Predicting colon cancer metastasis through spatial molecular characterization of the tumor immune microenvironment

Project Relevance

Every year, over 150,000 Americans are diagnosed with Colorectal Cancer (CRC), and annually over 50,000 individuals will die from CRC, necessitating improvements in screening, prognostication, disease management, and therapeutic options. Spatial patterns of tumor infiltrating lymphocytes (TIL) – T cell, NK cells and B cells near the primary tumor site – and their concomitant molecular alterations can determine whether there is potential for concurrent nodal and/or distant metastasis, which can indicate the risk of recurrence and mortality. In this proposal, we aim to develop low-cost assessments for TIL-specific whole transcriptomic molecular alterations that can predict whether a tumor has or will metastasize, and the risk of tumor recurrence to complement existing disease management strategies.

Project Summary
Colorectal Cancer (CRC) is both the third most common form of cancer and cause of cancer-related deaths in the United States. Examination of axillary lymph nodes at the time of surgical resection is essential for prognostication and while it is important to maximize the number of lymph nodes assessed, recent population-based studies have shown that evaluation of lymph node involvement is usually incomplete or inadequate. This can impact the accuracy of tumor staging and downstream disease management options, such as whether the patient should receive adjuvant chemotherapy. Developing alternative assessment methods which assess lymph node involvement through indirect mechanisms would be illuminating in cases where resection is inadequate. Tumor-infiltrating lymphocytes (TIL) and other immune cell types are important prognostic indicators in CRC. The type, density, and location of TILs with respect to the tumor, in addition to tumor-specific somatic alteration profiles, can determine TIL's effect on prognosis. Furthermore, spatially dependent, immune cell specific, proteomic and transcriptomic expression patterns inside and around tumor – the Tumor Immune Microenvironment (TIME) – can discern the coordinated immune response to tumor metastasis. The comprehensive characterization of TILs is possible using highly multiplexed spatial omics technologies, but high cost and low throughput prevent their clinical deployment. *Virtual staining* can infer molecular information at low cost from tissue histology where the morphology allows. We aim to design a low-cost *Virtual Staining* test, distilled from highly multiplexed spatial molecular information, that could complement surgical lymph node dissection for recurrence risk assessments and compete with other emerging predictors (e.g., circulating tumor DNA). In a set of stage III tumors with or without nodal and/or distant metastases, we will identify spatial proteomic and whole transcriptomic markers of metastasis with digital spatial profiling and Visium spatial transcriptomics of immune cells. We will also assess upstream cell-type specific DNA methylation alterations concomitant with spatial architectural TIME changes. Identified markers will be validated through lower-cost multiplexed immunofluorescence staining. Finally, we will establish histological correspondence to identified spatial metastasis markers and develop virtual staining algorithms to convert H&E-stained tissue into validated multiplexed immunofluorescent and whole transcriptomic markers. Spatial and cell-type specific patterns of molecular markers that indicate whether a patient has or is likely to develop metastasis will be identified under this framework. Inferring such information from tissue morphology can provide a low-cost and highly interpretable adjunct molecular assessment for lymph node resection, to predict recurrence risk and response to adjuvant chemotherapy. We expect that our findings will provide preliminary data for an R01 clinical trial to compare identified markers prospectively to independent metastasis predictors (e.g., liquid biopsy) for their ability to assess patient prognosis and treatment options.

Specific Aims

Colorectal cancer (CRC) has an annual incidence in the United States of approximately 150,000 new cases and a 63% 5-year survival rate. Successive invasion into epithelial, laminar, submucosal and other layers of colon is prognostic, where higher tumor stage reflects invasion depth. Although assessing regional lymph node involvement is also important for determining prognosis (e.g., risk of recurrence), variable resection quality can preclude adequate assessment for disease staging, necessitating the exploration of indirect molecular profiling methods to infer the presence of nodal metastasis[1,2].

Tumor-infiltrating lymphocytes (TIL) are an important prognostic indicator as they mediate direct antitumor cell immune responses and contribute to the recruitment [3], activation, and maturation of other immune cells [4,5]. Previous studies have shown that the type, functionality, density, and location of TILs within the tumor microenvironment (TME) inform the immune response and its anti-tumoral effectiveness. Various tumor-specific characteristics, including mismatch repair alterations, determine TIL's effects on TME and prognosis.

This proposal aims to understand how spatial patterns of highly multiplexed molecular markers (proteomic and whole transcriptomic) and cell-type specific (DNA methylation) can serve as indicators of concurrent nodal and distant metastasis and how highly multiplexed findings can be distilled into a low-cost adjunct test to improve recurrence risk assessment. For inadequately dissected lymph nodes, this adjunct test can communicate the confidence in the examination from spatial molecular information at the primary site. Morphological information from whole slide images combined with spatial transcriptomics will inform prognostication through low-cost and rapid histological inference of prognostic molecular information[6].

***Aim 1* will establish spatial proteomic and whole transcriptomic markers** associated with tumor metastasis. For stage III CRC tumors, spatial proteomic and transcriptomic profiling of immune cells within regions of interest will be conducted across age, sex, MMR alteration, and grade matched patients. Spatial profiling will consider macroarchitectural components labeled by the pathologist (intratumoral, tumor immune interface, and away from tumor). Mixed Effects Machine Learning methodologies will identify potential effect modifiers to follow up on for clinical findings. Spatial whole transcriptomic analyses will complement our spatial protein expression findings and can reveal upstream recruitment factors (e.g., cytokines).

***Aim 2* will establish proportions of 17 cell types in the TME and cell-specific DNA Methylation (DNAm) alterations** through application of a novel tumor deconvolution approach. DNAm cell-typing will yield cell-type specific metastasis-associated alterations upstream of proteomic and transcriptomic markers.

***Aim 3* will examine the reproducibility and scalability of identified markers** through concordance assessments with alternative multiplexed staining (e.g., multiplexed immunofluorescence). In addition, we will establish histological correspondence to identified spatial metastasis markers through joint modeling of the histomorphology, spatial transcriptomics, and protein staining. An ML model will be developed to predict identified markers from matched immunofluorescence stains through *Virtual Staining* of H&E slides (*VirtualProtein*) and will feature application of graph neural networks on cellular histomorphology to localize whole transcriptomic signatures at cellular resolution (*VirtualRNA*) [7–9].

This proposal will dovetail with efforts in the Center for Quantitative Biology (CQB) COBRE Single Cell Genomics Core (SCGC) to expand spatial molecular assessments and will facilitate collaboration among the CQB COBRE, the Pathology Shared Resource (PSR), Center for Molecular Epidemiology (Epi) COBRE Biorepository Core, and the bioMT COBRE Molecular Interactions & Imaging Core (MIIT) Microscopy Shared Resource (MSR). The PI has assembled a team of clinician scientist collaborators in the departments of Pathology, Dermatology, and Medical Oncology to increase the translational impact of study findings through a team science effort. These findings will be back translated to inform basic science research and the creation of computational biology tools that facilitate advanced spatial omics analyses. Access to developed software will be democratized through incorporation into the CQB COBRE Data Analytics Core's (DAC) analysis suite.

