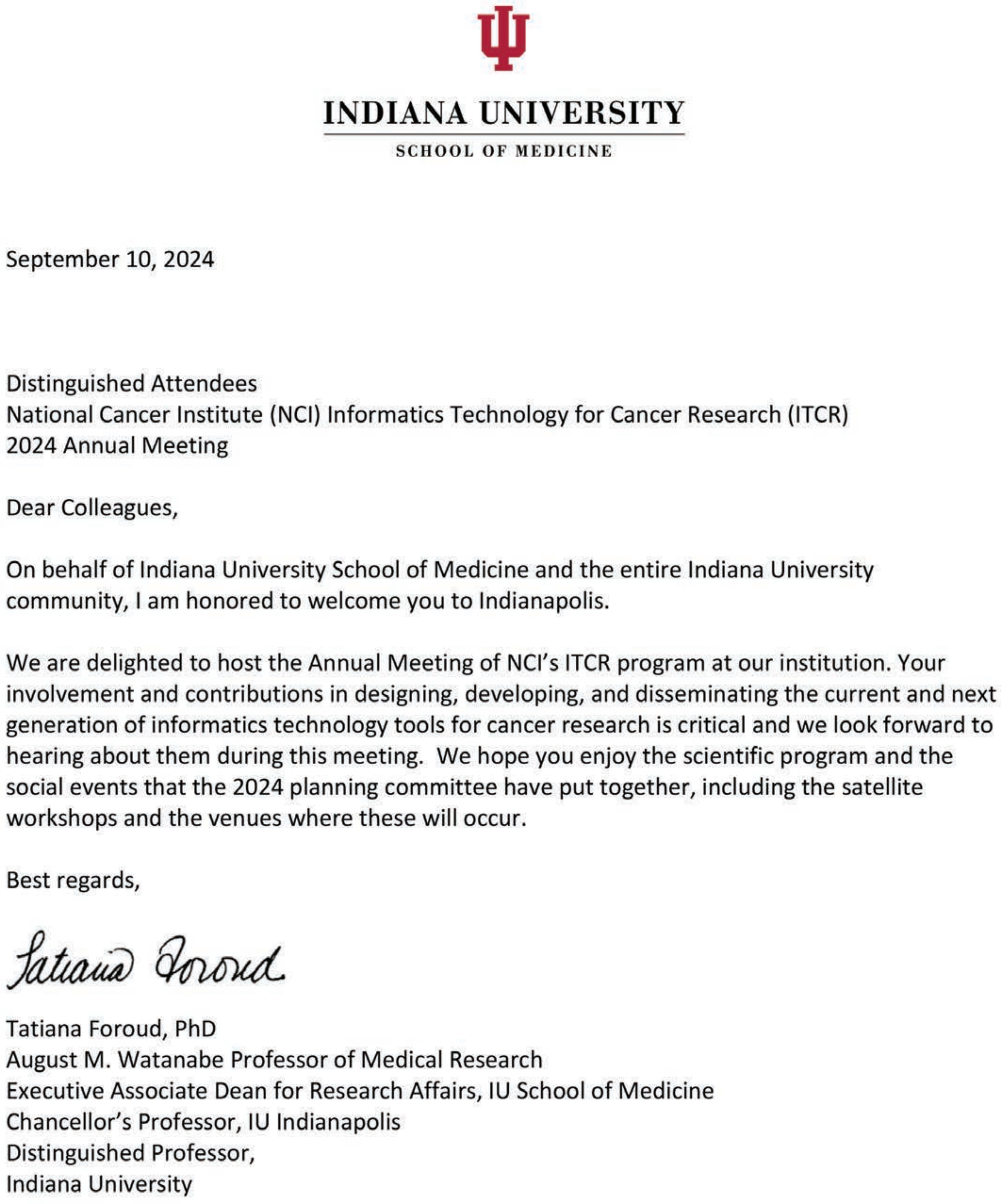
NIH/NCI INFORMATICS TECHNOLOGY FOR CANCER RESEARCH (ITCR) ANNUAL MEETING

September 16-19, 2024, Indianapolis, IN

Welcome by the Executive Associate Dean for Research Affairs



INDIANA UNIVERSITY

SCHOOL OF MEDICINE

September 10, 2024

Distinguished Attendees
National Cancer Institute (NCI) Informatics Technology for Cancer Research (ITCR)
2024 Annual Meeting

Dear Colleagues,

On behalf of Indiana University School of Medicine and the entire Indiana University community, I am honored to welcome you to Indianapolis.

We are delighted to host the Annual Meeting of NCI's ITCR program at our institution. Your involvement and contributions in designing, developing, and disseminating the current and next generation of informatics technology tools for cancer research is critical and we look forward to hearing about them during this meeting. We hope you enjoy the scientific program and the social events that the 2024 planning committee have put together, including the satellite workshops and the venues where these will occur.

Best regards,

Tatiana Foroud

Tatiana Foroud, PhD
August M. Watanabe Professor of Medical Research
Executive Associate Dean for Research Affairs, IU School of Medicine
Chancellor's Professor, IU Indianapolis
Distinguished Professor,
Indiana University

Table of Contents

General Information

Social Activities

Morning Runs: Meeting daily at 6am in the entrance of JW Marriott. This activity is subject to attendees registering for the activity by Monday 9/16 noon. If interested please indicate your interest by emailing us at comppath@iu.edu

Monday, 9/16/2024 @ 6pm: Evening Meet Up: Casual Pub Night
- Fat Roost Diner, Hyatt Regency Indianapolis, One South Capitol Ave.

Tuesday, 9/17/2024 @ 5pm: Poster Session, Hors d'oeuvres, & Dinner @ The Museum
- Indiana State Museum, 650 W Washington St, Indianapolis

Wednesday, 9/18/2024 @ 6pm: Shuttle service to a Surprise Reception!

Sponsors

Melvin and Bren Simon Comprehensive Cancer Center, Indiana University
Division of Computational Pathology, Dept of Pathology & Laboratory Medicine, IUSM
Research Center for Federated Learning, IUSM
Dept of Pathology & Laboratory Medicine, IUSM
Dept of Biostatistics & Health Data Science, IUSM
Center for Computational Biology & Bioinformatics, IUSM

Scientific Planning Committee

Spyridon Bakas	(Indiana University)
Travis S Johnson	(Indiana University)
Juli Klemm	(NIH/NCI)
Kristen Naegle	(The University of Virginia)
Prateek Prasanna	(Stony Brook University)
Veronica Rotemberg	(Memorial Sloan Kettering)
Candace Savonen	(Fred Hutchinson Cancer Center)
Carrie Wright	(Fred Hutchinson Cancer Center)

Advisory Committee (General Chairs of Prior ITCR Meetings)

Kooresh Shoghi	(Washington University in St Louis)
Samuel Volchenboum	(University of Chicago)

Administrative Support

Sally Atcheson	(Indiana University)
Angeliki Papavasileiou	(Indiana University)

Volunteers

(Division of Computational Pathology, Dept of P&LM, IUSM)

Sanyukta Adap
Bhakti Baheti
Ujjwal Baid
Shubham Innani
Sarthak Pati
Siddhesh Thakur
Jayden You

Video Contest Judges

Chelsea Allanigue	(IU Biomedical Informatics Club)
Bhakti Baheti	(Assistant Professor, IUSM)
Ujjwal Baid	(Assistant Professor, IUSM)
Mike Enberg	(Marketing Consultant)
Samanta Roth	(NCI Office of Communications & Public Liaison)
Christina Sisti	(NCI Patient Advocate)
Sahiti Somalraju	(IU Biomedical Informatics Club)

