Inaugural NCI Informatics Technology for Cancer Research (ITCR) Trainee Symposium (Virtual)

"Today's Computations - Tomorrow's Cures"

MAY 15-17, 2024

Host: Emory University

Atlanta GA, United States

Xinyi Yang. Learning without Real Data Annotations to Detect Hepatic Lesions in PET Images
Benjamin Parker. Intra- and peri-tumoral radiomic features are predictive of pathologic response to multiple neoadjuvant therapy regimen in rectal cancers via pre-treatment MRI
Harrison Yee. Utilizing Deep Learning and Channel-Wise Data Fusion for End-to-end Organelle-based Breast Cancer Cell Classification

Clinical and population research 17-May

Yingjie Qiu. Transparent and Efficient Adaptive Designs for Clinical Trials
Yi Lian. A Flexible and Adaptive Framework for High-Dimensional Fairness-Aware Integration of Data from Multiple Sites
Yiming Zhang. Disparities in the Documentation of Social Determinants of Health ICD-10 Z-Codes for Patients Diagnosed with Cancer: An Epic Cosmos Study
Julia Herriott. Pilot Testing the User Interface for a Novel Algorithm to Process Electronic Adherence Monitoring Device Data
Ramon Ortiz. The TPS2TOPAS interface tool for investigating new radiotherapy devices

NCI ITCR Trainee Symposium Organizing Committee

Welcoming Remarks

Juli Klemm
National Cancer Institute

Suresh S. Ramalingam
Emory University

Keynote Lectures

Jill Mesirov, Ph.D.
University of California San Diego

Keynote lecture: Computational Cancer Genomics: Applications, Methods, and Software

Day 1, May 15. 1:10PM

Nastaran Zahir, Ph.D.
National Cancer Institute

Keynote lecture: NCI Opportunities for Early Career Cancer Researchers

Day 2, May 16. 1:00PM

Anant Madabhushi, Ph.D.
Emory University

Keynote lecture: My ongoing career journey in AI and Medicine

Day 3, May 17. 1:00PM

Omics technologies across the cancer research continuum
15-May

Session Chairs:
Ramon Ortiz, Ph.D., University of California San Francisco
Zeynep Kosaloglu-Yalcin, Ph.D., La Jolla Institute for Immunology

AmpliconSuite: Analyzing focal amplifications in cancer genomes
Presenter: Jens Luebeck
Presenter's email: jluebeck@ucsd.edu
Institute: University of California San Diego
Principle Investigator: Vineet Bafna

Focal amplifications in cancer genomes, particularly extrachromosomal DNA (ecDNA) amplifications, are pivotal in cancer progression, enabling high amplification of oncogenes. Distinguishing these events with whole-genome sequencing (WGS) data is challenging due to their complex profiles of copy number (CN) and structural variation (SV). We present AmpliconSuite, a collection of tools enabling robust identification of focal amplifications from WGS data. At the core are the AmpliconArchitect (AA) and AmpliconClassifier (AC) methods, which detect and analyze SVs and CNs using WGS data to produce robust predictions of focal amplification types, including ecDNA and breakage-fusion-bridge (BFB) cycles. We combined these tools into a single, reproducible workflow, AmpliconSuite-pipeline, which is available through Nextflow, GenePattern and Bioconda. AmpliconSuite-pipeline also incorporates other upstream tools into the workflow to standardize inputs and improve filtering of inputs. It introduces our latest classification methods, such as the ecContext method within AC for classifying types of ecDNA based on their mechanisms of formation (chromothripsis, excision, etc.).
To foster collaboration, we also introduce a companion website, AmpliconRepository.org. This community-editable platform allows researchers to publicly share calls generated by AmpliconSuite. Notably, AmpliconRepository.org provides ecDNA predictions on 2,525 tumor samples from TCGA, PCAWG, and CCLE. The ongoing goal of this repository is to become the largest resource for focal amplifications in cancer, driven primarily by ecDNA but also including other mechanisms like BFB.
AmpliconSuite makes identification of focal amplifications reproducible and simple, and empowers users to share analyses publicly, representing a valuable resource to investigate the mechanisms of oncogene amplification.

FAST-scDECO: a flexible and adaptive scalable tool for single cell differential coexpression analysis
Presenter: Anderson Bussing
Presenter's email: abussing@email.sc.edu
Institute: University of South Carolina
Principle Investigator: Yen Yi Ho

Within a cell, genetic interactions are highly dynamic and tightly regulated in response to internal cellular signals and external stimuli. Evidence of these dynamic interactions can be seen in single-cell gene expression (scRNA-seq) data by examining dynamic coexpression changes. Existing approaches for studying these changes have utilized Bayesian frameworks, with parameter estimates and standard errors being obtained via Monte Carlo sampling. However, these methods are too slow to handle gene-gene interactions on a genome-wide scale. In this talk, we propose a frequentist perspective to speed up the parameter estimation process. Our proposed model can accommodate the zero-inflation and overdispersion commonly observed in scRNA-seq data, and it allows for covariate-dependent marginal parameters and dependence structures. We conducted simulations to compare our proposed approach to existing methods and evaluated the performance of our approach in terms of the computation time, coverage, power, and robustness. We also applied it to triple-negative breast cancer scRNA-seq data, where we performed whole-genome search to identify spatially significant gene pairs.

DSCC: Disease Subtyping using Consensus Network and Multi-omics Data Integration
Presenter: Ha Nguyen
Presenter's email: hvn0006@auburn.edu
Institute: Auburn University

Principle Investigator: Tin Nguyen

Cancer is a complex disease driven by numerous biological processes activating on multiple levels. Various genome-wide profiling techniques have been developed to capture the dynamics of these processes at the genomics, transcriptomics, epigenomics, and proteomics levels. Integrative analysis of data from these sources offers a comprehensive view that reveals connections unattainable through single-omic observations. In this study, we introduce DSCC (Disease Subtyping using Community detection from Consensus network), a novel approach aimed at discovering disease subtypes from multi-omics data. DSCC leverages pathway knowledge and exploits local patient relationships within each data type to construct a consensus network based on patient connectivities. Through an extensive analysis utilizing real multi-omics data encompassing over 15,000 patients across 33 cancers sourced from The Cancer Genome Atlas, METABRIC, and Gene Expression Omnibus, our findings demonstrate the robustness of DSCC against noise. Moreover, it achieves remarkable performance in identifying both known patient classes and novel subtypes, characterized by significant differences in survival profiles.

A Deep Learning Foundation Model for Predicting Responses to Genetic and Chemical Perturbations in Single Cancer Cells

Presenter: Farzan Taj
Presenter's email: farzan.taj@mail.utoronto.ca
Institute: University of Toronto
Principle Investigator: Lincoln Stein

In cancer treatment, interpatient variability presents a significant challenge, as patients with ostensibly similar profiles often exhibit divergent responses to identical therapies. This variability is primarily attributed to genetic and molecular differences among individuals and their tumors. Recognizing the importance of understanding the impact of these differences on treatment outcomes, research endeavors have increasingly focused on generating large-scale drug screening and pharmacogenomic data and integrating them into predictive computational models for drug response. Our recently published work, the Multi-Modal Drug Response Predictor (MMDRP), epitomizes this effort by identifying and alleviating prevalent limitations in generalizability, data processing, and representation in drug response prediction. The MMDRP model leverages new deep-learning methods and data types to enhance the prediction accuracy for novel drug and cell line combinations.
Building upon this foundation, our current research extends MMDRP's methodologies to single-cell resolution, introducing the Single-Cell Multi-Modal Perturbation Response Predictor (scMMPRP). This framework employs a transformer-based architecture to model the effects of not only chemical but also genetic perturbations at the single-cell level. By training on various large-scale single-cell perturbation datasets, scMMPRP aims to predict cellular outcomes from a variety of perturbation types, enhancing our understanding of cellular response dynamics. The model also has the capacity to reverse-engineer perturbation effects to achieve desired gene expression profiles. Specifically, in the context of oncology, scMMPRP could facilitate the discovery of therapeutic interventions that transform malignant cellular states into benign ones, thereby advancing personalized cancer treatment and contributing to precision medicine's evolution.

Data driven refinement of gene signatures for enrichment analysis and cell state characterization

Presenter: Alexander Wenzel
Presenter's email: atwenzel@ucsd.edu
Institute: University of California San Diego
Principle Investigator: Jill Mesirov

Gene Set Enrichment Analysis (GSEA) is a standard method for identifying pathway enrichment in gene expression data by testing whether a set of genes whose expression would indicate the activity of a specific process or phenotype are coordinately up- or downregulated more than would be expected by chance. As GSEA relies on high quality gene sets with coordinately regulated member genes, we maintain the Molecular Signatures Database (MSigDB) which contains 9 collections of curated gene sets representing different biological pathways and processes. Over time, we have observed that some of the MSigDB gene sets, especially those that are manually curated or defined in a very specific biological context, may not provide a sensitive and specific enough co-regulation signature. In response, we have created a data-driven, matrix-factorization-based refinement method to build more sensitive and specific gene sets. This method incorporates large-scale datasets from multiple sources such as the Cancer Dependency Map and can use existing signatures to generate one or more sensitive and context-specific gene sets. We will present the initial results of this refinement method and our ongoing work which will yield a new collection of refined gene sets that will be made freely available in MSigDB for use with GSEA and many other applications.