Identifying metastasis related molecular alterations at the primary site will inform the sufficiency of resected lymph nodes for recurrence risk assessment and CRC disease management options (i.e., adjuvant chemotherapy). Findings will elucidate how spatially dependent lymphocyte expression patterns inform a coordinated immune response to nodal metastasis. We expect research findings to fuel a clinical trial, funded through an R01 grant mechanism (e.g., https://grants.nih.gov/grants/guide/pa-files/PAR-22-131.html), which will compare the adjunct test prospectively to independent indicators of metastasis and recurrence risk: 1) CDX2 expression, 2) circulating tumor DNA (ctDNA) markers, and 3) immunoscore[10–12]. As compared to these technologies, we expect the adjunct test to balance tradeoffs in efficiency, cost, access, and comprehensiveness.

Research Strategy

Significance

Burden of Colon Cancer and Assessment Challenges. Colorectal cancer (CRC) is the third leading cause of cancer both worldwide and in the United States and accounts for approximately 8 percent of cancer-related deaths[13]. CRC incidence is shifting towards younger demographics who are not included in established screening programs[14,15]. While modifying specific risk factors (e.g., epigenetics, environment, diet)[16–18] can be effective in informing and/or curbing CRC incidence, there is a concurrent and vital need to develop more accurate, faster, and lower cost solutions for CRC screening and prognostication– and this is the goal of the work detailed in this proposal. Disease management of CRC often includes lymph node resection to determine N-stage as a proxy for recurrence risk, after resection at the primary site at the time of diagnosis. This is followed by adjuvant therapy for patients with positive lymph nodes. Lymph node resection, histology and grossing is often suboptimal outside of a subspecialist-driven academic medical *Center of Excellence*. As an example, one population-based study concluded that only 37% of colon cancer cases had adequate assessment of the regional lymph nodes (at least 12 nodes assessed) [19–22]. Inadequate resection and downstream analysis can impact prognostication and selection of relevant treatment options for clinical triage.

Role of Tumor Infiltrating Lymphocytes on Prognostication. The importance of Tumor Infiltrating Lymphocytes (TIL) on characterizing and modulating the Tumor Microenvironment (TME) and Tumor Immune Microenvironment (TIME)[23] to both prognosticate and establish novel immunotherapies cannot be understated but has been understudied. The TME is represented by an amalgamation of malignant and benign cells, blood vessels, and extracellular matrix, networked with complex communication patterns through the secretion of cytokine recruitment factors [23–25]. Many recent studies have shown that T cell, B cell, NK cell, and other monocyte/lymphocyte immune infiltrates and their spatial distribution, density and relationships play an important role in providing a coordinated antitumoral response, modified by Microsatellite Instability (MSI) status[26,27]. Information on cell-type specific molecular alterations (e.g., DNA Methylation, transcriptome expression) within these spatial arrangements related to colon cancer metastasis has not been fully elucidated.

Spatial Localization of Immune Signatures. The development of spatial omics technologies such as 10x Genomics Spatial Transcriptomics (ST) or GeoMX Digital Spatial Profiling (DSP) has enabled multiplexing findings (e.g., whole transcriptome, WTA) at incredible spatial resolution[28]. Existing applications of spatial profiling include comparisons of TIL subpopulations across the TME and automated scoring systems that have been developed which infer TIL information from standard morphology stains as a digital biomarker (*virtual stain*)[29–32], with limited translational application[33]. Despite the breadth of literature, few studies thus far have attempted to connect comprehensive, high-multiplexing of TILs offered by spatial omics technologies with the capacity to predict lymph node involvement and recurrence risk, and the potential to inform assessments at low-cost and high-throughput through molecular information inferred from tissue morphology [34,35].

Innovation

This proposal aims to improve disease prognostication through a low-cost and high throughput *virtual staining* spatial molecular assessment of tumor metastasis which can complement inadequate lymph node resection evaluations. The ***innovations*** of this proposal are three-fold: 1) investigate spatial cell-type specific proteomic, transcriptomic and DNA methylation changes associated with metastasis at an unprecedented scale, 2) explore a low-cost validation framework (i.e., adjunct test) that can infer the molecular associations from the primary site histology (i.e., *virtual staining*), and 3) develop innovative statistical and machine learning (ML) computational biology methods that can further inform the disease epidemiology (e.g., effect modifiers) while remaining cognizant of the potential impact of batch variation or spatial autocorrelation. These innovations are ***improvements*** on previous assessments as they: 1) provide higher multiplexing, assessed at a larger sample size, with more accurate cell-typing, 2) emphasize the translational potential of an informative *virtual* spatial molecular readout for oncologists, and 3) address the batch-level variation inherent to spatial omics assays. The far-reaching ***impact and applications*** of these innovations include: 1) serving as an adjunct assessment to lymph node resection procedures, augmenting traditional resection evaluation as a “second-check” mechanism for non-experts, or obviating the need to perform the invasive procedure entirely, 2) providing additional information for the oncologist to follow up for selection of optimal therapeutics, and 3) reducing cost barriers to local and distant metastasis assessment for global health applications. Conceptually, the ***goal of the adjunct test*** is to infer the degree of lymph node involvement or recurrence risk in situations where lymph node assessment is biased/inadequate while providing additional prognostic information, with similar power at lower cost as compared to highly multiplexed molecular profiling[36] and other independent prognostic indicators [10–12,37].

Approach

Overview. The aims of this proposal represent distinct assessment methods: Aim 1) spatial molecular profiling via a) lymphocyte-specific proteomic profiling and b) RNA profiling, Aim 2) assessment of cell-type specific DNA methylation (DNAm), and Aim 3) establishment of a low-cost adjuvant assessment method through a) independent validation with immunofluorescence (IF) and b) virtual staining with tissue morphology (H&E). Aims are divided into subaims, where appropriate (e.g., 1a–proteomic, 1b–RNA profiling). Subaims are broken down into subsections which detail: 1) tissue molecular profiling and imaging (**Profile**), 2) data analysis (**Assess**), 3) expected key findings (**Deliverable**), and 4) preliminary data, statistical power, limitations and minimum success criteria (**Feasibility**). Subsections will reference data collected from specific tissue sections (**Data Collection**).

Data Collection. IRB approval has been obtained for this study. In collaboration with the Pathology Shared Resource (PSR), we will access 150 tissues from stage III Colon adenocarcinoma patients biobanked in Pathology. Cases were identified through a retrospective search of pathology reports and slides at Dartmouth Hitchcock Medical Center (DHMC), a *Center of Excellence* with high quality nodal data, from 2016 to 2022. Stage was determined using the pTNM staging system, which balances local invasion versus metastasis as an overall prognostic marker. By restricting the stage, we can identify markers that provide prognostic value beyond that offered through the current prognostic staging system. Demographic information and tissue characteristics will be collected for each specimen as detailed in **Table 1**. Tissue biopsies from the patient will be partitioned into tissue blocks that will be sectioned into 5-micron thick layers (**Figure 1;** description of sections in **Table 1**).

