T cells are correlated with cyclooxygenase-2 expression and are associated with recurrence in resected non-small cell lung cancer. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* **5,** 585–590 (2010).
19. Chen, J. J. W. *et al.* Up-regulation of tumor interleukin-8 expression by infiltrating macrophages: its correlation with tumor angiogenesis and patient survival in non-small cell lung cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **9,** 729–737 (2003).
20. Varn, F. S., Wang, Y., Mullins, D. W., Fiering, S. & Cheng, C. Systematic Pan-Cancer Analysis Reveals Immune Cell Interactions in the Tumor Microenvironment. *Cancer Res.* **77,** 1271–1282 (2017).
21. Dupont, W. D. & Plummer, W. D. Power and sample size calculations for studies involving linear regression. *Control. Clin. Trials* **19,** 589–601 (1998).
22. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356,** eaah4573 (2017).
23. Venteicher, A. S. *et al.* Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* **355,** (2017).
24. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541,** 331–338 (2017).
25. Habib, N. *et al.* DroNc-Seq: Deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq. *bioRxiv* 115196 (2017). doi:10.1101/115196
26. Yosef, N. & Regev, A. Writ large: Genomic Dissection of the Effect of Cellular Environment on Immune Response. *Science* **354,** 64–68 (2016).
27. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34,** 1145–1160 (2016).
28. Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539,** 309–313 (2016).
29. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352,** 189–196 (2016).
30. Singer, M. *et al.* A Distinct Gene Module for Dysfunction Uncoupled from Activation in Tumor-Infiltrating T Cells. *Cell* **166,** 1500–1511.e9 (2016).
31. Shekhar, K. *et al.* Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166,** 1308–1323.e30 (2016).
32. Genshaft, A. S. *et al.* Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. *Genome Biol.* **17,** 188 (2016).
33. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167,** 1853–1866.e17 (2016).
34. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33,** 495–502 (2015).
35. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).
36. Kowalczyk, M. S. *et al.* Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* **25,** 1860–1872 (2015).
37. Gaublomme, J. T. *et al.* Single-cell Genomics Unveils Critical Regulators of Th17 cell Pathogenicity. *Cell* **163,** 1400–1412 (2015).
38. Avraham, R. *et al.* Pathogen Cell-to-cell Variability Drives Heterogeneity In Host Immune Responses. *Cell* **162,** 1309–1321 (2015).
39. Trombetta, J. J. *et al.* Preparation of Single-Cell RNA-Seq Libraries for Next Generation Sequencing. *Curr. Protoc. Mol. Biol.* **107,** 4.22.1-17 (2014).

40. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510,** 363–369 (2014).
41. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396–1401 (2014).
42. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498,** 236–240 (2013).
43. Moffitt, J. R. *et al.* High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci.* **113,** 11046–11051 (2016).
44. Moffitt, J. R. *et al.* High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl. Acad. Sci.* **113,** 14456–14461 (2016).
45. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348,** aaa6090 (2015).
46. Chen, F. *et al.* Nanoscale imaging of RNA with expansion microscopy. *Nat. Methods* **13,** 679–684 (2016).
47. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* **92,** 342–357 (2016).
48. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11,** 360–361 (2014).
49. Habib, N. *et al.* Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science* **353,** 925–928 (2016).
50. Tirosh, I et al. Single-cell RNA-seq supports a developmental hierarchy in IDH-mutant oligodendroglioma. *Nature* (In Press).
51. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10,** 623–629 (2013).
52. Thomsen, E. R. *et al.* Fixed single-cell transcriptomic characterization of human radial glial diversity. *Nat. Methods* **13,** 87–93 (2016).
53. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17,** 77 (2016).
54. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2,** 666–673 (2012).
55. Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7,** 709–715 (2010).
56. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33,** 503–509 (2015).
57. Larsson, C., Grundberg, I., Söderberg, O. & Nilsson, M. In situ detection and genotyping of individual mRNA molecules. *Nat. Methods* **7,** 395–397 (2010).