Integrating multi-omics analyses into agent-based models made easy: the Bioinformatics Walkthrough for PhysiCell

Presenter: Daniel Bergman
Presenter's email: dbergma5@jh.edu
Institute: Johns Hopkins University
Principle Investigator: Elana Fertig

Spatial multi-omics datasets provide unprecedented characterization of tumors. Current methods, however, only capture the tumor microenvironment (TME) at a single timepoint. To overcome this limitation, mechanistic mathematical models informed by these data can be used to predict TME evolution. Agent-based models are particularly well-suited as they represent cells as digital agents, each obeying its own set of rules as it interacts with other agents and the environment. Key to ABM success is accurately defining and calibrating agents to reflect reality, including initial tissue organization and cellular behaviors. Bioinformatics analysis of multi-omics data yields the quantitative insights into cell identity, function, and location necessary for model construction. While researchers have successfully realized this promise of bioinformatics, the bespoke nature of prior work and the increasing availability of sequencing data underscores a need to create robust and reusable tools to bridge these two fields We have taken a step towards realizing this by building a Bioinformatics Walkthrough, in which a user leverages analyzed multi-omics data to generate initial ABM organization using an intuitive GUI within the PhysiCell framework. Coupled with the well-developed PhysiCell ecosystem and the newly-released rules grammar, this enables ABM creation from omics data within minutes. We showcase its efficacy by generating patient-specific ABMs of pancreatic cancer from spatial transcriptomics datasets, predicting patient-level heterogeneity in TME evolution. This marks a crucial step towards leveraging multi-omics data to specify and initialize cancer ABMs effectively.

ImmunoPheno: A Data-Driven Bioinformatics Platform for the Design and Analysis of Immunophenotyping Experiments

Presenter: Sriya Potluri
Presenter's email: Sriya.Potluri@pennmedicine.upenn.edu
Institute: University of Pennsylvania Perelman School of Medicine
Principle Investigator: Pablo Gonzalez Camara

The tumor microenvironment influences cancer progression, treatment, and metastasis. Immunophenotyping provides valuable information to analyze dynamic interactions occurring in the tumor microenvironment and to identify cells based on antigens present on a cell's surface. As the standard method for phenotypic categorization of immune cell populations in tissue samples, multiplexed antibody-based cytometry techniques measure specific protein expression levels and algorithms use predefined sets of protein markers to cluster cells according to antigenic profile. However, current cytometry techniques are limited by the infeasibility of designing antibody panels that include markers for all cell types and states, especially for rare cell populations present in a sample, and many cell clusters overlap in a tissue only differ slightly in their antigenic profiles. Furthermore, accurately identifying cell populations typically depends on the data's annotation, which is typically a manual, subjective, and laborious process that can hinder the reproducibility and accuracy of the results. To overcome these challenges, we are developing ImmunoPheno, a Python library and online resource that leverages single-cell transcriptomic atlases, to assist and automate the identification and annotation of immune cell populations in cytometry data based on harmonized reference single-cell proteo-transcriptomic data. Built upon our previous work STvEA (Spatial Transcriptomics via Epitope Anchoring), ImmunoPheno enables the detection of subtle cell populations, spatial patterns of transcription, and candidate paracrine interactions in multiplexed antibody-based cytometry data. As an online database and web portal ImmunoPheno advances the design of new cytometry experiments and boosts the phenotypic resolution, accuracy, and reproducibility of multiplexed antibody-based cytometry data analyses.

Improving Tumour Subclonal Reconstruction via Phasing: A Case Study in Hereditary Leiomyomatosis and Renal Cell Cancer

Presenter: Selina Wu
Presenter's email: Selinawu@mednet.ucla.edu
Institute: University of California Los Angeles
Principle Investigator: Paul Boutros

Genetic mutations in tumor genomes are fundamental in subclonal reconstruction (SRC), helping to trace the evolutionary trajectory of cancer subclones and reveal drivers behind tumor initiation and evolution. A notable challenge in SRC is the unambiguous detection of branching phylogenies among subclones, which is important in guiding effective treatment strategies. Addressing this, we introduce a novel tool for single nucleotide variant (SNV) phasing, enabling the detection of branching and linear evolutionary patterns, thereby enhancing tumor

reconstruction accuracy. Leveraging read-level data to verify SNV co-occurrence and phylogeny among subclones, our phasing tool will become increasingly essential for accurate SRC as the adoption of long-read sequencing, which overcomes short-read limitations, expands. Additionally, in single-sample studies, where establishing branching phylogeny is especially challenging, our tool provides vital insights into discerning linear or branching relationships. We apply this tool in the SRC of hereditary leiomyomatosis and renal cell carcinoma (HLRCC) tumors, demonstrating its utility through a case study with high read depth and tumor purity in whole genome sequencing of tumor/normal pairs. Our analysis reveals late and limited subclonal diversification in HLRCC, detecting an average of four subclones per tumor. We also identify key mutational drivers of tumorigenesis, including loss of heterozygosity (LOH) on chromosome 1q, widespread copy number gains on chromosomes 2 and 16, and extensive clonal and subclonal LOH. This analysis highlights our tool's ability to uncover the subclonal architecture of this aggressive cancer and inform treatment strategies addressing the genetic diversity within tumors.

Precision oncology decision-support informatics approaches to match actionable genotypes with targeted therapies

Presenter: Jaime Wehr
Presenter's email: jwehr3@jh.edu
Institute: Johns Hopkins University
Principle Investigator: Valsamo Anagnostou and Taxiarchis Botsis

Jaime Wehr[1,2], Kory Kreimeyer[1,3], Maria Fatteh[1,2], Jonathan Spiker[1,3], Mimi Najjar[1,2], Jessica Tao[1,2], Rena Xian[4], Adrian Dobbs[5], Jenna Canzoniero[1,2], Taxiarchis Botsis[1,3] and Valsamo Anagnostou[1,2]
[1]Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD; [2]The Johns Hopkins Molecular Tumor Board, Johns Hopkins School of Medicine, Baltimore, MD; [3]Division of Quantitative Sciences, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD; [4]Department of Pathology, Johns Hopkins School of Medicine, Baltimore, MD; [5]Johns Hopkins Clinical Research Network, Johns Hopkins School of Medicine, Baltimore, MD

The genomic complexity of cancers has allowed for the delivery of precision oncology where the genotype of a tumor is used to match the patient to targeted therapies. There is an expanding list of genomic alterations linked with FDA-approved therapies, but the question we frequently encounter is how do we match "un-vetted" mutations detected by next-generation sequencing (NGS) with effective treatments. Despite the development of knowledgebases, meta-knowledgebases and computational algorithms for mutation pathogenicity assessment, an end-to-end information technology that matches genotypes with appropriate targeted therapies in the context of clinical trials does not currently exist and represents a major gap in the delivery of precision medicine. To this end, and within the NCI ITCR grant U01CA274631, we have commenced the development of a precision oncology platform that utilizes open-source tools, retrieves data from public repositories via APIs, processes and annotates NGS data together with structured and unstructured free-texts, extracts features including gene names, mutations, and treatments and leverages a common data model with standardized EHR mappings. Furthermore, we have built a user's interface to allow for interaction with the platform, data visualization and submission of user-generated information to the platform's local database. Ultimately actionable mutations are matched with genotype-tailored active clinical trials. Pilot testing of the clinical utility of this approach within the Johns Hopkins Molecular Tumor Board and the Johns Hopkins Community Research Network resulted in identification of patients that may benefit from targeted therapies and highlights the clinical value of our approach in enhancing enrollment in clinical trials.

CancerModels.Org - an open global cancer research platform for patient-derived cancer models.

Presenter: Tushar Mandloi
Presenter's email: tushar@ebi.ac.uk
Institute: EMBL-EBI
Principle Investigator: Zinaida Perova, PhD

CancerModels.Org (www.cancermodels.org) is a research platform that standardises, harmonises and integrates the data associated with Patient-Derived Cancer Models (PDCMs). The portal publishes over 8300 models - covering PDXs, organoids and cell lines - across 13 cancer types, including rare pediatric PDX models and models from minority ethnic backgrounds. This makes CancerModels.Org the largest free-to-consumer and open-access resource of this kind.
In the last year, the platform has been enhanced with new functionality and an updated user interface to cater to a more varied set of use cases. Users can search for models of interest by exploring molecular data summaries for models of specific cancer types, as well as by using the intuitive search and faceted filtering options of the web interface. The data is also accessible via REST API, hence enabling offline analyses. The underpinning data model supports gene expression, gene mutation, copy number alteration, immune-markers, biomarkers, patient treatment,

drug dosing studies and image data. For an improved prioritization of PDCMs we performed knowledge enrichment by linking to external resources, such as publication platforms, cancer-specific annotation tools (COSMIC, CIViC, OncoMX, OpenCRAVAT, ClinGen), and raw data archives (ENA, EGA, GEO, dbGAP). Finally, to streamline model and data submission, we built a Metadata dictionary and a Metadata validation service.
In conclusion, CancerModels.Org aggregates PDCMs from 38 academic and commercial providers, enabling users to search and compare models and associated molecular data. It increases the visibility and reusability of the models and facilitates collaboration across disciplines and geographies.