Social Media

Please use #ITCR2024

Program Overview

	Monday, Sep. 16, 2024		Tuesday, Sep. 17, 2024		Wednesday, Sep. 18, 2024		Thursday, Sep. 19, 2024	
	Location	Pre-Meeting Day	Location	Day 1	Location	Day 2	Location	Day 3
06:00-06:30			JW	Morning Run	JW	Morning Run	JW	Morning Run
06:30-07:00								
07:00-07:30								
07:30-08:00								
08:00-08:30			Eiteljorg Museum	Breakfast	Eiteljorg Museum	Breakfast	Eiteljorg Museum	Breakfast
08:30-09:00						Oral Session 4 (*Imaging*)		Oral Session 8 (*Clinical*)
09:00-09:30				Welcome / Opening Session		Oral Session 5 (*Spatial*)		Oral Session 9 (*Molecular*)
09:30-10:00				Keynote 1		Short Coffee Break		Short Coffee Break
10:00-10:30						Keynote 2		Keynote 3
10:30-11:00				Short Coffee Break				
11:00-11:30				Patient Advocacy Panel		Lightning Trainee Session 1		Lightning Trainee Session 2
11:30-12:00		Registration Desk Opens						
12:00-12:30	Eiteljorg Museum	ITCR Training Network Workshops		Lunch		Lunch		Lunch boxes
12:30-13:00						Video Contest		Spatial Transcriptomics Workshop
13:00-13:30				Oral Session 1 (Imaging)		Sustainability Session		
13:30-14:00								
14:00-14:30				Oral Session 2 (*Spatial*)				
14:30-15:00				Coffee Break		Coffee Break		
15:00-15:30				Oral Session 3 (*Omics*)		Oral Session 6 (*Omics*)		
15:30-16:00								
16:00-16:30				New Award Lightning Talks		Oral Session 7 (Clinical)		
16:30-17:00								
17:00-17:30			Indiana State Museum	Poster Session & Hors d'oeuvres				
17:30-18:00								
18:00-18:30	HYATT	Evening Meet Up: Casual Pub Night			JW	Shuttle service to reception		
18:30-19:00								
19:00-19:30				Dinner @ the Museum	Surprise	Reception @ Local Comm. Suggestion		
19:30-20:00								
20:00-20:30								
20:30-21:00								

Museum Information

Wi-Fi is free throughout select areas of the museums including the Museum cafés.

Photography with **no flash** is allowed for personal use in the **permanent galleries** unless noted otherwise (i.e. a label may say "no photo" on it). Video is not allowed in **permanent galleries**.

Phone: 317.636.9378 - Eiteljorg
317.232.1637- Indiana State Museum

Address: 500 W. Washington St. - Eiteljorg
650 W. Washington St. - Indiana State Museum
Indianapolis, IN 46204

Keynote Speakers

Debra JH Mathews, PhD, MA, is the Associate Director for Research and Programs for the Johns Hopkins Berman Institute of Bioethics, and a Professor in the Department of Genetic Medicine, Johns Hopkins University School of Medicine. Dr. Mathews runs the Genomics and Society Mentorship Program and serves as the Chair of the Berman Institute's Inclusion, Diversity, Anti-Racism, and Equity (IDARE) Committee. Within the Institute for Assured Autonomy (IAA), Dr. Mathews serves as the Ethics & Governance Lead. In this role, she leads work focused on the ethical, societal, and governance implications of autonomous systems, and identifies opportunities across IAA for the integration of ethics and governance work and priorities.

Dr. Mathews's academic work focuses on ethics and policy issues raised by emerging technologies, with particular focus on genetics, stem cell science, neuroscience, synthetic biology, and artificial intelligence. Dr. Mathews is a member of the steering committee of The Hinxton Group, an international collective of scientists, ethicists, policymakers and others, interested in ethical and well-regulated science, and whose work focuses primarily on stem cell research. She has been an active member of the International Neuroethics Society since 2006, has been on the Society's Board of Directors since 2015, and is currently serving as President of the Society. In addition to her academic work, Dr. Mathews has spent time at the Genetics and Public Policy Center, the US Department of Health and Human Services, the Presidential Commission for the Study of Bioethical Issues, and the National Academy of Medicine working in various capacities on science policy.

Dr. Mathews earned her PhD in genetics from Case Western Reserve University, as well as a concurrent Master's in bioethics. She completed a Post-Doctoral Fellowship in genetics at Johns Hopkins, and the Greenwall Fellowship in Bioethics and Health Policy at Johns Hopkins and Georgetown Universities.

Michael Feldman, MD, PhD, is an Endowed Chair Professor and the Chairman of the Department of Pathology & Laboratory Medicine at Indiana University School of Medicine in Indianapolis. Dr. Feldman's professional interests revolve around the development of institutional biobanking as well as the integration and adoption of information technologies in the discipline of healthcare, and particularly pathology. His research and technology interests are generally focused on bringing technology and tools out to clinical care but have widened to now thinking about this at the enterprise level and how to accommodate a wider range of needs, interests, capabilities, and varying enthusiasm for change.

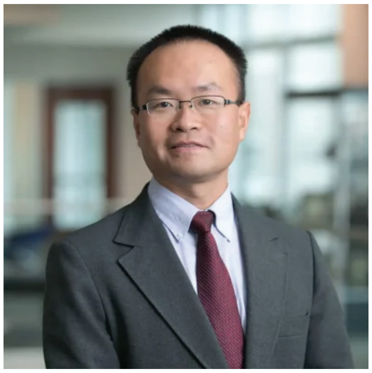
Dr. Feldman's work in the field of tissue banking included being an early funded adopter and tester in the NCI caBIG project, where they partnered on the testing of caTissue with the University of Pittsburgh and were the first cancer center using both caTissue and caTIES. caTIES has now morphed into a multicenter Tissue Collaborative Research Network (TCRN) and is NIH funded.

In the field of digital image analysis, Dr Feldman has been funded by the NIH, Synergy award from DOD, as well as industry-sponsored projects on several fronts including interactions between pathology/radiology (Radio-patho-genomics of prostate cancer and breast carcinoma), development and utilization of computer assisted diagnostic algorithms for machine vision in prostate and breast cancer. More recently he has been involved in the development of deep learning methods for complex interrogation of pathology slides both within the cancer domain, as well as in cardiovascular and renal pathology. He is also interested in solid tumor minimal residual disease. In this area he has been studying disseminated tumor cells (DTC) in the bone marrow of high-risk breast cancer patients. Using flow cytometry his group has developed high throughput flow to elucidate residual tumor burden in these patients. Randomized clinical trials are now in place to target these cells focusing on autophagy, MTOR and cMET pathways to interrupt these pathways that have been shown to play a role in DTC survival in both animal models, as well as single cell sequencing studies in these high-risk patients.

Dr. Feldman earned his MD, PhD in Pathology from Rutgers University. He completed a Post-Doctoral Fellowship in immunology at University of Pennsylvania. Before joining IU he has been acting as the Vice Chair for Clinical Services at the Department of Pathology of the Perelman School of Medicine at the University of Pennsylvania.

Peter Mattson, PhD, is the President of the Board of the MLCommons Association, a board member at the AI Verify Foundation, and a Senior Staff Engineer at Google. On the MLCommons board, he sits on the Compensation and Diversity, Equity, and Inclusion (DEI) committees. Dr. Mattson is also actively involved in the technical leadership of MLCommons, where he is currently Co-chair of the MLCommons AI Safety Benchmarking Working Group. In this role, Dr. Mattson is leading the effort to define global standard benchmarks for AI safety; he has given invited briefings on the effort to the NTIA, NIST, OSTP, US AISI, EU AI Office, and Singapore's IMDA. He previously co-founded the working groups for the MLPerf hardware speed benchmark, the MedPerf federated evaluation platform for healthcare, and the Croissant dataset metadata format.

Dr. Mattson's research focuses on evaluating the safety, quality and effectiveness of models and understanding how improving datasets can drive better training and evaluation. He is actively involved in the program committees of NeurIPS Datasets and Benchmarks track and the ICML DMLR workshop, as well as serving as a founding editor of the Journal of Data Central Machine Learning. Dr. Mattson sits on the National Science Foundation's Advisory Committee for Cyberinfrastructure. Previously, he led the Programming Systems and Applications Group at Nvidia Research, was VP of software infrastructure for Stream Processors Inc (SPI), and was a managing engineer at Reservoir Labs.
Dr Mattson earned his PhD in electrical engineering from Stanford University, as well as a Master's degree in electrical engineering. While there, he received concurrent National Science Foundation and Stanford Graduate Fellowships. He also earned a Bachelor of Science degree in electrical engineering from the University of Washington.