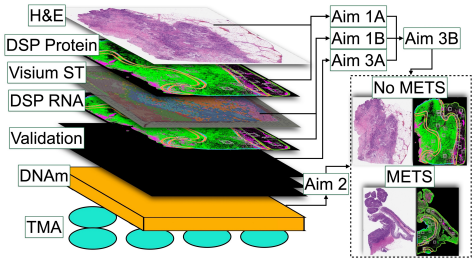


Figure 1: Overview of specimen sectioning

Table 1: Tabular description of demographics, tissue characteristics and assayed tissue sections

Demographics & Tissue Characteristics		Block Sections (5 um thick)	Assay	Associated Aim
Age	Invasion (T-Stage)	Layer 1 (n=150)	H&E	Aim 1
Sex	Nodal Metastasis (N-Stage)	Layer 2 (n=64)	Proteomics DSP, IF, Co-registered H&E	Aim 1a
Tumor Grade	Distant Metastasis (M-Stage)	Layer 3 (n=32)	Visium ST, Co-registered H&E	Aim 1b
Tumor Site (e.g., cecum, rectum)	Tissue Slides (Lymph Nodes)	Layer 4 (n=16)	RNA DSP, IF, Co-registered H&E	Aim 1b
Gross Tissue Dimensions	MMR Gene Loss Expression (MLH1, PMS2, MSH2, MSH6), using IHC	Layer 5 (n=150)	Validation IF Stains, Co-registered H&E	Aim 3
		Layer 6 (n=64)	DNAm (30-um thick)	Aim 2
Macroarchictectural Annotations	Time of Recurrence or Right Censoring	Layer 7 (n ≤ 64)	Tissue microarrays (TMAs), H&E scored	Aim 3, if needed/costs

DSP slides will be IF stained for the following markers: SYTO (nuclei), CD45 (immune cell), and PanCK (epithelial/tumor) for targeted profiling of lymphocytes. A GI pathologist will view H&E and IF whole slide images (WSI) simultaneously to annotate sections and label any spatially resolved information by three distinct macroarchitectural regions: intratumoral (*intra*), tumor-immune interface (*inter*) and away from tumor (*away*).

Aim 1 Establish key TME spatial molecular alterations of concurrent nodal and/or distant metastasis

Aim 1a: Establish lymphocyte-specific spatial immuno-oncology protein markers of metastasis

1a-Profile: Proteomic Digital Spatial Profiling (DSP) of lymphocytes within distinct macroarchitectural contexts. We will profile 32 patients without metastasis and 32 patients with nodal metastasis (half of these patients with both nodal and distant metastasis) (n=64 patients). For each patient, 24 regions of interest (ROI) will be placed among the 3 macroarchitectural regions (1,536 total ROIs/nested observations)[38]. Immune cells are isolated within each ROI via image segmentation of the CD45 stain and profiled through targeted ultraviolet (UV) cleavage of attached oligo tags. The Nanostring nCounter will quantify immune cell protein expression across 40 immuno-oncology markers. Returned data will include protein expression measurements for each ROI, tagged with x,y coordinate, an ROI-specific nuclei count, and co-registered H&E and IF slides from same section. ROIs will be filtered based expression relative to negative control, normalized, and log2 transformed.

1a-Assess. Develop methods to identify differentially expressed clinical markers of nodal/distant metastasis in the TME, accounting for batch and spatial autocorrelation. Report effect modifiers with markers and clinical characteristics. Identify groups of metastasis markers with gene modules. Bayesian hierarchical linear regression models will establish associations with metastasis (*mets* used to indicate nodal metastasis only, distant only, nodal or distant, nodal and distant): $log_2(protein_i) = \beta_0 + \beta_1 mets_i + \beta_2 TME_i + \beta_3 mets_i * TME_i + \beta_4 \text{MLH1 loss}_i + \beta_5 age_i + \beta_6 sex_i + \theta_{batch[i]} + \theta_{patient[i]} + \theta_{\Sigma(s_i,s_j)} + \epsilon_i$. An interaction term between *mets* and *TME* evaluates metastasis conditioned on the macroarchitecture (*intra*, *inter*, *away*), adjusting for potential confounding (MLH1 loss, age, sex) (**Figure 2A**)[39,40]. Batch and case-level variation

are captured with random intercepts, θ, and $\theta_{\Sigma(s_i,s_j)}$ represents a gaussian process that accounts for correlation between adjacent observations, s_i, s_j, as an inverse function of their distance. Effect estimates will be communicated using the median posterior sample of the effect estimate, 95% high-density posterior credible interval, posterior probability of direction (*pd*), with *post hoc* comparisons via *emmeans*[41,42].

Most ML methods for omics data assume independence of nested observations, which can distort study findings when applied to spatial omics. We will develop a classifier to estimate the probability of tumor metastasis based on all markers (x_i) by fitting tree boosting models, $f_\phi(\vec{x_i})$, in a Mixed Effects Machine Learning modeling framework (MEML), e.g., leveraging Gaussian Process Boosting and hierarchical Bayesian Additive Regression Tree (BART)[43,44]: $logit(p_i) = f_\phi(\vec{x_i}) + \vec{\beta} \cdot \vec{x} + \theta_{batch[i]} + \theta_{patient[i]} + \theta_{\Sigma(s_i,s_j)}$. Salient cross-level interactions identified from the MEML method will be used in a Bayesian hierarchical logistic regression model to report pertinent effect modifiers (e.g., effect of CD20, conditional on age) (**Figure 2A**)[38,45]: $logit(p_i) = \vec{\beta} \cdot \vec{x} + \theta_{batch[i]} + \theta_{patient[i]} + \theta_{\Sigma(s_i,s_j)}$. Interactions are encapsulated in $\vec{x}$. Performance will be evaluated using a C-statistic, the Bayes Factor and WAIC. Feature selection will incorporate the Horseshoe LASSO and projection predictive selection methods[46,47]. A weighted gene co-expression network analysis (WGCNA) will cluster genes into modules for association with disease metastasis (**Figure 2A**)[48].

1a-Deliverable: Identify stains for independent validation. As this is a proteomic assessment, **1a-Assess** findings will motivate lower-cost staining (*Aim 3A*) to recapitulate findings. Effect modifiers revealed in **1a-Assess** will suggest stains (or set of stains) specific to age, sex or pathway (e.g., MLH1-loss).

1a-Feasibility: Preliminary data collected on 32 patients informs study power and has revealed metastasis-related markers and effect modifiers, but more data is required to achieve significant findings. Potential limitations include batch effects, biased ROI selection, and potential for Type I error. To date, 840 ROI have been collected, leading to an initial publication of MEML effect modifiers (e.g., fibronectin and age found in **Figure 2A**) and a follow up study indicating clinical markers (e.g., intratumoral CD8/CD66b found in **Figure 2A**). An empirical power analysis was conducted based on 100 simulated datasets (binary outcomes drawn via $y_i \sim Binomial(1, p_i)$) using the statistical model in **1a-Assess** at a sample size of 64, given batch effects and estimated effect size of 0.2, yielded a power of 0.81 after multiplicity adjustment. While we plan to document DSP limitations through independent publications, AI automation techniques are being developed for the representative selection of ROI. Sensible weakly informative priors centered around 0 will guard against Type I error in addition to multiplicity adjustments[49]. **Minimum success criteria–** Identify at least three proteins or effect modifiers for validation.