Flash talks

16-May

Session Chair:
Debolina Chatterjee, Ph.D., Indiana University

cBioPortal for Cancer Genomics

Presenter: Ino de Bruijn
Presenter's email: debruiji@mskcc.org
Institute: Memorial Sloan Kettering Cancer Center
Principle Investigator: Nikolaus Schultz

cBioPortal for Cancer Genomics is an open-source platform for interactive, exploratory analysis of large-scale clinico-genomic data. cBioPortal provides a suite of user-friendly visualizations and analyses, including OncoPrints, mutation "lollipop" plots, variant interpretation, group comparison, survival analysis, expression correlation analysis, alteration enrichment analysis, cohort and patient-level visualization.
The public site (https://www.cbioportal.org) is accessed by >35,000 unique visitors each month and hosts data from >390 studies spanning individual labs and large consortia. All data is available in the cBioPortal Datahub: https://github.com/cBioPortal/datahub; in 2023 we added 35 studies (~16,000 samples). In addition, >86 instances of cBioPortal are installed at academic institutions and companies worldwide.
We have improved support for multimodal datasets incorporating derived data elements, including cell type counts and fractions per sample from imaging or single-cell data. The MSK-SPECTRUM ovarian cancer study has samples profiled with bulk sequencing, scRNASeq, H&E and mpIF imaging. Through integrations with CELLxGENE (single cell data) and Minerva (imaging), users can explore these data modalities in detail in isolation and query them jointly in cBioPortal. We also continue to support and improve integrations with several ITCR-funded projects including MusicaTK, Galaxy, NDEx, CIViC, IGV, Next Generation Clustered HeatMap, Cancer Digital Slide Archive, and Bioconductor.
cBioPortal is open source (https://github.com/cBioPortal/). Development is a collaborative effort among groups at Memorial Sloan Kettering Cancer Center, Dana-Farber Cancer Institute, Children's Hospital of Philadelphia, Princess Margaret Cancer Centre, Caris Life Sciences, Bilkent University and The Hyve.

pVACsplice: a bioinformatics tool for detecting and prioritizing tumor-specific splicing-derived neoantigen

Presenter: My Hoang
Presenter's email: hmy@wustl.edu
Institute: Washington University in Saint Louis
Principle Investigator: Malachi Griffith

Alternative splicing-derived neoantigens is a rich source of putative tumor-specific targets for immunotherapy. Tumor cells exhibit heightened mis-splicing events compared to normal tissues, generating diverse transcript isoforms that encode novel peptides. These peptides, especially ones derived from frameshifts, are highly dissimilar from self-antigens, hence presenting an opportunity for enhanced immune recognition.
Though neoantigens arising from somatic single-point mutations in coding regions have been widely targeted by cancer therapies, other neoantigen sources, including alternative splicing neoantigens haven't been as extensively exploited. Here, we develop pVACsplice, a tool that predicts and prioritizes cis-splicing associated neoantigen candidates. pVACsplice takes alternative transcripts as input, translates them into altered peptides, then constructs neoantigens of user-defined sizes. It then estimates binding affinities of neoepitopes with user-input MHC alleles, and prioritizes candidates based on various criteria (binding affinity, solubility, transcript quality, and more). We then apply pVACsplice to examine the splicing neoantigen landscape of a Small Cell Lung Cancer (SCLC) cohort. We find numerous neojunctions and neoantigen candidates associated with genes frequently mutated in this malignancy.

Elucidating immune-related gene transcriptional programs via factorization of large-scale RNA-profiles

Presenter: Shan He
Presenter's email: she4@mdanderson.org
Institute: MD Anderson Cancer Center
Principle Investigator: Ken Chen

Immune checkpoint blockade (ICB) and adoptive cell therapy (ACT) have revolutionized cancer immunotherapy, yet their full potential remains limited due to immune-related adverse events or treatment resistance, especially in solid tumors. To overcome these hurdles, understanding the molecular mechanisms behind treatment resistance and response is essential. However, the current lack of immunologically relevant gene transcriptional programs (GTPs) hampers data-driven immunological discovery. Constructing immunity-specific knowledgebases with rigorously curated gene sets (irGSs) and rich immunological language can unlock new insights, enabling interpretation of high-throughput immune microenvironment profiling studies and fostering personalized, effective cancer treatments.
We collected 83 BulkRNAseq datasets from the ImmuneSigDB. These datasets contain samples challenged with infections, cytokines or immunological perturbations of different kinds and magnitude, possessing yet-to-be discovered immune functions that lie beneath the transcriptomic profiles. Using non-negative matrix factorization (NMF), we identified gene sets with coordinated expression. We performed extensively validation using different data modalities under different immunological contexts.
We presented 19 lymphoid-data derived and 9 myeloid-data derived gene sets (irGSs), encompassing diverse immune functions. We revealed six irGSs-defined pan-cancer microenvironment subtypes with significantly distinct survival patterns. irGSs well predict ICB response in melanoma, liver, and lung cancer. Lastly, our gene sets effectively delineate tumor-immune boundaries in breast cancer spatial transcriptomics data.
By studying gene set activities, immunologists can gain profound insights into cancer survival drivers, unravel the intricacies of ICB treatment mechanisms, and potentially conquer therapeutic resistance. These translational utilities herald a transformative era in cancer immunotherapy and open new frontiers in the fight against cancer.

A meta-analysis of cancer neoantigens curated from the literature

Presenter: Zeynep Kosaloglu-Yalcin
Presenter's email: zeynep@lji.org
Institute: La Jolla Institute for Immunology
Principle Investigator: Bjoern Peters

The Cancer Epitope Database and Analysis Resource (CEDAR) is a newly developed database for cancer epitope data curated from peer-reviewed publications, including epitope-specific T cell, antibody, and MHC ligand assays. Neoantigens are epitopes that arise from somatic mutations in cancer cells and are highly tumor-specific. Neoantigens are of interest as emerging targets for personalized cancer immunotherapies and as predictors for survival prognosis and immune checkpoint blockade responses. Given the importance of neoantigens, the curation of this category of epitopes was prioritized for CEDAR. Around three thousand neoantigens with positive assay outcomes have been curated in CEDAR. Here, we present a meta-analysis of this set of curated neoantigens to shed light on their specific features.

CoRAL accurately resolves extrachromosomal DNA genome structures with long-read sequencing

Presenter: Kaiyuan Zhu
Presenter's email: kaiyuan-zhu@ucsd.edu
Institute: University of California San Diego
Principle Investigator: Vineet Bafna

Extrachromosomal DNA (ecDNA) is a central mechanism for focal oncogene amplification in cancer, occurring in approximately 15% of early stage cancers and 30% of late-stage cancers. EcDNAs drive tumor formation, evolution, and drug resistance by dynamically modulating oncogene copy-number and rewiring gene-regulatory networks. Elucidating the genomic architecture of ecDNA amplifications is critical for understanding tumor pathology and developing more effective therapies.
Paired-end short-read (Illumina) sequencing and mapping have been utilized to represent ecDNA amplifications using a breakpoint graph, where the inferred architecture of ecDNA is encoded as a cycle in the graph. Traversals of breakpoint graph have been used to successfully predict ecDNA presence in cancer samples. However, short-read technologies are intrinsically limited in the identification of breakpoints, phasing together of complex rearrangements and internal duplications, and deconvolution of cell-to-cell heterogeneity of ecDNA structures. Long-read technologies, such as from Oxford Nanopore Technologies, have the potential to improve inference as the longer reads

are better at mapping structural variants and are more likely to span rearranged or duplicated regions.
Here, we propose CoRAL (Complete Rconstruction of Amplifications with Long reads), for reconstructing ecDNA architectures using long-read data. CoRAL reconstructs likely cyclic architectures using quadratic programming that simultaneously optimizes parsimony of reconstruction, explained copy number, and consistency of long-read mapping. CoRAL substantially improves reconstructions in extensive simulations and 9 datasets from previously-characterized cell-lines as compared to previous short-read-based tools.
As long-read usage becomes wide-spread, we anticipate that CoRAL will be a valuable tool for profiling the landscape and evolution of focal amplifications in tumors.