58. Mignardi, M. *et al.* Oligonucleotide gap-fill ligation for mutation detection and sequencing in situ. *Nucleic Acids Res.* **43,** e151 (2015).
59. Ke, R. *et al.* In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10,** 857–860 (2013).
60. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).
61. Novershtern, N. *et al.* Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis. *Cell* **144,** 296–309 (2011).
62. Scott, M. A. *et al.* Ductal carcinoma in situ of the breast: reproducibility of histological subtype analysis. *Hum. Pathol.* **28,** 967–973 (1997).
63. Sneige, N. *et al.* Interobserver reproducibility of the Lagios nuclear grading system for ductal carcinoma in situ. *Hum. Pathol.* **30,** 257–262 (1999).
64. Allred, D. C. *et al.* Adjuvant tamoxifen reduces subsequent breast cancer in women with estrogen receptor-positive ductal carcinoma in situ: a study based on NSABP protocol B-24. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **30,** 1268–1273 (2012).
65. Allred, D. C. *et al.* HER-2/neu in node-negative breast cancer: prognostic significance of overexpression influenced by the presence of in situ carcinoma. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **10,** 599–605 (1992).
66. Santagata, S. *et al.* Taxonomy of breast cancer based on normal cell phenotype predicts outcome. *J. Clin. Invest.* **124,** 859–870 (2014).
67. Kerlikowske, K. *et al.* Biomarker expression and risk of subsequent tumors after initial ductal carcinoma in situ diagnosis. *J. Natl. Cancer Inst.* **102,** 627–637 (2010).

68. Dowsett, M. *et al.* Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **31,** 2783–2790 (2013).
69. Kim, E. K., Kim, K. A., Lee, C. Y. & Shim, H. S. The frequency and clinical impact of HER2 alterations in lung adenocarcinoma. *PloS One* **12,** e0171280 (2017).
70. Travis, W. D. *et al.* The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* **10,** 1243–1260 (2015).
71. Taguchi, R. *et al.* A case of anaplastic lymphoma kinase (ALK)-positive ciliated muconodular papillary tumor (CMPT) of the lung. *Pathol. Int.* **67,** 99–104 (2017).
72. Noroski, J., Mayo, D. & Kirschbaum, J. J. Liquid chromatographic resolution of the isomers of tipredane and phenylthioproline using urea-solubilized beta-cyclodextrin in the mobile phase. *J. Pharm. Biomed. Anal.* **10,** 447–455 (1992).
73. Lüttges, J., Zamboni, G., Longnecker, D. & Klöppel, G. The immunohistochemical mucin expression pattern distinguishes different types of intraductal papillary mucinous neoplasms of the pancreas and determines their relationship to mucinous noncystic carcinoma and ductal adenocarcinoma. *Am. J. Surg. Pathol.* **25,** 942–948 (2001).
74. Fukumura, Y., Nakanuma, Y., Kakuda, Y., Takase, M. & Yao, T. Clinicopathological features of intraductal papillary neoplasms of the bile duct: a comparison with intraductal papillary mucinous neoplasm of the pancreas with reference to subtypes. *Virchows Arch. Int. J. Pathol.* **471,** 65–76 (2017).
75. Abe, K. *et al.* Different patterns of p16INK4A and p53 protein expressions in intraductal papillary-mucinous neoplasms and pancreatic intraepithelial neoplasia. *Pancreas* **34,** 85–91 (2007).
76. Miyazaki, T. *et al.* Molecular Characteristics of Pancreatic Ductal Adenocarcinomas with High-Grade Pancreatic Intraepithelial Neoplasia (PanIN) Are Different from Those without High-Grade PanIN. *Pathobiol. J. Immunopathol. Mol. Cell. Biol.* **84,** 192–201 (2017).
77. Feig, C. *et al.* The pancreas cancer microenvironment. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **18,** 4266–4276 (2012).
78. Soliman, A. S. *et al.* Contrasting molecular pathology of colorectal carcinoma in Egyptian and Western patients. *Br. J. Cancer* **85,** 1037–1046 (2001).