Kun Huang, PhD, FAIMBE, Dr. Kun Huang received his BS degrees in Biological Science and Computer Science from Tsinghua University in 1996 and his MS degrees in Physiology, Electrical Engineering, and Mathematics all from University of Illinois at Urbana-Champaign (UIUC). He then received his PhD in Electrical and Computer Engineering from UIUC in 2004 with a focus on computer vision and machine learning. He was a faculty member in the Department of Biomedical Informatics at The Ohio State University (OSU) from 2004 to 2017 where he served as the Associate Dean for Genome Informatics in the College of Medicine. Currently he is the IUSM PHI Endowed Chair for Genomic Data Science, Professor and Chair of the Department of Biostatistics and Health Data Science at Indiana University School of Medicine and Fairbanks School of Public Health. He is also the Associate Director for Data Science of the IU Simon Comprehensive Cancer Center and a member of the Regenstrief Institute. His research interests include bioimage informatics, computational pathology, translational bioinformatics, and heath data science. He is an elected Fellow of American Institute of Medical and Biological Engineering (AIMBE) and he research has been funded by various funding agencies and programs including the NCI ITCR program. Dr. Huang will be giving the opening session remarks.

Patient Advocates

Susan Gubar, PhD, A Distinguished Professor Emerita of English at Indiana University, Susan Gubar has collaborated with Sandra Gilbert for half a century on a series of publications from The Madwoman in the Attic to the Norton Anthology of Literature by Women and Still Mad. Some of this work earned them the Ivan Sandrof Lifetime Achievement Award from the National Book Critics Circle. Her most recent solo books include two memoirs: Memoir of a Debulked Woman and Late-Life Love.

From 2012 until 2021, Susan wrote the column "Living with Cancer" for the online New York Times. She received the Natalie Davis Spingarn Writers Award from the National Coalition for Cancer Survivorship in 2014 and the Lifetime Scholarly Achievement Award from the Modern Language Association in 2021. Her current book-in-production focuses on very creative old ladies. Grand Finales: The Creative Longevity of Women Artists spotlights novelists, poets, painters, musicians, dancers, and sculptors who transformed the last stage of existence into a rousing conclusion. Susan lives in Bloomington, Indiana.
Photo: Julie Gray

Dana Cattani, A cancer survivor since 2012, Dana Cattani advocates for improved cancer support services at the individual, community, and national levels. Motivated by her own experience, she informally has mentored over 50 patients through a range of stages and outcomes: diagnosis, treatment, and survivorship, as well as recurrence and end-of-life decisions. She also serves on regional advisory and development committees at Cancer Support Community—South Central Indiana. In addition, she speaks to national cancer organizations about effective communication, mental health, and the importance of patient voices and perspectives.

Prior to this current work, Cattani spent more than 30 years teaching or volunteering at schools in California, North Carolina, and Indiana. She has taught language arts to elementary students, literature to high schoolers, professional writing to undergraduates, and research skills to graduate students. In addition, she has supervised student teachers, led pedagogy workshops, and coached communication and presentation skills in business executive education programs. In every context, she helps people sharpen their thinking and writing to enhance their effectiveness and expand their options for the future.

This summer, in celebration of another year of life, she reread George Elliot's Middlemarch, took beginning poker lessons, and learned to keep her balance (mostly) on a stand-up paddle board in a peaceful lake.

Patient Advocates

Dr. Terry Whitt Bailey, is admired as a community activist, arts administrator and advocate for education and health equity.

Terry currently serves as the President & CEO of Cancer Support Community Indiana, providing free cancer survivorship programs and services that empower and strengthen individuals who are impacted by cancer. While most services are provided at CSC's main campuses in Indianapolis and Bloomington, services are also offered at Franciscan Health Indianapolis and Lafayette, at Community Health Network (North, South, East & Anderson), IU Health North (Carmel) and Hendricks Regional Health.

Prior to her appointment, Terry had a successful career as an arts administrator, serving as President & Chief Executive Officer of the Historic Madame Walker Theatre Center in Indianapolis and Cornerstone Center for the Arts in Muncie (IN). Terry previously held administrative positions with Ball State University, most notably as the Director of Executive Staff & Administrative Affairs (Chief of Staff) for the Office of the President.

Terry is a Certified Nonprofit Consultant by the National Association of Nonprofit Organizations & Executives and currently serves as Co-Chair of the Indiana State Legislative Arts Advocacy Council. She has received three Mayor's Community Service Awards, the Indiana Community Arts Leadership Award, the Minority Achievement Award from Center for Leadership Development, the Very Important Volunteer Award (V!VA), the Outstanding Service Award from Rotary International and the prestigious Athena Award from Women in Business Unlimited (WIBU), Inc.

Terry currently serves on the following Boards: IU Health-East Central Region, Washington Township Schools Foundation and Meridian Health Services. In addition, she is a member of the Indianapolis Alumnae Chapter of Delta Sigma Theta Sorority, Inc. and the National Council of Negro Women-Indianapolis Section.

Terry earned two Bachelors degrees from Rutgers University and her Masters degree from the University of California, Los Angeles (UCLA). She completed post-graduate work in Management Development at Harvard University, and earned a Doctor of Ministry Degree (D.Min) from Newburgh Theological Seminary.

Janet Freeman-Daily, Janet Freeman-Daily is a writer, speaker, and international cancer research advocate who translates the experience and science of cancer for others. She was diagnosed with metastatic NSCLC in 2011, learned about biomarker testing and clinical trials from online patient communities, joined a clinical trial, and has been doing well on a targeted therapy for over 11 years. She is a co-founder of The ROS1ders, #LCSM Chat on Twitter (now X), and the IASLC STARS program for training research advocates in lung cancer. She has received the LUNGevity Hero award, coauthored articles in oncology journals, been an invited speaker at national and international cancer conferences, and served on committees and scientific advisory boards for cancer centers, national nonprofits, industry, and government agencies. Formerly an aerospace systems engineer, she holds engineering degrees from MIT (SB) and Caltech (MS, Engineer). Janet blogs at GrayConnections.net.

Detailed Program

Monday, Sep. 16, 2024 - Pre-Meeting Day

12:00-16:00

ITCR Training Network Workshops
(Location: Eiteljorg Museum)
Organized by Carrie Wright and Candace Savonen

16:00-18:00

18:00-21:00

Evening Meet Up: Casual Pub Night
(Location: Fat Roost Diner within the Hyatt Regency Indianapolis, One South Capitol Avenue, Indianapolis)
Dinner will be available

Tuesday, Sep. 17, 2024 - Day 1

06:00-07:00

MORNING RUN *(Meeting at the JW entrance)*
Organized by Prof. Yunlong Liu, Sally Atcheson, and Nick Green
This activity is subject to attendees registering for the activity by Monday 9/16 noon. If interested please indicate your interest by emailing us at comppath@iu.edu

07:00-08:00

08:00-09:00

BREAKFAST
(Location: Eiteljorg Museum)

09:00-09:45

WELCOME / OPENING SESSION
(Location: Eiteljorg Museum)
Spyridon Bakas, Kun Huang, Juli Klemm

09:45-10:45

KEYNOTE 1
(Location: Eiteljorg Museum)
Moderator: Juli Klemm

Debra JH Mathews, PhD, MA
Johns Hopkins University School of Medicine - Ethics & Governance Lead
“Ethical, societal, & governance implications of autonomous systems”

10:45-11:00

COFFEE BREAK
(Location: Eiteljorg Museum)

11:00-12:00

PATIENT ADVOCACY PANEL: “Connecting Researchers and Patient Advocates: Why and How”
(Location: Eiteljorg Museum)
Moderator: Juli Klemm
Dana Cattani
Janet Freeman-Daily
Susan D. Gubar
Terry Whitt Bailey

Detailed Program

Tuesday, Sep. 17, 2024 - Day 1

12:00-13:00

LUNCH
(Location: Eiteljorg Museum)

13:00-14:00

ORAL SESSION 1 (IMAGING)
(Location: Eiteljorg Museum)
Moderator: Prateek Prasanna