EMulSI-Phy: Efficient Multi-Sample Inference of Cancer Phylogeny

Presenter: Helena Winata
Presenter's email: hwinata@g.ucla.edu
Institute: University of California Los Angeles
Principle Investigator: Paul Boutros

Inferring tumor evolution from DNA sequencing data is becoming part of routine analysis in cancer research. This reconstruction is crucial in understanding key events that drive cancer progression and patterns of mutation co-occurrence within clones. Sequencing multiple tumor samples from a patient across different spatial locations or time points offers detailed insights into tumor evolution. Yet, existing methods fail to routinely and accurately reconstruct phylogenies for datasets with as few as ten subclones.
To address these challenges, EMulSI-Phy leverages fundamental principles of cancer biology and evolution to enhance phylogenetic inference in a computationally efficient way. We will utilize read-level data to identify branching or linear evolution between SNV pairs, further constraining the set of plausible solutions. The utility of this strategy will grow as long-read sequencing becomes more prevalent. Benchmarking against Pairtree, a leading multi-sample reconstruction method, EMulSI-Phy shows an 18X decrease in runtime across 576 simulated datasets. This trend is replicated in real-world application, enabling the reconstruction of up to 88 distinct clones from 36 tumor samples in patients with metastatic breast cancer. EMulSI-Phy's computational efficiency and accuracy represent significant progress, providing a robust, scalable tool to unravel complex tumor evolution dynamics.

Generating counterfactual explanations of tumor spatial proteomes to discover therapeutic strategies for enhancing immune infiltration

Presenter: Zitong Jerry Wang
Presenter's email: zwang2@caltech.edu
Institute: California Institute of Technology
Principle Investigator: Matt Thomson

Immunotherapies can halt or slow down cancer progression by activating either endogenous or engineered T cells to detect and kill cancer cells. For immunotherapies to be effective, T cells must be able to infiltrate the tumor microenvironment. However, many solid tumors resist T-cell infiltration, challenging the efficacy of current therapies. Here, we introduce Morpheus, an integrated deep learning framework that takes large scale spatial omics profiles of patient tumors, and combines a formulation of T-cell infiltration prediction as a self-supervised machine learning problem with a counterfactual optimization strategy to generate minimal tumor perturbations predicted to boost T-cell infiltration. We applied our framework to 368 metastatic melanoma and colorectal cancer (with liver metastases) samples assayed using 40-plex imaging mass cytometry, discovering cohort-dependent, combinatorial perturbations, involving CXCL9, CXCL10, CCL22 and CCL18 for melanoma and CXCR4, PD-1, PD-L1 and CYR61 for colorectal cancer, predicted to support T-cell infiltration across large patient cohorts. Our work presents a paradigm for counterfactual-based prediction and design of cancer therapeutics using spatial omics data.

General informatics resources and platforms
16-May

Session Chairs:
Debolina Chatterjee, Ph.D., Indiana University
Yi Lian, Ph.D., University of Pennsylvania

Expanding Gearbox: LLM and ontology based opportunities to enrich clinical trial mapping

Presenter: Lauren Chan
Presenter's email: laurenchan@uchicago.edu
Institute: University of Chicago
Principle Investigator: Sam Volchenboum

For pediatric oncology patients with refractory or relapsed disease, it is critical to select treatment quickly, including identifying potential clinical trials. Parsing and interpreting numerous lengthy and complex criteria can be challenging for both families and clinicians seeking suitable trials. Launched at UChicago in 2023, the Genomic Eligibility AlgoRithm for Better Outcomes (GEARBOx) is a self-service clinical trial matching tool for efficient identification of trial eligibility. GEARBOx currently includes trials for pediatric acute myeloid leukemia (AML) and uses natural language processing (NLP) plus human curation to ingest trial criteria and compose survey content and trial mapping logic. Areas for improvement include; 1) workflow automation, 2) inclusion of additional tumor types, and 3) data standards for downstream analysis of tool use and suitability. We are currently trialing large language models (LLMs) to extract structured clinical trial data from free-text inclusion and exclusion criteria. We hope that application of LLMs will reduce time needed to parse trial criteria, incorporate new trials into GEARBOx, and facilitate patient trial matching. In parallel, we are also utilizing biomedical ontologies and their standardized language content regarding diseases, phenotypes, genes and otherwise to map our survey questions and resulting data. Biomedical ontology alignment will aid LLM grounding to reduce "hallucinations" and improve output data quality. Additionally, ontology coordination of surveys can create downstream computational analysis opportunities using knowledge graphs or other methods. Through enhanced criteria mapping, we can incorporate more tumor types and continue optimizing patient trial matching using GEARBOx to support subsequent patient care outcomes.

Extraction of Chemotherapy Treatment Timelines from EHR Notes

Presenter: Jiarui Yao
Presenter's email: Jiarui.yao@childrens.harvard.edu
Institute: Boston Children's Hospital, Harvard Medical School
Principle Investigator: Guergana Savova

Detailed information about tumor biology and effects of anticancer therapy is key to understanding the outcomes of patients with cancer. To understand relationships between treatments and outcomes, researchers and clinicians must identify which treatments were given to patients and the contexts in which those treatments were given. The complex nature of cancer treatments makes this a challenging task, as treatments often follow protocols involving multiple drugs given according to detailed schedules lasting weeks, months, or even years. Understanding when treatments involving multiple agents correspond to a published protocol can be difficult, particularly when many well-studied protocols have multiple similar variants. As this information is generally not available in structured data, extraction from EHR notes is necessary. Our goal is to use natural language processing approach to build tools for extracting treatment data from cancer clinical notes, thus providing tools capable of providing greater context and understanding of treatments provided to patients with cancer.
We developed – to our knowledge – the first end-to-end pipeline to extract chemotherapy events from EHR clinical narrative and to assemble a patient timeline. Our labeled corpus consists of instance- and patient-level gold annotations of de-identified notes. Instance-level labels are gold annotations for EVENTs (in our case chemotherapy treatments), TIMEXs (temporal expressions) and pairwise temporal relations between an EVENT and a TIMEX. These annotations are the instance level evidence that supports the patient-level timelines.
Preliminary results across several cancer suggests a range of performance characteristics across cancers, with F1 0.62, 0.86, 0.62, and 0.32 for Colorectal, Breast, Ovarian and Melanoma cancer data respectively.

EnGens: a computational framework for generation and analysis of representative protein conformational ensembles

Presenter: Anja Conev
Presenter's email: ac121@rice.edu
Institute: Rice University
Principle Investigator: Lydia Kavraki

Proteins are dynamic macromolecules that perform vital functions in cells. A protein structure determines its function, but this structure is not static, as proteins change their conformation to achieve various functions. Understanding the conformational landscapes of proteins is essential to understand their mechanism of action. Sets of carefully chosen conformations can summarize such complex landscapes and provide better insights into protein function than single conformations. We refer to these sets as representative conformational ensembles. Recent

advances in computational methods have led to an increase in number of available structural datasets spanning conformational landscapes. However, extracting representative conformational ensembles from such datasets is not an easy task and many methods have been developed to tackle it. Our new approach, EnGens (short for ensemble generation), collects these methods into a unified framework for generating and analyzing protein conformational ensembles. In this work we: (1) provide an overview of existing methods and tools for protein structural ensemble generation and analysis; (2) unify existing approaches in an open-source Python package, and a portable Docker image, providing interactive visualizations within a Jupyter Notebook pipeline; (3) test our pipeline on a few canonical examples found in the literature. Representative ensembles produced by EnGens can be used for many downstream tasks such as protein-ligand ensemble docking, Markov state modeling of protein dynamics and analysis of the effect of single-point mutations.

APE-Gen2.0: Expanding Rapid Class I pMHC Modeling to Post-Translational Modifications and Non-canonical Peptide Geometries

Presenter: Romanos Fasoulis
Presenter's email: rf27@rice.edu
Institute: Rice University
Principle Investigator: Lydia E. Kavraki

The recognition of peptides bound to class I major histocompatibility complex (MHC-I) receptors by T-cell receptors (TCRs) is a determinant of triggering the adaptive immune response. While the exact molecular features that drive the TCR recognition are still unknown, studies have suggested that the geometry of the joint peptide–MHC (pMHC) structure plays an important role. As such, there is a definite need for methods and tools that accurately predict the structure of the peptide bound to the MHC-I receptor. In the past few years, many pMHC structural modeling tools have emerged that provide high-quality modeled structures in the general case. However, there are numerous instances of non-canonical cases in the immunopeptidome that the majority of pMHC modeling tools do not attend to, most notably, peptides that exhibit non-standard amino acids and post-translational modifications (PTMs) or peptides that assume non-canonical geometries in the MHC binding cleft. Such chemical and structural properties have been shown to be present in neoantigens; therefore, accurate structural modeling of these instances can be vital for cancer immunotherapy. To this end, we have developed APE-Gen2.0, a tool that improves upon its predecessor and other pMHC modeling tools, both in terms of modeling accuracy and the available modeling range of non-canonical peptide cases. Some of the improvements include (i) the ability to model peptides that have different types of PTMs such as phosphorylation, nitration, and citrullination; (ii) a new and improved anchor identification routine in order to identify and model peptides that exhibit a non-canonical anchor conformation; and (iii) a web server that provides a platform for easy and accessible pMHC modeling. We further show that structures predicted by APE-Gen2.0 can be used to assess the effects that PTMs have in binding affinity in a more accurate manner than just using solely the sequence of the peptide. APE-Gen2.0 is freely available at https://apegen.kavrakilab.org.