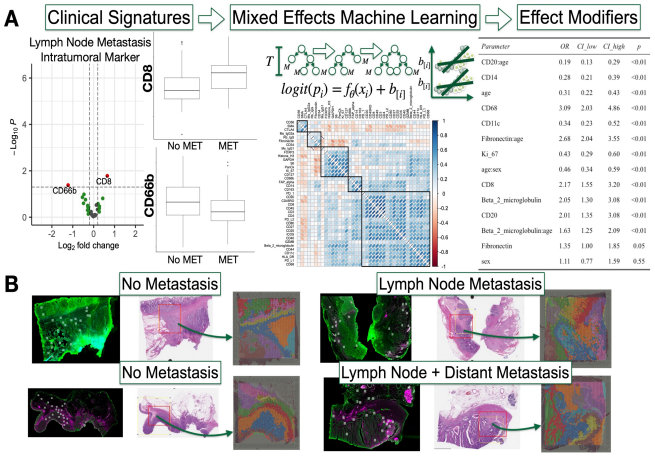


Figure 2: Aim 1: A) Identify clinical signatures (e.g., CD8/CD66b as intratumoral markers of metastasis) and effect modifiers via MEML (e.g., CD20 and age); B) Matched Visium mRNA profiles for WTA

Aim 1b: Uncover spatial cell-type transcriptomic signatures.

1b-Profile: Comprehensive transcriptomic profiling using the DSP RNA assay and Visium Spatial Transcriptomics (ST) platform. Up to 18,000 genes can be profiled using ST to complement proteomics. Thirty-two Visium ST slides (16 patients with nodal/distant metastasis), subset from the proteomics cohort, will be assayed. Capture areas (red squares; **Figure 2B**) will be placed by a pathologist based on overlap with existing proteomics measurements and to ensure adequate overlap with previous tissue macroarchitectural annotations (*intra/inter/away*). Capture areas will be manually cut by a histotechnician from the PSR and sent to the SCGC for profiling. ST uses spatially barcoded spots to capture polyadenylated mRNA molecules and register their coordinates within the capture area. ST will be compared to 16 DSP RNA slides (ST subset; 8 metastasis patients) profiled with in situ hybridization probes paired with an NGS readout. We favor ST over the DSP as is it does not require ROI selection. We plan to focus on ST because it offers a greater chance to elucidate spatial variation due to the comprehensive coverage of barcoded spots and to assist the SCGC in deploying the FFPE Visium protocol at scale.

1b-Assess: Apply methods from 1a-Assess. In addition– 1) correlate and compare DSP with ST to motivate transition from DSP to ST, 2) identify ST genes that exhibit spatial variation related to

metastasis, 3) infer spatial signaling patterns between cells, adjust for and estimation effect modification by cell-types through deconvolution, and 4) identify metastasis-related pathways. We will profile 384 DSP ROI based on representative clusters of ST expression patterns. Concordance between DSP and ST will be established through a negative binomial regression model, where DSP expression is the dependent variable, conditional on the nearest ST spot's expression as the independent variable, considering batch effects, nested observations from each profiling technique and information from adjacent spots in the form of lagged exogenous terms[50]. A negative binomial likelihood will be used to model mRNA expression. Separately, assays will be compared based on variance of effect estimates after applying **1a-Assess** methods.

We will leverage statistical and ML methods (e.g., SPARK, Moran's I, cluster analysis, Bayesian hierarchical gaussian process, graph convolutional networks) to identify spatially variable genes within and across slides by incorporating information from adjacent spots (i.e., spatial autocorrelation)[51–55]. Indices of spatial clustering will be compared through statistical comparison of Bayesian posterior distributions of autocorrelation parameters ($\rho_{\mathrm{I(mets)}}$) to identify which spatially variable genes exhibit similar/different clustering patterns between patients grouped by metastasis status and/or macroarchitecture.

Spots will be deconvolved into constituent immune cell-type proportions ($\vec{p}$) through established deconvolution approaches, some of which leverage spatial information[52,56–58]. Cell-cell graphs will be formed by constructing an adjacency matrix based on spatial proximity[59,60]. Methods which integrate information on ligand-receptor and cytokine expression with co-localized expression patterns will inform which cell-types are likely to interact. Effect estimates from **1b-Assess** will be compared to adjustment for and effect modification by cell-type (e.g., $mets_i * TME_i * p_{i,j}$; *jth*-cell type). Significant genes from **1b-Assess** will undergo gene set enrichment (GSEA) and pathway analyses via hypergeometric tests on well-known pathway databases (e.g., KEGG)[61].

1b-Deliverable: Similar to 1a-Deliverable, including spatially variable/interacting genes/cells, and pathways will inform knowledge on TIL cell-type specific molecular changes.

1b-Feasibility: Similar to 1a-Feasibility. Preliminary data collected in collaboration with PSR and SCGC. We have profiled four ST slides, according to **1b-Profile**, through a pilot initiative among the Levy Lab, PSR, DAC and SCGC. We successfully tested ST FFPE protocol and compared unsupervised gene expression clusters across slides (**Figure 2B**). Reagents for four matched DSP RNA WTA slides were collected for initial concordance assessments. We are generating preliminary findings from **1b-Assess**. Our team is well-positioned to execute on *Aim 1b* pending cohort expansion to capture patient-level variation. Multiplicity issues can be ameliorated through gene set weighting and Bayesian methods[62–64].

<u>Aim 2: Cell-Type Specific DNA Methylation Alterations Related to Metastasis</u>.

Cell typing with DNA Methylation (DNAm) is far more accurate than protein/RNA and DNAm alterations can set gene expression programming[65–68]. We will develop a DNAm assay for cell-type specific metastasis alterations that can be cheaper/easier to deploy than an RNA-based assay. We will tie in with *Aim 1* findings by correlating deconvolved RNA and DNAm cell types and binning identified ST genes by potential for DNAm dysregulation.

2-Profile: Genome-wide DNAm profiling. Epi COBRE Biorepository and CQB COBRE SCGC will use DNA for all cases (n=64) for bisulphite conversion and DNAm profiling by the Illumina EPIC methylation array [69]. Profiling will yield proportion of methylated alleles across the mixture (beta-value) at 850k CpG sites. DAC and Salas labs will assist in DNAm preprocessing[70].

2-Assess: Accurately resolve TME-related cell-types. Apply cell-types to identify epigenome-wide DNAm metastasis signatures (CpGs). Identify effect modifiers similar to 1a-Assess and pathways similar to 1b-Assess, without spatial adjustment. We will utilize a novel approach – **HiTIMED**: **H**ierarchical **T**umor **I**mmune **M**icroenvironment **E**pigenetic **D**econvolution – to deconvolve solid tumors into seventeen constituent cell types: tumor, epithelial, endothelial, stromal, basophil, eosinophil, neutrophil, monocyte, dendritic cell, B naïve and memory, CD4T naïve and memory, CD8T naïve and memory, T regulatory, and natural killer cells [71]. Metastasis-related differentially methylated CpGs will be identified through an epigenome wide association study (EWAS)[72], using the statistical model: $\{M_i\}_j = \beta_0 + \beta_1 mets_i + \beta_2 \mathrm{MLH1\ loss}_i + \beta_3 age_i + \beta_4 sex_i + \overrightarrow{\beta_{cell}} \cdot \overrightarrow{p_{cell}} + \theta_{batch[i]} + \epsilon_i$. *M* is the beta-value (CpG *j*) transformed to a normal distribution[73] to address heteroskedasticity. Models will adjust for and condition on cell type proportions ($\overrightarrow{p_{cell}}$; *K* cell-types; *K-1* assessed due to simplex constraint $\sum_i p_i = 1$). Effect estimates will be compared to unadjusted models. Conditional cell-type specific effects will use the CellDMC and Tensor Composition Analysis (TCA) approaches[74,75], which estimates differentially methylated cell types through the addition of interactions ($\overrightarrow{\beta_{cell-interact}} \cdot \overrightarrow{p_{cell}} * mets_i$) to the statistical model. MEML models (**1a-Assess**) incorporating public data (**2-Feasibility**) will identify additional

CpG-cell-type (e.g., $M_{ij} * p_{ik}$) and CpG-covariate interactions (e.g., $M_{ij} * z_{il}$), reported via: $logit(mets_i) = \beta_0 + \beta_1 M_{ij} + \beta_2 p_{ik} + \overrightarrow{\beta_3} \cdot \overrightarrow{z_i} + \beta_4 M_{ij} * p_{ik} + \beta_5 M_{ij} * z_{il} + \cdots + \theta_{batch[i]}$ for identified combinations of CpG *j* and cell-type *k* or covariate *l*. ML models will be compared to MethylCapsNet (developed by PI to report pathways) and Elastic Net [76]. CpGs will be associated with genes for pathway analysis after FDR adjustment[77,78].