Development of an automated method for replicating expert performance in image-guided MRS planning in brain tumors (ID: 36)
Patrick Bolan, Sangyoon Lee, Francesca Branzoli, Thanh Nguyen, Ovidiu Andronesi, Clark Chen, Alexander Lin, Roberto Liserre, Gerd Melkus, James Hodges, Yu-Hui Huang, **Malgorzata Marjanska**

Improving Model Generalization for Neuroendocrine Tumor Identification in PET Images (ID: 56)
Fuyong Xing, Xinyi Yang, Michael Silosky, Jonathan Wehrend, Daniel Litwiller, Debashis Ghosh, Bennett Chin

High-Dimensional Deep Learning of Recurrent vs Non-Recurrent Lung Cancer with Mass Spectrometry Imaging (ID: 45)
Chau Tran, Yik Siu, Manor Askenazi, Wenke Liu, Tianxiao Zhao, Lia Ficaro, Harvey Pass, David Fenyo, Drew Jones

RadxTools for assessing tumor treatment response on imaging (ID: 80)
Satish Viswanath

14:00-14:45

ORAL SESSION 2 (SPATIAL)
(Location: Eiteljorg Museum)
Moderator: Travis S Johnson

Data-driven analyses of immunophenotyping data using ImmunoPheno (ID: 23)
Lincoln Wu, Sriya Potluri, Zhangliang Yang, **Pablo Gonzalez Camara**

Single Cell Visualizations and Analyses on UCSC Xena (ID: 47)
Mary Goldman, Brian Craft, **Jing Zhu**

Visual Analytics for Exploration and Hypothesis Generation Using Highly Multiplexed Spatial Data of Tissues and Tumors (ID: 81)
Jeremy Muhlich

14:45-15:15

COFFEE BREAK
(Location: Eiteljorg Museum)

15:15-16:00

ORAL SESSION 3 (-OMICS)
(Location: Eiteljorg Museum)
Moderator: Kristen M Naegle

Tuesday, Sep. 17, 2024 - Day 1

Accelerating the expert-crowdsourcing of cancer variant interpretation in CIViC (ID: 30)
Obi Griffith, Kilannin Krysiak, Arpad Danos, Jason Saliba, Joshua McMichael, Adam Coffman, Susanna Kiwala, Cameron Grisdale, Caralyn Reisle, Mariam Khanfar, Steven Jones, Alex Wagner, Malachi Griffith

Customize your variant interpretation workflow with OpenCRAVAT (ID: 67)
Jasmine Baker, Kyle Moad, Madison Larsen, Kyle Anderson, Supra Gajjala, **Rachel Karchin**

Using Large Language Models to Make Galaxy more Useful (ID: 24)
Junhao Qui, **Jeremy Goecks**

16:00-17:00

NEW AWARD LIGHTNING TALKS
(Location: Eiteljorg Museum)
Moderator: Spyridon Bakas

The immunoPeptidoGenomic (iPepGen) informatics resource for immuno-oncology research (ID: 25)
Subina Mehta, Reid Wagner, Fengchao Yu, Alexey Nesvizhskii, Pratik Jagtap, **Tim Griffin**

Computational framework for inference of genetic ancestry from cancer-derived molecular data (ID: 26)
Pascal Belleau, Astrid Deschênes, Laine Marrah, David Tuveson, **Alexander Krasnitz**

Estimating the Distribution of Ratio of Paired Event Times in Phase II Oncology Trials (ID: 79)
Li Chen, Mark Burkard, Jianrong Wu, Jill Kolesar, **Chi Wang**

Methods for characterizing mechanobiology of the tumor microenvironment landscape (ID: 83)
Shikhar Uttam

A Multilevel Data Analytic Solution to Advance Population Cancer Research (ID: 84)
Johnnie Rose

Enhancement and further development of informatics methods for long-read cancer sequencing (ID: 85)
Katie Campbell

Informing mechanistic rules of agent-based models with single-cell multi-omics (ID: 86)
Paul Macklin, Elana Fertig

Detailed Program

Tuesday, Sep. 17, 2024 - Day 1

Structure-guided cancer immunotherapy design with HLA-Arena and CrossDome (ID: 89)
Martiela Vaz de Freitas
Cancer Genomics: Integrative and Scalable Solutions in R/Bioconductor (ID: 82)
Sean Davis

Integrative Analysis and Visualization Platform for Cancer Regulatory Genomics (ID:90)
Zhiping Weng

17:00-19:00 **POSTER SESSION**
(Location: Indiana State Museum)
Note: *Abstracts listed below do not represent the complete list of posters included in this session, but only those with just a poster presentation form. The poster session will include posters from all submitted abstracts, including those presented as a talk."*

Expanding Genetic Test Result Delivery with a Hybrid Rule-Based/Large Language Model Chatbot for Return of Positive Results (ID: 4)
Emma Coen, Guilherme Del Fiol, Kim Kaphingst, Caitlin Allen

Developing a Statistical and Visualization Tool for Cancer Registries to Detect Cancer Hot Spots for Small Geographic Areas (ID: 11)
Jacob Oleson

Topological uncertainty for vascular segmentation (ID: 13)
Saumya Gupta, Prateek Prasanna, Chao Chen

EMERSE (the text processing/searching tool) updates for 2024 (ID: 15)
David Hanauer, Lisa Ferguson, Kellen McClain, Guan Wang

CancerModels.Org - an open global cancer research platform for patient-derived cancer models. (ID: 18)
Zinaida Perova, Mauricio Martinez, Tushar Mandloi, Marcelo Rios Almanza, Steven Neuhauser, Dale Begley, Debra Krupke, Carol Bult, Helen Parkinson

Joint estimation of signatures across mutation modalities using multi-modal non-negative matrix factorization (ID: 22)
Kelly Geyer, Masanao Yajima, Jonathan Huggins, Joshua Campbell

Multiomics2Targets: Computational Workflow to Identify Targets for Cancer Cohorts Profiled with Transcriptomics, Proteomics, and Phosphoproteomics (ID: 32)
Giacomo Marino, Eden Deng, Daniel Clarke, Ido Diamant, Avi Ma'ayan

Data Explorer 2.0: a more powerful tool to interrogate the Cancer Dependency Map (ID: 35)

Tuesday, Sep. 17, 2024 - Day 1

Joshua Dempster, Randy Creasi, Yvonne Blanco, Barbara De Kegel, Mustafa Kocak, Francisca Vazquez, Philip Montgomery, Caterina Campbell
Updates from The Cancer Proteome Atlas: a new platform for animal model data and the continued chatbot development (ID: 50)
Jun Li, Wei Liu, Yitao Tang, Yiling Lu, Han Liang

Uncovering Druggable Vulnerabilities in Cancer with the AVERON Notebook (ID: 51)
Hongyue (Nicole) Chen, Brian Revennaugh, Haian Fu, Andrey Ivanov

Inclusion of new clinically actionable variant types into the CIViC data model (ID: 52)
Arpad Danos, Kilannin Krysiak, Jason Saliba, Adam Coffman, Susanna Kiwala, Joshua McMichael, Mariam Khanfar, Cameron Grisdale, Malachi Griffith, Obi Griffith

Integrating multi-omics analyses into agent-based models with the Bioinformatics Walkthrough to predict evolution of pancreatic ductal adenocarcinoma (ID: 55)
Daniel Bergman, Paul Macklin, Elana Fertig

Extracting Social Determinants of Health from Pediatric Patient Notes Using Large Language Models: Novel Corpus and Methods (ID: 57)
Yujuan Fu, Giridhar Ramachandran, Nicholas Dobbins, Namu Park, Michael Leu, Abby Rosenberg, Kevin Lybarger, Fei Xia, Ozlem Uzuner, Meliha Yetisgen

Developing a Rule-Based Algorithm to Identify Recurrent Non-Hodgkin Lymphoma in Electronic Health Data (ID: 58)
Mara Epstein, Feifan Liu, Laura Susick, Yanhua Zhou, Shane Bole, Lydia Goldthwait, Wendy Haykus, Muthalagu Ramanathan, George Divine, Christine Johnson