Evaluating Usability and Feasibility of Implementing Cancer Epidemiology Maps for SEER Registry End Users

Presenter: Erin Wissler Gerdes
Presenter's email: erin-wisslergerdes@uiowa.edu
Institute: University of Iowa
Principle Investigator: Jacob Oleson, Sarah Nash, and Mary Charlton

Reporting cancer incidence and mortality in smaller geographic areas is a challenge due to low case counts and the associated potential for identification of cases. In collaboration with the Iowa Cancer Registry, Kentucky Cancer Registry and New Mexico Tumor Registry, the University of Iowa developed an innovative mapping tool that uses a Bayesian hierarchical model to estimate age-adjusted cancer rates and cancer risk in smaller geographic regions, including Counties and ZIP Code Tabulation Areas. The current iteration of the tool includes mortality, incidence, and late-stage incidence data for eight common cancers. In order to assess feasibility and usability of the tool, we conducted key informant interviews with staff at each partner registry. These interviews provided initial feedback around map interpretability and identified potential end users across Iowa, New Mexico, and Kentucky to further inform usability. Five focus groups were then conducted with 19 end users who work in research, public health practice, and cancer patient advocacy. Focus group feedback highlighted different applications of the tool, suggested adjustments to the graphical interface, and recommended additional functionality for future iterations of the tool. In this presentation, we will discuss the different uses for the tool, consider the potential the tool will have on collaboration between different cancer control entities, and describe the next steps for usability testing and implementing changes to the tool.

An Interactive Analysis on Variant Callers and Performance

Presenter: Bryan Zhu
Presenter's email: bzhu@nygenome.org
Institute: New York Genome Center
Principle Investigator: Giuseppe Narzisi

With the recent advances both in sequencing technology and variant calling methods, different variant callers have shown promising progress in identifying as many somatic variants as possible while avoiding false positives. Here, we present our in-house developed benchmarking tool that we utilize to help us better understand how different callers are performing against each other. The key feature of the benchmarking tool is the ability to generate publication-quality reports where the user can explore the underlying data in an interactive way. This is achieved by making extensive use of the Plotly Open Source Graphing Library. The reports illustrate things like caller performance across various variant allele frequency ranges, dynamic stratification by variant types, scoring profiles for each caller, and tendencies for callers to either perform well or perform poorly on certain types of variants or variants in certain regions of the genome. A short demo will be presented to demonstrate the power of this approach on a typical use-case scenario to compare state-of-the-art somatic variant callers on matched tumor/normal pairs from recent high-depth sequencing studies. With this benchmarking tool, we aim to provide a user-friendly and efficient solution to quickly understand how different calling methods affect performance as well as indicators for areas where a caller can improve.

Discovering targetable Vulnerabilities in Cancer Using the Averon Notebook

Presenter: Hongyue (Nicole) Chen
Presenter's email: nicole.chen@emory.edu
Institute: Emory University
Principle Investigator: Andrey Ivanov

Cancer is a global health menace driven by genomic alternations, which results in over 10 million annual deaths. The genetic mutations disrupt crucial cellular processes, especially the intricacies of protein-protein interaction (PPI) networks, which play a pivotal role in this devastating illness. The advances in high-throughput screening technologies enabled comprehensive profiling of mutant-dependent PPIs in cancer cells (e.g., Mo et al., Cell 2022). However, elucidation of functional consequences of mutant-induced changes in PPI networks and their impact on clinical outcomes of cancer patients is highly challenging. To address this unmet challenge, we develop a novel computational platform, termed AVERON, to identify Actionable Vulnerabilities Enabled by Rewired Oncogenic Networks. AVERON is implemented as a Python Jupyter Notebook and serves as a tool for investigating both mutant-induced PPIs and PPIs lost due to the mutations. Based on experimentally determined networks of mutant-directed or neomorph PPIs (neoPPIs), AVERON employs innovative algorithms and statistical techniques to assess the levels of neoPPIs in cancer patients. It examines and visualizes neoPPI impact on clinical outcomes and identifies distinctive sets of signature genes and oncogenic pathways regulated by individual neomorph PPIs. Furthermore, the AVERON can uncover clinically significant neoPPI-regulated genes with available clinical compounds and approved drugs. Together, the AVERON Notebook provides a powerful computational platform to discover molecular mechanisms of neoPPI-dependent tumorigenesis, identify draggable vulnerabilities enabled by mutant-directed PPIs, and inform therapeutic strategies in patients with mutated tumor driver genes.

Latent embeddings of accessible loci for Automatic Cell-Type Annotation in scATAC-seq Data

Presenter: Ariaki Dandawate
Presenter's email: ariad@ds.dfci.harvard.edu
Institute: Dana Farber Cancer Institute
Principle Investigator: Clifford Meyer

The current single-cell ATAC-seq annotation techniques rely on either estimating gene expression from accessible loci or label-transfer from paired scRNA-seq datasets. However gene expression is not always highly correlated with chromatin accessibility, and paired scRNA-seq might not always be readily available or clearly alignable. Consequently, there is a growing need for an automatic and independent method of cell-type identification. However, annotating single-cell ATAC-seq data poses unique challenges primarily because of data sparsity. On the other hand, bulk ATAC-seq data is widely available and provides a strong basis for defining cell types. We present a method to annotate single-cell ATAC-seq data that learns latent embeddings of accessible regions from bulk and pseudo-bulk ATAC-seq data. Using publicly available data representing diverse cell types, we demonstrate that modeling patterns of accessibility is effective for capturing latent relationships between regions, leading to robust cell-type

identification. This automatic annotation method provides an advantage over existing ones that require the user to have pre-existing knowledge of the cell types or systems present in the data. Applied to various single-cell ATAC-seq datasets from common and rare tissue types, the method not only provides accurate annotation, but also provides meaningful insights into complex cell types and states.

Ontology-leveraged metadata harmonization to improve AI/ML application on omics databases

Presenter: Kaelyn Long
Presenter's email: Sehyun.Oh@sph.cuny.edu
Institute: City University of New York
Principle Investigator: Levi Waldron

Efforts to establish comprehensive biological data repositories have been significant at both national and institutional levels. Despite the large volumes of datasets collected from diverse studies, the potential for cross-study analysis within these repositories remains largely untapped due to heterogeneity in metadata structures. This lack of metadata harmonization hinders the application and development of machine learning tools, which can serve a pivotal role in managing the complexity and high dimensionality of multi-omics datasets.
To address this, we initiated the OmicsMLRepo project, aiming to harmonize metadata from various omics data repositories. This project involved manual review of metadata schema, consolidation of similar or identical information, and incorporation of ontologies where applicable. As a result, we have harmonized hundreds of studies on metagenomics and cancer genomics datasets, accessible through the R/Bioconductor packages, curatedMetagenomicData and cBioPortalData. Furthermore, we developed a software package that leverages the ontology incorporated into our curated metadata schema, enhancing the usability of our harmonized metadata. These informatics infrastructures equip large omics datasets for AI/ML applications, making them more accessible to a wide range of research communities.
In summary, the OmicsMLRepo project simplifies the process of data analysis in biological research by providing a unified and user-friendly platform for accessing and analyzing omics data. This can potentially accelerate discoveries and innovations in the field.

terraTCGAdata: accessing and representing TCGA on the AnVIL

Presenter: Marcel Ramos
Presenter's email: marcel.ramos@sph.cuny.edu
Institute: CUNY School of Public health
Principle Investigator: Levi Waldron

NHGRI's Analysis Visualization and Informatics Lab-space (AnVIL) project aims to invert the distributed data warehouse model by centralizing and democratizing large publicly available and restricted-access multi-omics datasets. AnVIL makes use of workflow languages and of data models to connect data storage buckets to compute resources, both of which present challenges for researchers unfamiliar with these systems. Here, we leverage the AnVIL to increase the utility of NCI-funded datasets, starting with TCGA, to allow cancer researchers to conduct resource-intense workflows without large data transfers or in-house compute resources. We present terraTCGAdata, an R / Bioconductor package that interfaces with data stored on AnVIL, providing single-command construction of integrative MultiAssayExperiment representations of multi-omic and pan-cancer TCGA datasets. terraTCGAData on the AnVIL provides an interactive and reproducible environment enabling rapid access and analysis of TCGA public-access data without necessarily learning a workflow language.