2-Deliverable: Report dysregulated pathways by cell-type and corroborate with 1b-Deliverable. Validation of HiTIMED will encourage adoption as an immunomethylomics prognostication tool. We will overlap DNAm markers with ST genes by CpG promoter islands to inform a custom lower-cost Illumina BeadChip array.

2-Feasibility: We have profiled 32 samples. Incorporating TCGA data, Bayesian approaches, relaxing FDR adjustments, variance filtering, and assessing coarse cellular hierarchy will strengthen power of findings for 2-Assess. Further sample profiling will leverage SCGC resources and involve collaboration with PSR, DAC, and Epi COBRE Biorepository. Bayesian modeling approaches will allow us to set informative priors based on estimation of a posterior parameter estimates based similar patient characteristics from The Cancer Genome Atlas (TCGA), which will be compared to findings with non-informative priors. The ML parameters can be similarly initialized through TCGA pretraining. Using pwrEWAS, collecting 64 samples will increase statistical power from 0.62 (n=32) to 0.76 for the ability to find 2,500 differentially methylated CpGs from a set of 100,000 CpGs ($\beta = 0.3$) [79]. **Minimum success criteria–** apply HiTIMED to compare TME cell-type proportions.

Aim 3: Developing and Scaling Low-Cost Adjuvant Assessments as Clinical Trial Precursor

Aim 3a: Independent Validation of Stains Predictive of Metastasis

3a-Profile: Perform fluorescence imaging of top 10 mRNA/protein markers from 1-Deliverables. IF stains will be selected based on effect size, statistical significance, and perceived pathway relevance from TME literature. Two 6-plex IF stains (one stain for nuclei) will be imaged using the PerkinElmer Vectra3 (bioMT MSR) for 150 cases (TMAs if needed). Cells will be separated through image segmentation of the nuclear stain and tagged by patient ID, macroarchitecture, slide coordinates (x,y) and a 10-dimensional vector representing protein expression represented by 16-bit precision image intensity, thresholded to indicate positive staining.

3a-Assess: Recapitulate 1-Assess findings and demonstrate prognostic potential of low-cost stains through recurrence prediction. Differential expression will be compared to ordinal stain scores assigned by collaborating pathologists. A Cox proportional hazards model will predict recurrence risk from cells, covariate adjusted (conditional mean, **1-Assess**). As cases are from 2016-2021, we will assess 2/5-year recurrence[36].

3a-Deliverable: Similar to 1-Deliverable. Significant findings for 3a-Assess, compared to N-stage, would indicate added prognostic value. Effect estimates are expected to have greater statistical power than *Aim 1*.

3a-Feasibility: Existing IF stains collected with DSP profiling can trial 3a-Assess. RNA may be more difficult to validate in IF. For example, presence of PanCK at the tumor-immune interface quantified using the DSP was associated (**1a-Feasibility**; not pictured) with metastasis. PanCK target identified with DSP has matched IF stain for comparison. Validation of selected mRNA targets will depend on degree to which RNA translates to protein and may require **3b-Assess**. **Statistical power** is expected to be greater than that in *Aim 1* (**3a-Deliverable**). **Minimum Success Criteria–** Six of ten stains confirm equivalent/added prognostic value.

Aim 3b: *Virtual Staining* of Validation Stains and ST for Low-Cost Cellular Molecular Assessment

3b-Profile: Co-register same-layer morphology (H&E) and molecular (IF, ST spots; 1,3A-Profile) information. Set aside 150 serial section H&E WSI for validation (**Table 1**).

3b-Assess: Develop two Virtual Staining tools, *VirtualProtein* and *VirtualRNA*, to infer spatial molecular information from H&E, and assess imputed spatial molecular information similar to 3a-Assess. Develop tool for accurate detection of nuclei from tissue morphology to permute Virtual Staining at the cellular resolution. ***VirtualProtein*** **recapitulates spatial protein expression from H&E.** ***VirtualRNA*** **recapitulates spatial mRNA expression from H&E at cellular resolution.** *VirtualProtein* will be trained using a generative adversarial network (GAN) which takes as input the H&E WSI and outputs the pixel-wise presence and intensity of the multiplexed stains from **3A-Profile** (**Figure 3A**). Since GANs cannot operate on the entire WSI, the GAN will be modified to incorporate macroarchitectural information with Graph Neural Networks (GNN) for visual/predictive consistency [7–9,80–82].

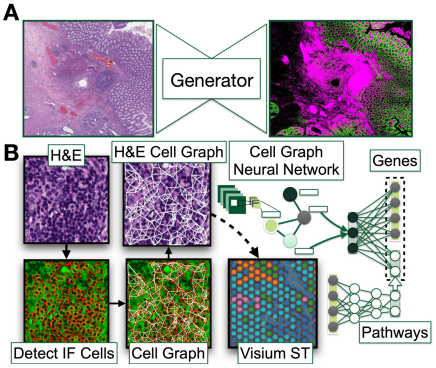


Figure 3: Aim 3– Virtual: A) Protein; B) RNA

We will use co-registered *IF* and *H&E* stains of nuclei and immune cells to build a highly accurate immune cell prediction tool by training an object detection neural network using the Detectron2 framework[83,84].

Graphs will be constructed over cells, identified by applying the detection network on **1B-Profile** H&E images, to form *cell graphs* based on spatial proximity. A convolutional neural network (CNN) will extract cell-level information from cell morphology and surrounding tissue context, stored as node attribute vectors [7,85–87].

VirtualRNA will train GNN to pool information from *cell graphs* to predict the mRNA expression patterns at an ST spot (**Figure 3B**). Gradient-based techniques will identify correspondent cells and their morphology by backpropagating information from predicted genes to the original cell morphology graphs/images. As inference on 18,000 genes is intractable, we will predict individual gene expression (ranked from **1-Assess**) and pathway-level mRNA expression (aggregated across relevant genes) using a *sparse pathway neural network*[76,88–91].

3b-Deliverable: Similar to 3a-Deliverable. In addition, develop an image viewer oncologists can use to interface with *Virtual Stain* results[92]. We anticipate *Virtual Stains* can be ordered at even lower cost compared to IF stains from *3A*. We aim to identify proteomic and mRNA markers which can be predicted from morphology and used to prognosticate at similar power as **3a-Deliverable**. We expect effect estimates to be more precise than Aims *1-2*, while circumventing costs, labor and turnaround time associated with staining.