Preclinical Imaging XNAT-Enabled Informatics (PIXI): An open-source resource to support cloud-based computational workflows for preclinical imaging (ID: 59)
Andrew W. Lassiter, Stephen M. Moore, James D. Quirk, Richard Laforest, William Horton, Daniel S. Marcus, Kooresh I. Shoghi

Enhanced MSBooster for Sensitive HLA Peptide Identification (ID: 60)
Fengchao Yu, Kevin Yang, Pratik Jagtap, Reid Wagner, Timothy Griffin, Alexey Nesvizhskii

Discovery of Novel CDK Inhibitors with ARCHS4/RummaGEO and the LINCS L1000 Datasets (ID: 61)
John Erol Evangelista, Alexander Lachmann, Daniel Clarke, Avi Ma'ayan

Introduction to Bioinfor-omics: Online Course about the Application of Bioinformatics Methods to Multi-Omics Datasets (ID: 62)
Heesu Kim, Daniel Clarke, Giacomo Marino, Zhuorui Xie, Alexander Lachmann, Ido Diamant, John Erol Evangelista, Eden Deng, Stephanie Olaiya, Sherry Jenkins,

Tuesday, Sep. 17, 2024 - Day 1

Avi Ma'ayan

Topological Features for Histopathology Modeling (ID: 63)
Meilong Xu, Prateek Prasanna, Chao Chen

Harmonizome 3.0: Integrating Knowledge about Genes and Proteins from Diverse Multi-Omics Resources (ID: 64)
Ido Diamant, Daniel Clarke, Nathania Lingam, Avi Ma'ayan

CARS-TEA: Utilizing ARCHS4 to Enable Enrichment Analysis at the Transcript Level (ID: 65)
Anna Byrd, Giacomo Marino, Avi Ma'ayan

Using ARCSH4 for scRNA-seq Imputation and Cell Type Identification (ID: 66)
Nasheath Ahmed, Giacomo Marino, Billal Ali, Sophie Gideon, Avi Ma'ayan

A novel transformer-based deep learning model for predicting binding interactions between HLA class I molecules and peptides (ID: 68)
Kun Hee Kim, Xianli Jiang, Yukun Tan, Jae Jun Ku, Shaoheng Liang, Maura Gillison, Ken Chen

The DepMap Portal: Enabling explorations of cancer cell lines and dependencies (ID: 69)
Ali Mourey, Jessica Cheng, Nayeem Aquib, Philip Montgomery, Randy Creasi, Sarah Whitaker, Josh Dempster, Lauren Golden, Yvonne Blanco, Katie Campbell, Francisca Vazquez

Metabolomics Outcomes as a Function of Biological Complexity from Colorectal Cancer-Related Microbiome Samples (ID: 71)
Jennifer Nguyen, Joseph Krampen, Jungmoo Huh, Kathryn McBride, Patrick Schloss, Marcy Balunas

Investigating epithelial-mesenchymal transition and endothelial-mesenchymal transition in the tumor microenvironment of HPV+ head and neck cancer (ID: 72)
Catherine Zhou

Overture: An Open-Source Genomics Data Platform (ID: 73)
Christina Yung, Mitchell Shiell, Jon Eubank, Justin Richardsson, Brandon Chan, Robin Haw, Lincoln Stein, Melanie Courtot, Overture Team

Towards an automated expert knowledge base reviewer using natural language processing (ID: 74)
Caralyn Reisle, Cameron J. Grisdale, Kilannin Krysiak, Arpad M. Danos, Mariam Khanfar, Erin Pleasance, Jason Saliba, Melika Bonakdar, Malichi Griffith, Obi L. Griffith, Steven J.M. Jones

Image Embeddings to Reduce Image Feature Variance (ID: 75)

Tuesday, Sep. 17, 2024 - Day 1

Torop Max, Jennifer Dy, Veronica Rotemberg, Kivanc Kose

The Integrative Genomics Viewer (IGV) for Cancer Research (ID: 76)
James Robinson, Helga Thorvaldsdottir, Jill Mesirov

WebMev: leveraging cloud and containerization tools for performant, open platform for exploratory omics research (ID: 77)
Derrick DeConti, Brian Lawney, Ilya Sytchev, Saron Nhong, John Quackenbush

The GenePattern ecosystem for cancer bioinformatics (ID: 78)
Michael Reich, Thorin Tabor, Ted Liefeld, Edwin Huang, Forrest Kim, Helga Thorvaldsdottir, Jill Mesirov

Determining tissue-independent N6-methyladenosine (m6A) epitranscriptome and its regulatory role in cancer (ID: 88)
Sumin Jo, Ting-He Zhang, Wen Meng, Jianqiu Zhang, Shou-Jiang Gao, Yufei Huang

RummagenexRummaGEO: Crossing the Rummagene and RummaGEO Gene Sets for Novel Hypotheses Generation (ID: 91)
Eugenia Ampofo

19:00-21:00 **DINNER @ THE MUSEUM**
(Location: Indiana State Museum)

Detailed Program

Wednesday, Sep. 18, 2024 - Day 2

06:00-07:00

MORNING RUN *(Meeting at the JW entrance)*
Organized by Prof. Yunlong Liu, Sally Atcheson, and Nick Green
This activity is subject to attendees registering for the activity by Monday 9/16 noon. If interested please indicate your interest by emailing us at comppath@iu.edu

07:00-08:00

08:00-08:30

BREAKFAST (Optional meeting of the OPEN Working Group. All are invited)
(Location: Eiteljorg Museum)

08:30-09:15

ORAL SESSION 4 (IMAGING)
(Location: Eiteljorg Museum)
Moderator: Prateek Prasanna

Preclinical Imaging XNAT-Enabled Informatics (PIXI): An open-source resource to support cloud-based computational workflows for preclinical imaging (ID: 48)
Kooresh Shoghi, Andrew Lassiter, Stephen Moore, James Quirk, Richard Laforest, William Horton, Daniel Marcus

Advancing Medical Image Visualization: Integrating Polymorph Segmentation in Cornerstone3D for Enhanced OHIF Viewer Capabilities (ID: 21)
Gordon Harris, Alireza Sedghi, James Hanks, Dan Rukas, Rob Lewis, Chris Hafey, Trinity Urban, Erik Ziegler

Differential Privacy for Privacy-aware AI in Computational Pathology: Tool or Toy? (ID:37)
Sarthak Pati, Spyridon Bakas

09:15-10:00

ORAL SESSION 5 (SPATIAL)
(Location: Eiteljorg Museum)
Moderator: Travis S Johnson

Democratizing spatial transcriptomics analysis with spatialGE (ID: 2)
Oscar Ospina, Roberto Manjarres-Betancur, Guillermo Gonzalez-Calderon, Alex Soupir, Inna Smalley, Kenneth Tsai, Xiaoqing Yu, Brooke Fridley

Engineering model-based systems to monitor and steer subclonal dynamics (ID: 16)
Thomas Veith, Saeed Alahmari, Vural Tagal, Richard Beck, Issam El Naqa, **Noemi Andor**

Quantifying spatial tumor heterogeneity (ID: 41)
Cong Ma, Uthsav Chitra, **Benjamin Raphael**

10:00-10:15

SHORT COFFEE BREAK
(Location: Eiteljorg Museum)

Detailed Program

Wednesday, Sep. 18, 2024 - Day 2

10:15-11:15

KEYNOTE 2
(Location: Eiteljorg Museum)
Moderator: Spyridon Bakas

Michael D Feldman, MD, PhD
Indiana University School of Medicine - Chairman of Pathology & Laboratory Medicine
"Where Computational Analytics Meet Clinical Requirements"

11:15-12:00

LIGHTNING TRAINEE SESSION 1
(Location: Eiteljorg Museum)
Moderator: Kooresh Shoghi

The three-dimensional structure of extrachromosomal DNA reveals novel conformation changes (ID: 7)
Biswanath Chowdhury, Kaiyuan Zhu, **Chaohui Li**, Jens Luebeck, Katerina Kraft, Shu Zhang, Lukas Chavez, Paul S. Mischel, Howard Y. Chang, Vineet Bafna