Patterns in Flanking Bases of Mutation Signatures

Presenter: Haoyue Feng
Presenter's email: fengx456@bu.edu
Institute: Boston University
Principle Investigator: Joshua Campbell

A range of external exposures or internal biological processes can contribute to the total mutational burden seen in human tumors. Various mutational patterns, known as "mutational signatures," have been identified across diverse tumor types. These signatures can serve as a record of environmental exposure and offer insights into the causes of cancer development. Most studies to date have identified these signatures in tumors using Non-Negative Matrix Factorization (NMF) or a related method. NMF can decompose a matrix of mutation counts, X, into an N x K matrix of mutational signatures and a K x M matrix of mutational exposures, where N is the number of samples, M is the number of mutation types, and K is the number of estimated mutational processes. Each "mutation motif" typically

includes 6 different kinds of mutations (C>A, C>G, C>T, T>A, T>C, T>G) along with the bases immediately surrounding the mutation (i.e., the trinucleotide context). Considering 4 possible bases at each flank, the total number of mutation types reaches 96 (4 bases x 6 mutations x 4 bases). However, NMF fails in characterizing motifs beyond penta-nucleotide context due to the exponentially increasing computational burden as we incorporate flanking bases. For example, a penta-nucleotide context has 1536 mutation types, while there are 24576 mutation types considering its immediate left and right context. To overcome the limitation to NMF, we propose a scalable and efficient Bayesian algorithm, which characterizes novel signatures utilizing a Poisson-based model and further refines mutations under each mutational signature into a group of motifs via a Bayesian mixture model. We have tested the algorithm using bladder cancer data from 412 patients. Considering a varying number of pairs of flanking bases, our current model consistently revealed novel motifs e.g. motifs centered at SBS 28, SBS 10a and SBS 10b along with distinctive base patterns. Overall, this model will provide a streamlined framework for cancer researchers to characterize the mutational patterns with customized length of contexts.

Flash talks

17-May

Session Chair:
Daniel Bergman, Ph.D., Johns Hopkins Medicine

Developing and Deploying the UCSC Xena Gene Set Enrichment Analysis Appyter

Presenter: Cally Lin
Presenter's email: clin125@ucsc.edu
Institute: UC Santa Cruz Genomics Institute
Principle Investigator: Jingchun Zhu

UCSC Xena is a web-based visual integration and exploration tool for multiomic data and associated clinical and phenotypic annotations. It showcases seminal cancer genomics datasets from TCGA, the Pan-Cancer Atlas, GDC, PCAWG, ICGC, and more; a total of more than 1500 datasets across 50 cancer types. Researchers can easily explore public data, their own private data, or both using the Xena Browser. Xena users wanted to be able to run Gene Set Enrichment Analysis (GSEA) on their dynamically created subgroups but the traditional implementation of GSEA was not performant enough for a web application. To enable this functionality for users, we leveraged blitzGSEA, a more performant version of GSEA, implementing it within the Appyter framework developed by the Ma'ayan lab. Appyter allows users to use Jupyter Notebooks to run bioinformatic methods without having to interact with the code directly and also allows developers to dynamically code inputs, such as the subgroups created by a user. The Appyter framework also handles users running analyses concurrently and will cache results if two users run the same analysis, such as part of a workshop. I first compared blitzGSEA to traditional GSEA, ensuring that it provided similar, biologically consistent results. I then developed a blitzGSEA Appyter and implemented a Plotly Dash page to provide users with interactive plots and tables. The Appyter was then deployed on the production Xena Browser using Docker and an AWS EC2 instance. The blitzGSEA Appyter is currently run an average of 500 times a month.

TCPAplus: An LLM-empowered Chatbot for Analyzing a Large Protein Expression Atlas of Human Cancers

Presenter: Wei Liu
Presenter's email: wliu15@mdanderson.org
Institute: MD Anderson Cancer Center
Principle Investigator: Han Liang

Functional proteomics data provide profound insights into cancer mechanisms, aiding the development of novel biomarkers and therapeutic approaches. Here we present a comprehensive functional proteomics resource comprising nearly 8,000 patient samples from The Cancer Genome Atlas and close to 900 samples from Cancer Cell Line Encyclopedia using reverse phase protein arrays. Our protein panel, consisting of approximately 500 high-quality antibodies, spans all major cancer hallmark pathways and is rich in therapeutic targets and biomarkers. To maximize the utility of this invaluable resource, we introduce an intuitive analytical tool, TCPAplus. This platform harnesses state-of-the-art large language models and enables researchers to seamlessly query protein-centric cancer omics data, perform various analyses, visualize findings, and engage in discussions using natural language, thereby significantly easing the translation from intricate proteogenomic data to valuable biological insights.

Optimizing Sample Size for Statistical Learning in Bulk Transcriptome Sequencing: A Learning Curve Approach

Presenter: Yunhui Qi
Presenter's email: qyh601088611@gmail.com
Institute: Iowa State University
Principle Investigator: Li-Xuan Qin

Accurate outcome prediction from transcriptomics data is pivotal for personalized medicine. The success of such endeavors depends on determining a suitable sample size, ensuring adequate statistical power without unnecessary resource allocation or ethical concerns. Current sample size calculation methods for outcome prediction rely on assumptions and algorithms that may not align with modern machine and deep learning techniques. This paper addresses this methodology gap by developing computational approaches to determine the required number of samples for accurate predictions in transcriptomics studies using statistical learning. Our approach establishes the power-versus-sample-size relationship by employing a data augmentation strategy followed by fitting a learning curve. We evaluated its performance for both microRNA and RNA sequencing using data from the Cancer Genome Atlas, considering various data characteristics (such as sample size, marker filtering, and depth normalization) and algorithm configurations (such as model selection, hyperparameter tuning, and offline augmentation), based on a range of evaluation metrics. Python and R codes for implementing our proposed approach are freely available on GitHub. Our study is expected to advance the adoption of statistical learning in biomedical transcriptomics studies and accelerate their reproducible translation into clinically useful predictors, thereby enhancing better disease treatment.

Single cell transcriptomics level surface protein abundance estimation using STREAK.

Presenter: Azka Javaid
Presenter's email: Azka.javaid.gr@dartmouth.edu
Institute: Dartmouth College
Principle Investigator: Hildreth Rob Frost

Current supervised surface protein abundance estimation approaches are limited in that they target a small number of receptors. Second, algorithms that use deep learning-based approaches are not easily interpretable or customizable by medical practitioners. Our method, STREAK (gene Set Testing-based Receptor abundance Estimation using Adjusted distances and cKmeans thresholding), includes two components. The first component is receptor gene set construction, which creates the gene sets for subsequent single cell-level scoring. In this first step, we construct the sets by computing the Spearman rank correlation between the normalized reduced-rank reconstructed RNA transcript counts and between the centered log-ratio-normalized ADT transcript counts using joint scRNA-seq/CITE-seq data. We next perform inference on this set of genes using the Variance-adjusted Mahalanobis (VAM) method. Lastly, we perform a clustered thresholding step using the Ckmeans.1d.dp package. We compare STREAK against both unsupervised and supervised receptor abundance estimation strategies, which include SPECK, the normalized RNA transcript and the Random Forest (RF), Support Vector Machines (SVM) and the cTP-net algorithms. We perform evaluation by quantifying the expression concordance of the estimated abundance values with CITE-seq ADT training counts using two evaluation strategies and six joint CITE-seq datasets, which represent human and mouse tissue types such as the peripheral blood mononuclear cells, mesothelial cells, monocytes and lymphoid tissue. Overall, we observe STREAK to have a superior performance relative to comparative methods. Additionally, we note that since it allows specification of customizable gene sets, STREAK has tremendous clinical utility as an interpretable abundance estimation approach.

Identification of High-Risk Cells in Spatially Resolved Transcriptomics of Cancer Biopsies Using Deep Transfer Learning

Presenter: Debolina Chatterjee
Presenter's email: dchatter@iu.edu
Institute: Indiana University School of Medicine
Principle Investigator: Travis S. Johnson

Spatially resolved transcriptomics is an emerging field in biomedical informatics, encompassing technologies broadly categorized into sequencing-based and imaging-based platforms. The examination of high-risk cells and regions in tissue samples offers meaningful insights into specific disease prognoses. Previously our team developed DEGAS (Diagnostic Evidence Gauge of Single-cells), a sophisticated deep transfer learning algorithm designed to identify high-risk components in single-cell RNA sequencing data from tumors. DEGAS employs latent representation and domain adaptation to link disease attributes in patients to individual cells.
We propose that by integrating spatial location information from spatially resolved transcriptomics platforms, DEGAS can not only identify high-risk components in tissue samples but also pinpoint locations within the slides associated

with disease status, outcomes, and pathology image features. To gauge DEGAS's versatility across diverse platforms, we conducted experiments by overlaying patient risk scores on Triple Negative Breast Cancer data from the 10X Genomics Visium platform collected at the Indiana University School of Medicine. Additionally, we applied DEGAS to publicly available Nanostring's CosMx FFPE samples from normal and Hepatocellular Carcinoma samples, revealing high-risk regions within tissue that align with proliferation signatures, lymphocyte infiltration patterns, and regions of angiogenesis. The results indicate that the high-risk regions are frequently enriched for tumor tissue. Within these tumor regions, DEGAS reveals heterogeneity in risk that correlates with markers for aggressive disease and cell type heterogeneity adding additional nuance to our understanding of these biopsies.