3b-Feasibility: Similar to 3a-Feasibility. The PI has previously demonstrated the feasibility of *Virtual Staining*, GNN, and sparse pathway neural networks for digital pathology through published research. *VirtualRNA* places emphasis on cell-level features. The use of graph-based neural networks bridges the gap between macroarchitectural and cell-level features[7–9,83].

Collaborators/Mentors. Translational research team (e.g., oncology experts) described in budget justification.

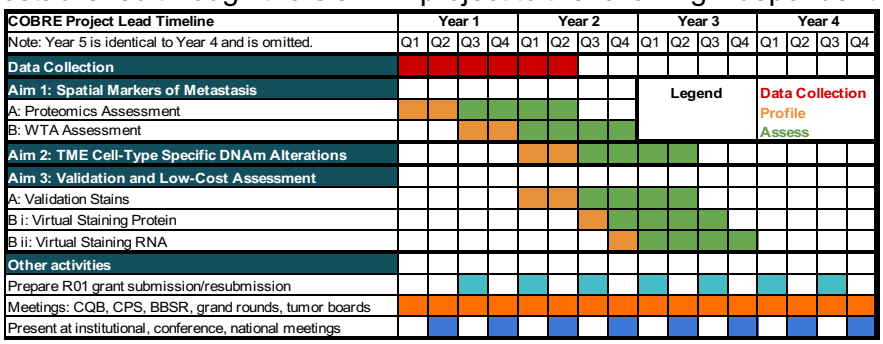
Plan to Leverage CQB Resources. The PI has met with faculty in SCGC and DAC to understand areas where the project will interface CQB cores. Chiefly, grant resources will be reinvested into the COBRE cores for large-scale Visium profiling, DNAm profiling, RNA quality control and data preprocessing/analysis help. Summarized are discussions between PI and CQB Core faculty: **SCGC** (Fred Kolling, PI)– 1) incorporate newly acquired technologies (merScope, HD Visium ST, spatial CITE-Seq) and deploy FFPE ST protocol, 2) up to 1/3 discount on large scale Visium ST, 3) collaborate on methods to ameliorate batch effects, 4) form interest group that connects DH faculty interested in spatial omics, 5) specimen triage; **DAC** (James O'Malley, PI)– 1) guidance on statistical methodology for grant aims, 2) DAC PI is interested in developing BART model where tree structure varies spatially instead of treating spatial clustering as a nuisance; PI Levy has previously published MEML interaction method with O'Malley, 3) benchmark ST against DSP, 4) software developed in COBRE project will be made publicly available through DAC HTML interface (Tim Sullivan), and 5) DAC (Owen Wilkins) will identify/leverage public single cell RNASeq data (Colon) to expand/impute ST markers[51,59,93–96].

Timeline: We have put together a timeline which details completion of the grant aims, publications, attendance/presentations at institutional group meetings and national conferences, and R01 submissions:

Publication Potential: We anticipate that project lead results will generate 10-15 publications that will both reflect well on the COBRE CQB and serve as initial data for further extramural grants. We plan to pursue an aggressive publication timeline that will prioritize intermediate findings from the grant aim deliverables. PI has a record of publication productivity (i.e., 18 first/senior author publications over past 3 years).

Presentation and Tumor Boards: The PI will solicit feedback from clinical and basic science investigators in the form of regular attendance and presenting his work at: 1) Oncology and GI Pathology tumor boards, 2) the COBRE CQB meetings, 3) Biostatistics and Bioinformatics Shared Resource (BBSR), 4) Cancer Population Sciences (CPS), 5) Cancer Center and Pathology grand rounds, and 6) national conferences on computational biology and bioinformatics. Attendance at these events will solidify the PI's understanding of pressing implementation challenges and presentations will facilitate collaboration amongst institutional stakeholders.

R01 Clinical Trial: PI and research team plan to submit R01 applications semi-annually. Multi-center R01 validation trial will compare the adjunct test established through the COBRE project to the following independent recurrence predictors: 1) CDX2 expression, 2) circulating tumor DNA (ctDNA) markers, and 3) immunoscore[10–12]. PI plans to interface with Mentoring and Administrative COBRE Core, who can provide guidance on building and maintaining national interinstitutional collaborations. PI has reached out and met with GrantsGPS for R01 preplanning.



References

1. Senthil, M., Trisal, V., Paz, I. B., *et al.* Prediction of the Adequacy of Lymph Node Retrieval in Colon Cancer by Hospital Type. *Archives of Surgery* **145**, 840–843 (2010). PMID: 20855753
2. Kamal, Y., Schmit, S. L., Hoehn, H. J., *et al.* Transcriptomic Differences between Primary Colorectal Adenocarcinomas and Distant Metastases Reveal Metastatic Colorectal Cancer Subtypes. *Cancer Research* **79**, 4227–4241 (2019). PMID: 31239274
3. Idos, G. E., Kwok, J., Bonthala, N., *et al.* The Prognostic Implications of Tumor Infiltrating Lymphocytes in Colorectal Cancer: A Systematic Review and Meta-Analysis. *Scientific Reports* **10**, 3360 (2020). PMID: 32099066
4. Bear, H. D. & Chin, C. S. 35 - Approaches to Adoptive Immunotherapy. in *Surgical Research* (eds. Souba, W. W. & Wilmore, D. W.) 415–434 (Academic Press, 2001). doi:10.1016/B978-012655330-7/50037-X.
5. Tseng, D., Schultz, L., Pardoll, D., *et al.* 6 - Cancer Immunology. in *Abeloff's Clinical Oncology (Sixth Edition)* (eds. Niederhuber, J. E., Armitage, J. O., Kastan, M. B., et al.) 84-96.e5 (Elsevier, 2020). doi:10.1016/B978-0-323-47674-4.00006-2.
6. Zhang, M., Sheffield, T., Zhan, X., *et al.* Spatial molecular profiling: platforms, applications and analysis tools. *Brief Bioinform* doi:10.1093/bib/bbaa145. PMID: 32770205
7. Levy, J., Haudenschild, C., Barwick, C., *et al.* Topological Feature Extraction and Visualization of Whole Slide Images using Graph Neural Networks. *Pac Symp Biocomput* 285–296 (2021) doi:10.1101/2020.08.01.231639. PMID: 33691025
8. Levy, J. J., Azizgolshani, N., Andersen, M. J., *et al.* A large-scale internal validation study of unsupervised virtual trichrome staining technologies on nonalcoholic steatohepatitis liver biopsies. *Modern Pathology* **34**, 808–822 (2021). PMID: 33299110
9. Levy, J., Jackson, C., Sriharan, A., *et al.* Preliminary Evaluation of the Utility of Deep Generative Histopathology Image Translation at a Mid-sized NCI Cancer Center. *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020) - Volume 3: BIOINFORMATICS* **3**, 302–311 (2020).
10. Bruni, D., Angell, H. K. & Galon, J. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. *Nat Rev Cancer* **20**, 662–680 (2020). PMID: 32753728
11. Dalerba, P., Sahoo, D., Paik, S., *et al.* CDX2 as a Prognostic Biomarker in Stage II and Stage III Colon Cancer. *N Engl J Med* **374**, 211–222 (2016). PMID: 26789870
12. Tarazona, N., Gimeno-Valiente, F., Gambardella, V., *et al.* Detection of postoperative plasma circulating tumour DNA and lack of CDX2 expression as markers of recurrence in patients with localised colon cancer. *ESMO Open* **5**, e000847 (2020). PMID: 32967918
13. Wong, M. C., Huang, J., Lok, V., *et al.* Differences in incidence and mortality trends of colorectal cancer worldwide based on sex, age, and anatomic location. *Clinical Gastroenterology and Hepatology* **19**, 955–966 (2021). PMID: 32088300
14. Kasi, P. M., Shahjehan, F., Cochuyt, J. J., *et al.* Rising proportion of young individuals with rectal and colon cancer. *Clinical Colorectal Cancer* **18**, e87–e95 (2019). PMID: 30420120
15. Patel, S. G. & Ahnen, D. J. Colorectal cancer in the young. *Current gastroenterology reports* **20**, 1–12 (2018). PMID: 29616330
16. Lao, V. V. & Grady, W. M. Epigenetics and colorectal cancer. *Nature reviews Gastroenterology & hepatology* **8**, 686–700 (2011). PMID: 22009203
17. Cheng, E., Ou, F.-S., Ma, C., *et al.* Diet-and Lifestyle-Based Prediction Models to Estimate Cancer Recurrence and Death in Patients With Stage III Colon Cancer (CALGB 89803/Alliance). *Journal of Clinical Oncology* JCO-21 (2022). PMID: 34995084
18. Slattery, M. L. Diet, lifestyle, and colon cancer. *Semin Gastrointest Dis* **11**, 142–146 (2000). PMID: 10950460
19. Baxter, N. N., Virnig, D. J., Rothenberger, D. A., *et al.* Lymph node evaluation in colorectal cancer patients: a population-based study. *J Natl Cancer Inst* **97**, 219–225 (2005). PMID: 15687365