RummaGEO: Automatic Mining of Human and Mouse Gene Sets from GEO (ID: 20)
Giacomo Marino, Daniel Clarke, Eden Deng, Avi Ma'ayan

Strategic Patch-based Deep Learning Workflow for Accurate Breast Cancer Cell Classification from Microscopy Images (ID: 27)
Harrison Yee, Kailie Matteson, Joshua Goldwag, John Lamar, Margarida Barroso, Xavier Intes, Uwe Kruger

OmicsMLRepo: Ontology-leveraged metadata harmonization to improve AI/ML-readiness of omics data in Bioconductor (ID: 33)
Sehyun Oh, Kaelyn Long, Kai Gravel-Pucillo, Levi Waldron, Sean Davis

Playbook Workflow Builder: Interactive Construction of Bioinformatics Workflows from a Network of Microservices (ID: 39)
Daniel Clarke, Avi Ma'ayan

Immunotherapy Treatment Outcome Prediction in Small Cell Lung Cancer (SCLC) through Topo-Geometric Characterization of Pulmonary Artery Vasculature: A proof-of-concept study (ID: 54)
Moinak Bhattacharya, Jiachen Yao, Shirish M Gadgeel, Chao Chen, Prateek Prasanna

Identification of High-Risk Cells in Spatially Resolved Transcriptomics of Cancer Biopsies Using Deep Transfer Learning (ID: 70)
Debolina Chatterjee, Justin Couetil, Tianhan Dong, Jie Zhang, Kun Huang, Chao Chen, Travis Johnson

12:00-13:00

LUNCH
(Location: Eiteljorg Museum)

Detailed Program

Wednesday, Sep. 18, 2024 - Day 2

13:00-13:15

VIDEO CONTEST
(Location: Eiteljorg Museum)
Candace L Savonen, Carrie Wright

13:15-14:45

SUSTAINABILITY SESSION
(Location: Eiteljorg Museum)
Session Overview & Moderator: Juli Klemm
Project overviews by: Ilija Dukovski (COMETS), Jacob Oleson (Small Area Risk Maps), Rachel Karchin (Open CRAVAT), Gordon Harris (OHIF), David Hanauer (EMERSE)
Facilitated Group Discussions
Rapid Group Report-Outs

14:45-15:15

COFFEE BREAK
(Location: Eiteljorg Museum)

15:15-16:00

ORAL SESSION 6 (-OMICS)
(Location: Eiteljorg Museum)
Moderator: Kristen M Naegle

Fifteen Years of cBioPortal for Cancer Genomics (ID: 42)
*Ino de Bruij*n, Tali Mazor, Rima AlHamad, Calla Chennault, Corey Dubin, Jeremy Easton-Marks, Zhaoyuan Fu, Benjamin Gross, Charles Haynes, David M Higgins, Jason Hwee, Jagannathan K Prasanna, Mirella Kalafati, Karthik Kalletla, James Ko, Tim Kuijpers, Sowmiyaa Kumar, Priti Kumari, Ritika Kundra, Bryan Lai, Xiang Li, James Lindsay, Aaron Lisman, Qi-Xuan Lu, Ramyasree Madupuri, Angelica Ochoa, Yusuf Ziya Özgül, Oleguer Plantalech, Matthijs Pon, Baby A Satravada, Jessica Singh, S Onur Sumer, Pim van Nierop, Floris Vleugels, Avery Wang, Manda Wilson, Hongxin Zhang, Gaofei Zhao, Ugur Dogrusoz, Allison Heath, Adam Resnick, Trevor J Pugh, Chris Sander, Ethan Cerami, Jianjiong Gao, Nikolaus Schultz

Pathway-guided Feature Selection and Integration for Cancer Subtyping (ID: 17)
Ha Nguyen, Dung Pham, Hung Nguyen, Dao Tran, *Tin Nguyen*

Informatics tools for neoantigen characterization and therapeutic translation (ID: 31)
Susanna Kiwala, Huiming Xia, My Hoang, Kartik Singhal, Evelyn Schmidt, Mariam Khanfar, Joshua McMichael, Jasreet Hundal, Thomas Mooney, Jason Walker, S. Peter Goedegebuure, Christopher Miller, Todd Fehniger, Robert Schreiber, William Gillanders, Obi Griffith, *Malachi Griffith*

16:00-17:00

ORAL SESSION 7 (CLINICAL)
(Location: Eiteljorg Museum)
Moderator: Veronica Rotemberg
Distributed multiple imputation for correlated incomplete data based on federated generalized linear mixed model (ID: 40)
Yi Lian, **Xiaoqian Jiang**, Qi Long

Wednesday, Sep. 18, 2024 - Day 2

Matching Genotypes with Personalized Therapies: Development of a Decision Support Infrastructure to Augment the Value of Precision Medicine (ID: 9)
Taxiarchis Botsis, Kory Kreimeyer, Jonathan Spiker, Maria Fatteh, Jamie Wehr, Mimi Najar, Jessica Tao, Nicole Imamovic, Ander Pindzola, Rena Xian, Adrian Dobs, Jenna Canzoniero, Valsamo Anagnostou

GARDE: Scalable Clinical Decision Support for Individualized Cancer Risk Management (ID: 3)
Guilherme Del Fiol, Caitlin Allen, Ravi Sharaf, Richard Bradshaw, Muhammad Danyal Ahsan, Emerson Borsatto, Elena Elkin, Melissa Frey, Kevin Hughes, Wendy Kohlmann, Polina Kukhareva, Anne Madeo, Che Martin, Chelsey Schlechter, Kimberly Kaphingst, Kensaku Kawamoto

Demonstrating the Value of DeepPhe for Translational studies in Breast/Ovarian Cancer and Melanoma
Alexander van Helene, Harry Hochheiser, Jiarui Yao, Eli Goldner, Sean Finan, John Levander, Dennis Johns, David Harris, Piet de Groen, Elizabeth Buchbinder, Danielle Bitterman, Jeremy Warner, Guergana Savova

17:00-18:00

17:30 SHUTTLE SERVICE TO RECEPTION
(Meeting at the JW Entrance)

19:00-21:00 RECEPTION @ SURPRISE EVENT
Location: Surprise!

Thursday, Sep. 19, 2024 - Day 3

06:00-07:00 MORNING RUN *(Meeting at the JW entrance)*
Organized by Prof. Yunlong Liu, Sally Atcheson, and Nick Green
This activity is subject to attendees registering for the activity by Monday 9/16 noon. If interested please indicate your interest by emailing us at comppath@iu.edu

07:00-08:00

08:00-08:30 BREAKFAST
(Location: Eiteljorg Museum)

08:30-09:15 ORAL SESSION 8 (CLINICAL)
(Location: Eiteljorg Museum)
Moderator: Veronica Rotemberg

GEARBOx: automated, patient-centric clinical trials matching (ID: 5)
Luca Graglia, Brian Furner, Jooho Lee, Lauren Chan, Steve Krasinsky, Tomasz Oliwa, Enal Hindi, Michael Watkins, Kirk Wyatt, Samuel Volchenboum

Thursday, Sep. 19, 2024 - Day 3

Evaluating a Novel Algorithm to Process Electronic Adherence Monitoring Device Data (ID: 12)
Meghan McGrady, Kevin Hommel, Constance Mara, Michal Kouril

Facilitating androgen deprivation therapy treatment discussions between prostate cancer patients and their physicians via a comprehensive prognosis model that outputs personalized treatment benefit estimates based on cancer-related, genetic, and non-cancer risk factors (ID: 43)
Jessica Aldous, Matthew Schipper, Ralph Jiang, Robert Dess, Krithika Suresh, Elizabeth Chase, William Jackson

09:15-10:00 **ORAL SESSION 9 (MOLECULAR)**
(Location: Eiteljorg Museum)
Moderator: Carrie Wright

Inferring Kinase Activity from Tumor Phosphoproteomic Data (ID: 10)
Sam Crowl, Candace Lei, Gabriela Salazar Lopez, Joseph-Levi Custer, **Kristen Naegle**