Prediction of Multi-Class Peptides by T-cell Receptor Sequences with Deep Learning

Presenter: Phi Le
Presenter's email: philong.le@ucsf.edu
Institute: University of California San Francisco
Principle Investigator: Li Zhang

Predicting T-cell receptor (TCR) binding to antigen peptides is crucial for understanding the immune system and developing new treatments for diseases like cancer. As experimental procedures are expensive, machine learning methods are gaining interest in learning TCR-antigen peptide bindings. However, existing computational tools to predict the binding of a pair of TCR and antigen peptide depend heavily on how the methods artificially generate non-binding pairs for training. Moreover, there has been little discussion on prediction based on integrating TCR sequences, HLA types, and V /J genes. To overcome these limitations, we propose a two-step framework to predict multiple antigen peptides recognized by TCR sequences. The first step involves a feature engineering process to deal with two different types of variables: 1) proposing various types of neural networks inspired by language representation models to learn embeddings of letter-based TCR and peptide amino acid sequences; and 2) employing a categorical encoding method for HLA types and V/J genes. The second step pertains to building a prediction model to identify the specific antigen to which a TCR binds. We trained our models on three large publicly available databases and validated the performances of encoding methods and prediction models on the scenarios in which our feature engineering saw and unsaw the peptides. Our results show that this feature engineering improves our classification results compared to other methods and including HLA and V/J information significantly improves prediction performance with an AUC of at least 0.96 for predicting antigen classes for even out-of-distribution data.

Image-based approaches

17-May

Session Chair:
Alexander Wenzel, Ph.D., University of California San Diego

Human-in-the-loop deep learning-based segmentation of rectal tumors on MRI

Presenter: Michael Kong
Presenter's email: mfk58@case.edu
Institute: Case Western Reserve University
Principle Investigator: Satish Viswanath

Precise localization and segmentation of rectal cancer tumors on routine MRI is critical for accurate clinical staging and treatment selection. In this study, we present an integrated human-in-the-loop deep learning (DL) model for semi-automated segmentation of rectal tumors on pre-treatment T2-weighted MRI scans across multiple acquisition planes (axial, coronal).
Using a retrospective, multi-center dataset of baseline MRI scans from rectal cancer patients, a DL tumor segmentation model was trained using available radiologist tumor annotations on axial and coronal T2-weighted MRIs from two institutions. On a separate cohort for human-in-the-loop (HITL) learning, DL-generated delineations were manually refined by two radiology readers. Refined annotations were then used to re-optimize the model to output final tumor segmentations on MRI. Model performance was evaluated before and after HITL refinement against “ground truth” radiology annotations within a holdout validation cohort.
381 MRI scans from 231 rectal cancer patients were included. After one round of HITL refinement, DL-based segmentations yielded the best overall performance on the external validation cohort (N=20, two institutions) for both axial (DSC = 0.76) and coronal scans (DSC=0.75). Comparatively, the pre-HITL model yielded significantly worse performance on axial (DSC=0.60, p=0.0362) and coronal (DSC=0.62, p=0.0186) scans. HITL refinement for rectal

tumor annotation also significantly improved reader efficiency by 20 minutes ($p<0.001$).
Human-informed DL models can aid in accurately delineating tumors on MRI scans from rectal cancer patients, with model generalizability and reader efficiency across multi-center cohorts.

Topological uncertainty for vascular segmentation

Presenter: Saumya Gupta
Presenter's email: saumgupta@cs.stonybrook.edu
Institute: Stony Brook University
Principle Investigator: Chao Chen

Modeling vascular structures is pivotal in cancer imaging informatics, as it aids in understanding tumor biology and enhancing patient care. In particular, vasculature reveals angiogenesis within and around tumors, provides insights into the aggressiveness of cancer, and its potential for growth and metastasis, and thus helps understanding tumor behavior and devising targeted therapies. Their segmentation, however, is challenging due to relatively weak signals and complex geometry/topology. Furthermore, these fine-scaled 3D structures are hard to annotate even for humans. Thus obtaining large-scale annotated datasets for supervised training of deep learning models is challenging. To facilitate this, it is necessary to incorporate smart annotation strategies, to efficiently leverage human input.
In this work, we focus on estimating the uncertainty of deep learning segmentation models, so that highly uncertain, and thus error-prone structures can be identified for human annotators to verify. Unlike existing works, which provide pixel-wise uncertainty maps, we stipulate it is crucial to estimate uncertainty in units of topological structures, e.g., small pieces of connections and branches. To achieve this, we leverage tools from topology, specifically discrete Morse theory (DMT), to capture the structures, and then reason about their uncertainties. To model this, we propose a joint prediction model that estimates the uncertainty of a structure while taking the neighboring structures into consideration (inter-structural); and propose a novel ProbabilisticDMT to model the inherent uncertainty within each structure (intra-structural) by sampling its representations via a perturb-and-walk scheme. On various datasets, our method produces better structure-wise uncertainty maps compared to existing works, paving the way for improved diagnostic and treatment strategies.

Topology-Guided Vasculature Analysis

Presenter: Michael Yao
Presenter's email: jiacyao@cs.stonybrook.edu
Institute: Stony Brook University
Principle Investigator: Chao Chen

Research has established a link between vasculature morphology and the advancement of cancer prediction and patient care. While medical image segmentation has seen significant progress, there remains a lack of investigation into how precise vascular segmentation could enhance the classification of medical imaging. Innovative methods like ensemble, feature fusion, knowledge distillation, and gating have been introduced to merge additional information for improved image classification. These strategies incorporate vessel segmentation into image analysis but lack interpretability and overlook the spatial relationship between vascular structures and the image.
Addressing this, our study introduces a novel topology-guided framework that utilizes the vessel mask as auxiliary input, emphasizing the spatial interplay between vessel features and image patches. This approach enhances spatial contextual awareness by focusing on local topological features extracted from the vessel mask. We validate the importance of these features with significant p-value findings in hypothesis testing. Our "topology-guided attention" model then utilizes these features to direct the attention of a transformer to regions dense in vascular structures. This method aims to bolster diagnostic accuracy and efficiency by improving the network's capacity to identify and assess vascular-rich areas in medical images, demonstrating a novel way to integrate vascular information into image classification. We demonstrate consistent performance across two datasets: a 2D Retinal dataset, featuring retinal vessels affected by various diseases, and a 3D dataset that examines lung vessel structures in relation to lung cancer treatment outcomes.

Learning without Real Data Annotations to Detect Hepatic Lesions in PET Images

Presenter: Xinyi Yang
Presenter's email: xinyi.2.yang@cuanschutz.edu
Institute: University of Colorado
Principle Investigator: Fuyong Xing

Background and Significance: Deep neural networks have been recently applied to lesion identification in fluorodeoxyglucose positron emission tomography (PET) images, but they typically rely on a large amount of well-annotated data for model training. This is extremely difficult to achieve for neuroendocrine tumors (NETs), because of low incidence of NETs and expensive lesion annotation in PET images. This study designs a novel, adaptable deep learning method, which uses no real lesion annotations but instead low-cost, list mode-simulated data, for hepatic lesion detection in real-world clinical NET PET images. The method significantly reduces human effort for data annotation.
Methods: We first propose a region-guided generative adversarial network (RG-GAN) for lesion-preserved image-to-image translation. Then, we design a specific data augmentation module for our list-mode simulated data and incorporate this module into the RG-GAN to improve model training. Finally, we combine the RG-GAN, the data augmentation module, and a lesion detection neural network into a unified framework for joint-task learning to adaptatively identify lesions in real-world PET data.
Results: The proposed method outperforms recent state-of-the-art lesion detection methods in real clinical 68Ga-DOTATATE PET images, and produces very competitive performance with the target model that is trained with real lesion annotations. With RG-GAN modeling and specific data augmentation, we can obtain good lesion detection performance without using any real data annotations.
Impact: Our method eliminates the need of out-of-domain target data annotation for automated hepatic lesion detection in PET images and thus reduces the cost of data preparation for AI model deployment in real applications. In addition, it improves model generalizability for cross-domain/-scanner lesion detection with PET imaging.