20. Hartgrink, H. H., Velde, C. J. van de, Putter, H., *et al.* Extended lymph node dissection for gastric cancer: who may benefit? Final results of the randomized Dutch gastric cancer group trial. *77* (2004). PMID: 15082726
21. Schofield, J. B., Mounter, N. A., Mallett, R., *et al.* The importance of accurate pathological assessment of lymph node involvement in colorectal cancer. *Colorectal Disease* **8**, 460–470 (2006). PMID: 16784464
22. Ong, M. L. & Schofield, J. B. Assessment of lymph node involvement in colorectal cancer. *World journal of gastrointestinal surgery* **8**, 179 (2016). PMID: 27022445
23. Binnewies, M., Roberts, E. W., Kersten, K., *et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat Med* **24**, 541–550 (2018). PMID: 29686425
24. Jakubowska, K., Koda, M., Kisielewski, W., *et al.* Tumor-infiltrating lymphocytes in primary tumors of colorectal cancer and their metastases. *Exp Ther Med* **18**, 4904–4912 (2019). PMID: 31807155
25. Zhang, Y. & Zhang, Z. The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. *Cellular & molecular immunology* **17**, 807–821 (2020). PMID: 32612154
26. Kamal, Y., Schmit, S. L., Frost, H. R., *et al.* The tumor microenvironment of colorectal cancer metastases: opportunities in cancer immunotherapy. *Immunotherapy* **12**, 1083–1100 (2020). PMID: 32787587
27. Kamal, Y., Dwan, D., Hoehn, H. J., *et al.* Tumor immune infiltration estimated from gene expression profiles predicts colorectal cancer relapse. *OncoImmunology* **10**, 1862529 (2021). PMID: 33763292
28. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat Methods* **18**, 9–14 (2021). PMID: 33408395
29. Nalisnik, M., Amgad, M., Lee, S., *et al.* Interactive phenotyping of large-scale histology imaging data with HistomicsML. *Sci Rep* **7**, 14588 (2017). PMID: 29109450
30. SAHA, M., GUO, X. & SHARMA, A. TilGAN: GAN for Facilitating Tumor-Infiltrating Lymphocyte Pathology Image Synthesis With Improved Image Classification. *IEEE Access* **9**, 79829–79840 (2021). PMID: 34178560
31. He, B., Bergenstråhle, L., Stenbeck, L., *et al.* Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering* 1–8 (2020) doi:10.1038/s41551-020-0578-x. PMID: 32572199
32. Rao, A., Barkley, D., França, G. S., *et al.* Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021). PMID: 34381231
33. Liu, H., Zhao, Y., Yang, F., *et al.* Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer with Deep Learning. *BME Frontiers* **2022**, (2022).
34. Monjo, T., Koido, M., Nagasawa, S., *et al.* Efficient prediction of a spatial transcriptomics profile better characterizes breast cancer tissue sections without costly experimentation. *Scientific reports* **12**, 1–12 (2022).
35. Li, Y., Stanojevic, S. & Garmire, L. X. Emerging Artificial Intelligence Applications in Spatial Transcriptomics Analysis. *arXiv:2203.09664 [cs, q-bio]* (2022).
36. Uttam, S., Stern, A. M., Sevinsky, C. J., *et al.* Spatial domain analysis predicts risk of colorectal cancer recurrence and infers associated tumor microenvironment networks. *Nature Communications* **11**, 3515 (2020). PMID: 32665557
37. Reinert, T., Henriksen, T. V., Christensen, E., *et al.* Analysis of Plasma Cell-Free DNA by Ultradeep Sequencing in Patients With Stages I to III Colorectal Cancer. *JAMA Oncology* **5**, 1124–1131 (2019). PMID: 31070691
38. Levy, J. J., Bobak, C. A., Nasir-Moin, M., *et al.* Mixed Effects Machine Learning Models for Colon Cancer Metastasis Prediction using Spatially Localized Immuno-Oncology Markers. *Pac Symp Biocomput* **27**, 175–186 (2022). PMID: 34890147
39. Bürkner, P.-C. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* **80**, 1–28 (2017).

40. Carpenter, B., Gelman, A., Hoffman, M. D., *et al. Stan: A Probabilistic Programming Language*. *Grantee Submission* vol. 76 1–32 (2017).
41. Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., *et al.* Indices of Effect Existence and Significance in the Bayesian Framework. *Front. Psychol.* **10**, 2767 (2019). PMID: 31920819
42. Searle, S. R., Speed, F. M. & Milliken, G. A. Population marginal means in the linear model: an alternative to least squares means. *The American Statistician* **34**, 216–221 (1980).
43. Sigrist, F. Latent Gaussian Model Boosting. *arXiv:2105.08966 [cs, stat]* (2021).
44. Tan, Y. V. & Roy, J. Bayesian additive regression trees and the General BART model. *Statistics in Medicine* **38**, 5048–5069 (2019). PMID: 31460678
45. Levy, J. J. & O'Malley, A. J. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. *BMC Med Res Methodol* **20**, 171 (2020). PMID: 32600277
46. Carvalho, C. M., Polson, N. G. & Scott, J. G. Handling Sparsity via the Horseshoe. in *Artificial Intelligence and Statistics* 73–80 (PMLR, 2009).
47. Bartonicek, A., Wickham, S. R., Pat, N., *et al.* The value of Bayesian predictive projection for variable selection: an example of selecting lifestyle predictors of young adult well-being. *BMC Public Health* **21**, 695 (2021). PMID: 33836714
48. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008). PMID: 19114008
49. Gelman, A. & Tuerlinckx, F. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* **15**, 373–390 (2000).
50. Doreian, P. Network autocorrelation models: Problems and prospects. *Spatial statistics: Past, present, future* 369–89 (1989). PMID: 21909184
51. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat Methods* 1–13 (2022) doi:10.1038/s41592-022-01409-2.
52. Hu, J., Li, X., Coleman, K., *et al.* SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat Methods* **18**, 1342–1351 (2021). PMID: 34711970
53. Zhao, E., Stone, M. R., Ren, X., *et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol* **39**, 1375–1384 (2021). PMID: 34083791
54. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods* **17**, 193–200 (2020). PMID: 31988518
55. Zhu, J., Sun, S. & Zhou, X. SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology* **22**, 184 (2021). PMID: 34154649
56. Song, Q. & Su, J. DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Briefings in Bioinformatics* (2021) doi:10.1093/bib/bbaa414. PMID: 33480403
57. Dong, R. & Yuan, G.-C. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biology* **22**, 145 (2021). PMID: 33971932
58. Danaher, P., Kim, Y., Nelson, B., *et al.* Advances in mixed cell deconvolution enable quantification of cell types in spatial transcriptomic data. *Nat Commun* **13**, 385 (2022). PMID: 35046414
59. Cang, Z. & Nie, Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature Communications* **11**, 2084 (2020). PMID: 32350282
60. Armingol, E., Officer, A., Harismendy, O., *et al.* Deciphering cell–cell interactions and communication from gene expression. *Nat Rev Genet* **22**, 71–88 (2021). PMID: 33168968
61. Reimand, J., Isserlin, R., Voisin, V., *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc* **14**, 482–517 (2019). PMID: 30664679
62. Frost, H. R. Tissue-adjusted pathway analysis of cancer (TPAC). *bioRxiv* (2022).
63. Frost, H. R. Analyzing cancer gene expression data through the lens of normal tissue-specificity. *PLoS Computational Biology* **17**, e1009085 (2021). PMID: 34143767