Lancet2: improved performance and genotyping of somatic variants using localized genome graphs (ID: 14)
Rajeeva Musunuri, Bryan Zhu, Wayne Clarke, Timothy Chu, Jennifer Shelton, Dickson Chung, Shreya Sundar, Adam Novak, Benedict Paten, Nicolas Robine, **Giuseppe Narzisi**

Integrated metabolic and mechanical modeling for spatio-temporal simulations of tumors in their microenvironment (ID: 28)
Ilija Dukovski, Louis Brezin, Kirill Korolev, Daniel Segrè

10:00-10:15 **COFFEE BREAK**
(Location: Eiteljorg Museum)

10:15-11:15 **KEYNOTE**
(Location: Eiteljorg Museum)
Moderator: Spyridon Bakas

Peter Mattson, PhD
MLCommons - President
"Benchmarking, Governance, & Orchestration of Emerging Informatics Technologies"

11:15-12:00 **LIGHTNING TRAINEE SESSION 2**
(Location: Eiteljorg Museum)
Moderator: Kooresh Shoghi

Detailed Program

Thursday, Sep. 19, 2024 - Day 3

AmpliconSuite: Analyzing focal amplifications in cancer genomes (ID: 8)
Jens Luebeck, Edwin Huang, Forrest Kim, Ted Liefeld, Bhargavi Dameracharla, Rohil Ahuja, Kaiyuan Zhu, Soyeon Kim, Hoon Kim, Roel G.W. Verhaak, Michael Reich, Paul S. Mischel, Jill Mesirov, Vineet Bafna

Advancing Precision Oncology: Collaborative Efforts in the Curation and Classification of Somatic Cancer Variants by ClinGen Somatic and CIViC (ID: 34)
Mariam Khanfar, Jason Saliba, Arpad Danos, Kilannin Krysiak, Adam Coffman, Susanna Kiwala, Joshua McMichael, Cameron J Grisdale, Ian King, Shamini Selvarajah, Rashmi Kanagal-Shamanna, Laveniya Satgunaseelan, David Meredith, Madina Sukhanova, Charles G Mullighan, Mark G Evans, Yassmine Akkari, Gordana Raca, Angshumoy Roy, Ramaswamy Govindan, Jake Lever, Alex H Wagner, Obi L Griffith, Malachi Griffith

DataCrossways: A Unified, Scalable Data Management Layer Applied to Enhance the ARCHS4 Resource (ID: 38)
Alexander Lachmann, Avi Ma'ayan

Building Contextual m6A Knowledge Graph in Cancer through Literature Analysis with reguloGPT (ID: 87)
Xidong Wu, **Sumin Jo**, Yiming Zeng, Arun Das, Ting-He Zhang, Parth Patel, Yuanjing Wei, Lei Li, Shou-Jiang Gao, Jianqiu Zhang, Dexter Pratt, Yu-Chiao Chiu, Yufei Huang

pVACview: Visualization Tool for Neoantigen Prioritization (ID: 44)
Evelyn Schmidt, Huiming Xia, My Hoang, Susanna Kiwala, Joshua McMichael, Zachary L. Skidmore, Bryan Fisk, Jonathan J. Song, Jasreet Hundal, Thomas Mooney, Jason R. Walker, S. Peter Goedegebuure, Christopher A. Miller, William E. Gillanders, Obi L. Griffith, Malachi Griffith

Immune validation of neoantigen vaccines through clonal TCR analysis in patients with pancreas cancer (ID: 46)
Kartik Singhal, Felicia Zhang, Xiuli Zhang, S. Peter Goedegebuure, Christopher A. Miller, Gue Su Chang, Jasreet Hundal, John Garza, Mike D. McLellan, William E. Gillanders, Obi L. Griffith, Malachi Griffith

pVACsplice: predicting and prioritizing tumor-specific splicing antigens (ID: 49)
My Hoang, Miller Richters, Susanna Kiwala, Jeffrey P Ward, Ramaswamy Govindan, Obi L Griffith, Malachi Griffith

12:00-12:30 LUNCH BOXES
(Location: Eiteljorg Museum)

12:30-14:00 SPATIAL TRANSCRIPTOMICS WORKSHOP
(Location: Eiteljorg Museum)
Oscar Ospina (in partnership with the ITCR Training Network (ITN))

Abstracts

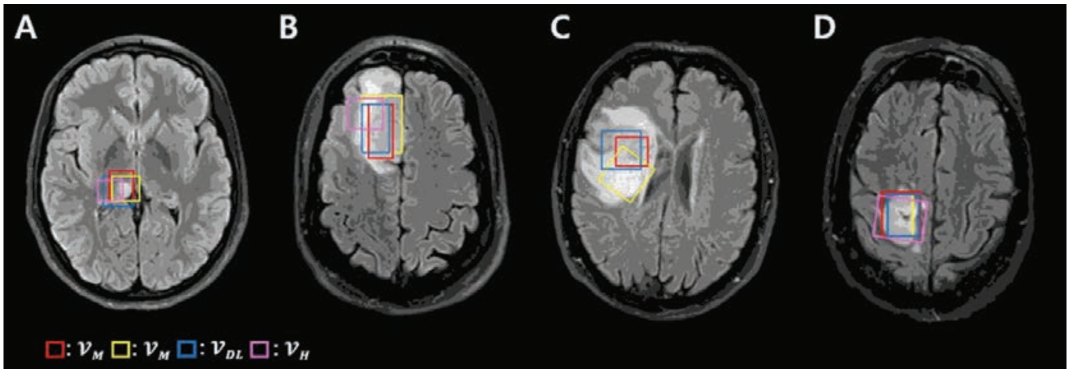
Oral Session 1 - Imaging

Development of an automated method for replicating expert performance in image-guided MRS planning in brain tumors

Authors: Patrick Bolan, Sangyoon Lee, Francesca Branzoli, Thanh Nguyen, Ovidiu Andronesi, Clark Chen, Alexander Lin, Roberto Liserre, Gerd Melkus, James Hodges, Yu-Hui Huang, Malgorzata Marjanska
Submitter: Patrick Bolan
Submitter Email: bola0035@umn.edu

Magnetic resonance spectroscopy (MRS) can provide useful metabolic information for diagnosis, treatment response, and prognosis of brain tumors. Single-voxel MRS measurements can produce quantitative information on up to 20 brain metabolites relevant to tumor type and progression. Integrating this non-invasive method for biochemical assessment with existing imaging modalities has shown promise in enhancing diagnosis and treatment response monitoring. Single-voxel MRS requires precise planning of the acquisition volume to produce a high-quality signal localized in the pathology of interest. Unfortunately, the placement of the acquisition volume is challenging and limit the clinical adoption of the technique: it requires a precise positioning of the acquisition voxel (a 3D cuboid) to optimally sample the tumor while avoiding non-involved tissues (e.g., bone, CSF) and sources of artifacts (e.g., hemorrhage, necrosis), which requires real-time assessment of MRI images and expertise in MRS methodology. Voxel placements vary with tumor size, location, stage, as well as the voxel placer's experience and approach. Specific aim 1 (SA1) of our U01 project is the development of an automated software tool to help standardize the voxel placement process and reduce the dependence on real-time placement by a specially trained expert. The first phase of SA1 was to generate a dataset of brain tumor cases with associated expert voxel placements and precise tumor segmentations. We collected MR images from 125 patients with gliomas, anonymized the image data, and distributed these images to five expert individuals (neuroradiologists or physicists with substantial clinical MRS experience) to perform MRS voxel placement. Each expert performed voxel placements on 25 datasets using a custom software tool intended to simulate the MRS planning procedure on a MR scanner. Experts also provided quality ratings on a 1-3 scale for their own placements, and those of 50 other placements from other experts. We also generated pixelwise three-class semantic segmentation of each image set, classifying each pixel as background, tumor core, whole tumor, and "penalty" regions (areas to avoid with MRS: CSF, necrosis, and cysts). Image segmentation was initialized by training a neural network (based on nnUNet) using the publicly-available BRATs brain tumor dataset, and manually refined by a neuroradiologist. The second phase of SA1 is to train an algorithm to generate voxel placements on unseen, prospectively acquired brain tumor MR images, with performance as similar as possible to that of our expert MRS placers. Using our dataset of brain tumor images with associated voxel placements, quality scores, and pixelwise segmentations, we are building automated software to perform the voxel placements accurately and quickly. The third phase of SA1 is to integrate the automated voxel placement tools with our MR scanner to enable voxel placements in real time, so they can be used immediately for subsequent MRS measurements in the patient.