Intra- and peri-tumoral radiomic features are predictive of pathologic response to multiple neoadjuvant therapy regimen in rectal cancers via pre-treatment MRI

Presenter: Benjamin Parker
Presenter's email: lkb44@case.edu
Institute: Case Western Reserve University
Principle Investigator: Satish Viswanath

Rectal cancers undergo neoadjuvant chemoradiation prior to resection, with increasing evidence that neoadjuvant chemotherapy or immunotherapy can boost response rates. This study evaluated whether radiomic features derived from intra- and peri-tumoral regions on pretreatment MRI scans could identify which patients will achieve complete response.
Retrospective, multi-center study of baseline MRI scans from rectal cancer patients who had undergone routine neoadjuvant chemoradiation or experimental immunotherapy + chemoradiation regimen. 916 radiomic features were extracted from annotated tumor and a 10mm peritumoral region. Top 5 radiomic features were identified and used to train a machine learning model to predict likelihood of achieving complete response. Model performance was validated via ROC analysis against response groups defined via (i) tumor regression grade for patients sent to surgery, and (ii) 1-year clinical complete response for patients who underwent non-operative management. Clinical variables were also statistically compared between response groups.
Training cohort comprised 64 patients (2 institutions) who underwent chemoradiation alone prior to surgery, where radiomic machine classifier yielded the best performance for identifying pathologic complete responders via baseline MRIs (AUC=0.765 ± 0.054) . In an external validation cohort of 37 patients who underwent experimental immunotherapy prior to chemotherapy+chemoradiation, this classifier maintained excellent performance in identifying clinical complete responders (AUC=0.7). Baseline CEA levels (p=0.3385) and clinical stage (p=0.3386) lacked statistical significance as a predictor of response.
Intra- and peri-tumoral radiomic features from baseline MRI scans can predictively identify clinical or pathologic response after neoadjuvant treatment regimen in rectal cancers.

Utilizing Deep Learning and Channel-Wise Data Fusion for End-to-end Organelle-based Breast Cancer Cell Classification

Presenter: Harrison Yee
Presenter's email: yeeh@rpi.edu
Institute: Rensselaer Polytechnic Institute
Principle Investigator: Magarida Barroso

Microscopy-based classification of cancer cells has traditionally focused on easily observable features, such as cell morphology and pleomorphism. Recently, we have used handcrafted feature extractors and machine learning classifiers on fluorescent confocal microscopy images of cultured breast cancer cells to identify specific features related to subcellular organelle morphology and organelle topology (inter-organelle distance distribution). Notably, the most discriminative features between breast cancer cell lines were related to organelle topology rather than

morphology. However, the previous approach required manual annotations and extensive image processing steps through external software for organelle object 3D-rendering. Herein, we introduce a deep learning approach using a patch-based convolutional neural network (CNN) with channel-wise intermediate data fusion to perform end-to-end breast cancer cell classification from fluorescent confocal microscopy images. First, microscopy images are preprocessed into 2D patches through patch extraction and threshold-based sparsity filtering. Subsequently, a channel-wise marginal intermediate fusion network is implemented to perform separate feature analyses of each organelle of interest, ultimately leading to the classification of each patch. By incorporating patch extraction and data fusion, we alleviate common issues with data scarcity and image preprocessing while primarily focusing analysis on features related to specific organelles of interest. Our methodology outperforms typical single organelle-based image classifiers and early fusion methods when tested on a dataset comprised of microscopy images from 6 different breast cancer cell lines, achieving a classification accuracy of 95.7%. These findings support the potential of this approach for more robust microscopy-based organelle analysis for cancer cell line research in the future.

Clinical and population research

17-May

Session Chair:
Daniel Bergman, Ph.D., Johns Hopkins Medicine

Transparent and Efficient Adaptive Designs for Clinical Trials

Presenter: Yingjie Qiu
Presenter's email: yingqiu@iu.edu
Institute: Indiana University
Principle Investigator: Yong Zang

Abstract not available

A Flexible and Adaptive Framework for High-Dimensional Fairness-Aware Integration of Data from Multiple Sites

Presenter: Yi Lian
Presenter's email: yi.lian@pennmedicine.upenn.edu
Institute: University of Pennsylvania
Principle Investigator: Qi Long

In healthcare data analytics, fairness issues can arise from the heterogeneity between the groups defined by sensitive attributes such as race/ethnicity, and the imbalance in their sample sizes. If the groups are modeled jointly, the inference or prediction results are likely dominated by the majority group thus biased for the minority groups. If modeled separately, although enjoying more flexibility, the minority groups may not have enough data volume to generate accurate and reliable results. These issues can happen even if the data is a representative sample of the population therefore fairness-aware statistical analysis methods are needed. In this work, we propose a flexible multitask regression framework that improves the prediction fairness for the minority groups, by adapting to the latent between-group heterogeneity in a data-driven manner, under high-dimensional settings. In addition, our work is in line with the notion of individual fairness in machine learning. Through extensive numerical experiments, we show that our method can improve the prediction accuracy for minority groups comparing to ordinary modeling strategies.

Disparities in the Documentation of Social Determinants of Health ICD-10 Z-Codes for Patients Diagnosed with Cancer: An Epic Cosmos Study

Presenter: Yiming Zhang
Presenter's email: yiming.zhang14@umassmed.edu
Institute: UMass Chan Medical School
Principle Investigator: Feifan Liu

Abstract not available

Pilot Testing the User Interface for a Novel Algorithm to Process Electronic Adherence Monitoring Device Data

Presenter: Julia Herriott
Presenter's email: julia.herriott@cchmc.org
Institute: Cincinnati Children's Hospital Medical Center
Principle Investigator: Meghan E. McGrady

Medication adherence is a National Cancer Institute priority, with electronic adherence monitoring devices (EAMDs) being the preferred research measure. However, widespread EAMD adoption is hindered by inadequate tools for accurately converting raw EAMD actuations into adherence data. This abstract reports on the results of pilot testing of a user interface designed to execute a novel EAMD data processing algorithm.
Methods: Usability data were obtained via an observational study with likely end-users (n = 5 Research Coordinators; n = 1 Principal Investigator) with 0.25-16 years' EAMD experience (M[SD] years = 4.05[6.70]). Participants were asked to "think aloud" as they processed a test dataset and then completed the System Usability Scale, System Usability Scale Adjective Rating, Net Promotor Score, and investigator-created measures assessing usability refinements and satisfaction. Results were summarized using descriptive statistics.
Results: 83.33% of users described the product's user-friendliness as "excellent," with System Usability Scale (SUS)

scores exceeding recommended cut-points (M[SD] SUS = 86.67[11.37]; M[SD] SUS Adjective Rating = 5.83[0.41]). 83% of users agreed the product saves time/resources and 100% agreed it produced more accurate and reproducible data. All users (100%) indicated interest in using the product (Net Promoter Score = 83%). Despite these strengths, the majority of users identified additional features related to tracking changes, integrating self-report adherence data, and indicating changes to medication regimen as "absolutely necessary."
Conclusions: Users are excited about the promise of the new algorithm and associated user interface but identified critical features that need to be integrated prior to widespread uptake.

The TPS2TOPAS interface tool for investigating new radiotherapy devices

Presenter: Ramon Ortiz
Presenter's email: ramon.ortizcatalan@ucsf.edu
Institute: University of California San Francisco
Principle Investigator: Bruce Faddegon

The Monte Carlo (MC) method, a key tool for radiotherapy dose calculations, remains underutilized in research for clinical medical physics. A reason for its underuse is the time-consuming execution and user error-prone conversion from treatment plan parameters of DICOM-RT files to MC parameters. To address the latter, we developed the TPS2TOPAS interface, which facilitates exporting DICOM-RT parameters from clinical treatment planning systems (TPS), to ready-to-run TOPAS parameter control files (PCFs). These parameters include multileaf-collimator (MLC) and jaw positions, gantry rotation, beam MU, patient setup, a selection of different multileaf-collimator models, variance reduction technique parameters, dedicated scoring quantities and output formats. We verified TPS2TOPAS using the Truebeam 120 millenium MLC, and several clinical cases including 3D-CRT, IMRT and VMAT. Dose distributions computed using these PCFs agreed with plan doses from RayStation TPS (>98% pass ratio in the 3%/3mm gamma test) across the various clinical cases and techniques, confirming the correct plan parameterization. The interface is poised to expand TOPAS use in research for clinical medical physics. To illustrate its utility, we assessed the dosimetric impact and potential neutron contamination of a temporary tissue expander (TTE), composed of high-density materials (neodymium, 7.4 g/cm3) and used in postmastectomy radiotherapy. We found that the presence of the TTE has a non-negligible impact on the dose received by patient structures, reducing up to a 19.3% the dose to the breast tissue distal to the ports. Additionally, we found no photoneutrons were produced by the TTE, having no effect on the equivalent neutron dose.

NCI ITCR Trainee Symposium Organizing Committee

Andrey Ivanov, Emory University, Chair

Juli Klemm, NIH/NCI

Candace Savonen, Fred Hutch Cancer Center

Carrie Wright, Fred Hutch Cancer Center

Drew Jones, New York University

Greg Caporaso, Northen Arizona University

Jeff Buchsbaum, NIH/NCI

Li Zhang, University of California San Francisco

Li-Xuan Qin, Memorial Sloan Kettering Cancer Center

Rao Divi, NIH/NCI

Rudolf Pillich, University of California San Diego

Satish Viswanath, Case Western Reserve University

Taxiarchis Botsis, Johns Hopkins University School of Medicine