64. Frost, H. R. Computation and application of tissue-specific gene set weights. *Bioinformatics* **34**, 2957–2964 (2018). PMID: 29659714
65. Titus, A. J., Gallimore, R. M., Salas, L. A., *et al.* Cell-type deconvolution from DNA methylation: a review of recent applications. *Hum. Mol. Genet.* **26**, R216–R224 (2017). PMID: 28977446
66. Salas, L. A., Koestler, D. C., Butler, R. A., *et al.* An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biol* **19**, (2018). PMID: 29843789
67. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**, 484–492 (2012). PMID: 22641018
68. Ili, C., Buchegger, K., Demond, H., *et al.* Landscape of Genome-Wide DNA Methylation of Colorectal Cancer Metastasis. *Cancers (Basel)* **12**, E2710 (2020). PMID: 32971738
69. Pidsley, R., Zotenko, E., Peters, T. J., *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology* **17**, 208 (2016). PMID: 27717381
70. Zhou, W., Triche, T. J., Jr, Laird, P. W., *et al.* SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Research* **46**, e123 (2018). PMID: 30085201
71. Salas, L. A., Zhang, Z., Koestler, D. C., *et al.* Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. *Nat Commun* **13**, 761 (2022). PMID: 35140201
72. Salas, L. A., Wiencke, J. K., Koestler, D. C., *et al.* Tracing human stem cell lineage during development using DNA methylation. *Genome Res* **28**, 1285–1295 (2018). PMID: 30072366
73. Du, P., Zhang, X., Huang, C.-C., *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010). PMID: 21118553
74. Zheng, S. C., Breeze, C. E., Beck, S., *et al.* Identification of differentially methylated cell-types in Epigenome-Wide Association Studies. *Nat Methods* **15**, 1059–1066 (2018). PMID: 30504870
75. Rahmani, E., Schweiger, R., Rhead, B., *et al.* Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nat Commun* **10**, 3417 (2019). PMID: 31366909
76. Levy, J. J., Chen, Y., Azizgolshani, N., *et al.* MethylSPWNet and MethylCapsNet: Biologically Motivated Organization of DNAm Neural Networks, Inspired by Capsule Networks. *npj Syst Biol Appl* **7**, 1–16 (2021). PMID: 34417465
77. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32**, 286–288 (2016). PMID: 26424855
78. Geeleher, P., Hartnett, L., Egan, L. J., *et al.* Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics* **29**, 1851–1857 (2013). PMID: 23732277
79. Graw, S., Henn, R., Thompson, J. A., *et al.* pwrEWAS: a user-friendly tool for comprehensive power estimation for epigenome wide association studies (EWAS). *BMC Bioinformatics* **20**, 218 (2019). PMID: 31035919
80. Zhu, J., Park, T., Isola, P., *et al.* Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. in *2017 IEEE International Conference on Computer Vision (ICCV)* 2242–2251 (2017). doi:10.1109/ICCV.2017.244.
81. Zhang, Y., de Haan, K., Rivenson, Y., *et al.* Digital synthesis of histological stains using micro-structured and multiplexed virtual staining of label-free tissue. *Light: Science & Applications* **9**, 78 (2020). PMID: 32411363
82. Rivenson, Y., de Haan, K., Wallace, W. D., *et al.* Emerging Advances to Transform Histopathology Using Virtual Staining. *BME Frontiers* **2020**, (2020).
83. Levy, J., Liu, X., Marotti, J. D., *et al.* Uncovering Additional Predictors of Urothelial Carcinoma from Voided Urothelial Cell Clusters Through a Deep Learning Based Image Preprocessing Technique. 2022.04.30.490136 (2022) doi:10.1101/2022.04.30.490136.

84. Levy, J. J. & Vaickus, L. J. Artificial Intelligence in Anatomic Pathology. *Advances in Molecular Pathology* **4**, 145–171 (2021).
85. Jaume, G., Pati, P., Bozorgtabar, B., *et al.* Quantifying explainers of graph neural networks in computational pathology. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8106–8116 (2021).
86. Ciga, O., Xu, T. & Martel, A. L. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications* **7**, 100198 (2022).
87. Lu, M. Y., Chen, R. J., Wang, J., *et al.* Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arXiv preprint arXiv:1910.10825* (2019).
88. Zheng, H., Momeni, A., Cedoz, P.-L., *et al.* Whole slide images reflect DNA methylation patterns of human tumors. *npj Genomic Medicine* **5**, 1–10 (2020). PMID: 32194984
89. Pope, P. E., Kolouri, S., Rostami, M., *et al.* Explainability Methods for Graph Convolutional Neural Networks. in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10764–10773 (2019). doi:10.1109/CVPR.2019.01103.
90. Ahmedt-Aristizabal, D., Armin, M. A., Denman, S., *et al.* A Survey on Graph-Based Deep Learning for Computational Histopathology. *arXiv:2107.00272 [cs, q-bio]* (2021).
91. Jaume, G., Pati, P., Anklin, V., *et al.* Histocartography: A toolkit for graph analytics in digital pathology. in *MICCAI Workshop on Computational Pathology* 117–128 (PMLR, 2021).
92. OpenSeadragon. http://openseadragon.github.io/.
93. Bergenstråhle, J., Larsson, L. & Lundeberg, J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC genomics* **21**, 1–7 (2020). PMID: 32664861
94. Andersson, A., Bergenstråhle, J., Asp, M., *et al.* Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun Biol* **3**, 1–8 (2020). PMID: 33037292
95. Yuan, Y. & Bar-Joseph, Z. GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biology* **21**, 300 (2020). PMID: 33303016
96. Lopez, R., Nazaret, A., Langevin, M., *et al.* A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv preprint arXiv:1905.02269* (2019).