In this presentation, we will present details of the expert voxel placement dataset and show initial results of the automated voxel placement algorithms.



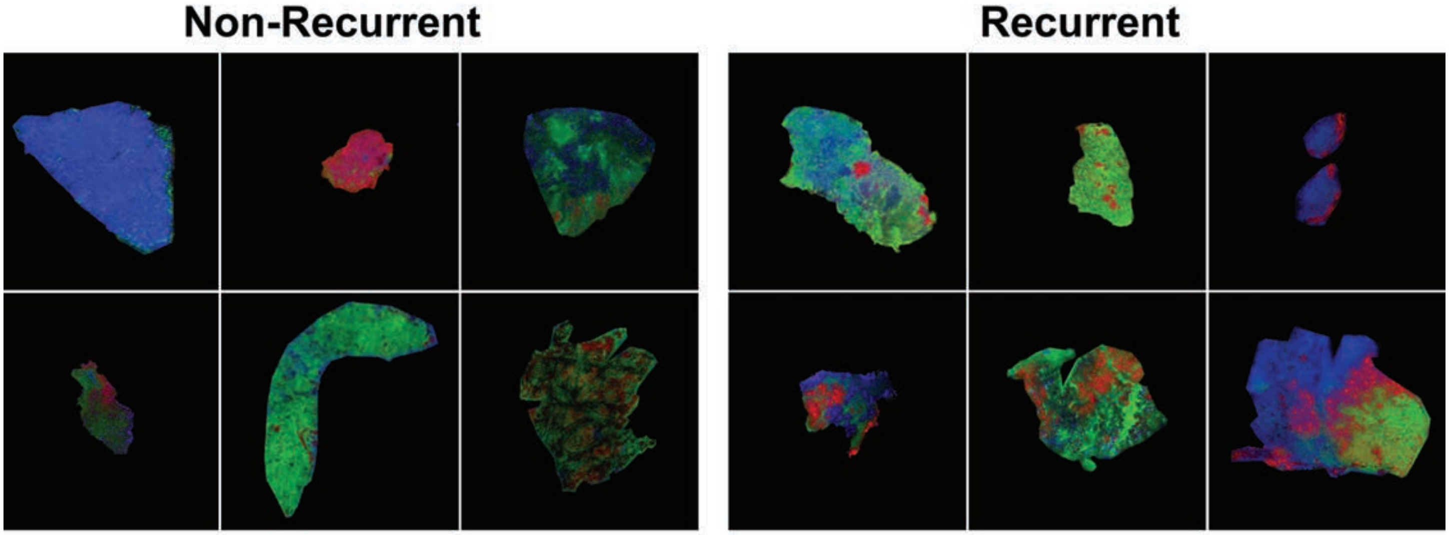
Comparisons of manual expert voxel placements (VM , red and yellow), with two automated voxel placement methods based on deep learning (VDL, blue) and a heuristic objective function (VH, pink) overlaid on T2w-FLAIR images.

Oral Session 1 - Imaging

High-Dimensional Deep Learning of Recurrent vs Non-Recurrent Lung Cancer with Mass Spectrometry Imaging

Authors
Chau Tran, Yik Siu, Manor Askenazi, Wenke Liu, Tianxiao Zhao, Lia Ficaro, Harvey Pass, David Fenyo, Drew Jones
Submitter: Drew Jones
Submitter Email: drew.jones@nyulangone.org

Mass spectrometry imaging (MSI) is a rapidly evolving technique that offers spatial resolution and quantification of metabolites, lipids, peptides, and glycans in snap-frozen tissue cryosections. MSI involves scanning biological tissues with a mass spectrometer to obtain detailed information about the molecular composition at different locations within the sample. The result is a set of data that can be reconstructed into an image, where each pixel represents the mass spectrum of that specific point in the sample. For each scanned tissue, MSI technique can provide thousands of layered images, where each image can be thought of as a separate data channel, providing highly descriptive spatial information about the molecular composition of the tissue in a much higher dimension than traditional optical images of histopathological stains represented in 3-channels (red, green, blue) with typical machine applied to image data. The nature of the dataset presents both challenges and opportunities to use machine learning approaches for classification tasks. Following our previous success in developing Panoptes, a deep learning algorithm that utilizes Convolutional Neural Network (CNN) architecture to predict the histological subtype and molecular subtype using high-resolution images of endometrial cancer tissues, we expand the algorithm's capabilities from 3-channel to n-channel. We applied our modified Panoptes-MSI CNN to study the technique's efficacy in predicting cancer recurrence compared to non-recurrent subjects in a human cohort of lung adenocarcinoma patients. Our dataset consists ~1TB of raw data across 24 histopathological sections containing both regions of cancer cells and normal tissue. We separated these data equally into training and testing datasets using a tiling approach and random assignment of the tiles. In this work, we discuss the key modifications to the architecture of Panoptes, and the performance trade-offs between using high resolution and high dimension sets of images as input. We assessed learning based on the model residuals over iterations with respect to tile size, multiple resolutions, and number of data-channels corresponding to "single molecule images". We further evaluated the impact of data pre-processing parameters and developed an algorithm to extract features based on their signal-to-noise ratio on a per-pixel basis to ensure that training was focused on high information content data. Continuing work is focused on extraction of discriminating model features, and on the alignment of other imaging modalities external to MSI to enable training on multi-modal cancer-imaging data.



Principal Components Summary Image of Human Recurrent vs Non-Recurrent Lung Adenocarcinoma

Abstracts

Oral Session 1 - Imaging

Improving Model Generalization for Neuroendocrine Tumor Identification in PET Images

Authors: Fuyong Xing, Xinyi Yang, Michael Silosky, Jonathan Wehrend, Daniel Litwiller, Debashis Ghosh, Bennett Chin
Submitter: Fuyong Xing
Submitter Email: fuyong.xing@cuanschutz.edu

Background: Gastroenteropancreatic neuroendocrine tumors (GEP-NETs) are difficult-to-detect tumors which commonly present at advanced stages. 68Ga and 64Cu DOTATATE positron emission tomography-computed tomography (PET/CT) are the most sensitive methods to identify somatostatin receptor subtype 2 positive GEP-NETs (majority of GEP-NETs). Currently, no standardized methods exist to quantify residual tumor burden to assess response to therapy. There is an unmet medical need of a robust, automated lesion identification system to provide uniform and objective quantification of residual tumor burden for NETs with PET imaging.

Deep neural networks have been recently applied to lesion identification in PET images, but they typically rely on a large amount of well-annotated data for model training. This is extremely difficult to achieve for NETs, because of low incidence of NETs and expensive lesion annotation with PET imaging. In addition, current deep models often do not consider domain shifts, e.g., testing data has a different distribution from training data. This leads to significant performance degradation when applying models to cross-site/-scanner data.

Methods: We design a novel deep learning-based imaging informatics system for automated, generalizable lesion detection in DOTATATE PET images. Specifically, we develop an adversarial domain generalization module that relies on patch-based gradient reversal learning and is generalizable to unseen resource PET image data. We also develop a new domain adaptation module that uses a region-guided generative adversarial network (GAN) for adaptive lesion detection in specific target datasets.

Results: Using list mode-simulated PET images as training data, our framework significantly outperforms the counterpart without domain generalization/adaptation in real clinical 68Ga-DOTATATE PET images. In addition, the framework produces very competitive performance with target models that are trained with real-world PET image.

Conclusion: With domain generalization and adaptation, deep learning models can be applied to cross-site/-scanner PET imaging data, significantly reducing human effort for data annotation and improving model generalizability for lesion identification in PET images. Thus, it will reduce the cost of applying deep learning algorithms to PET image quantification.

Abstracts

Oral Session 1 - Imaging

RadxTools for assessing tumor treatment response on imaging

Authors: Satish Viswanath
Submitter: Satish Viswanath
Submitter Email: satish.viswanath@case.edu

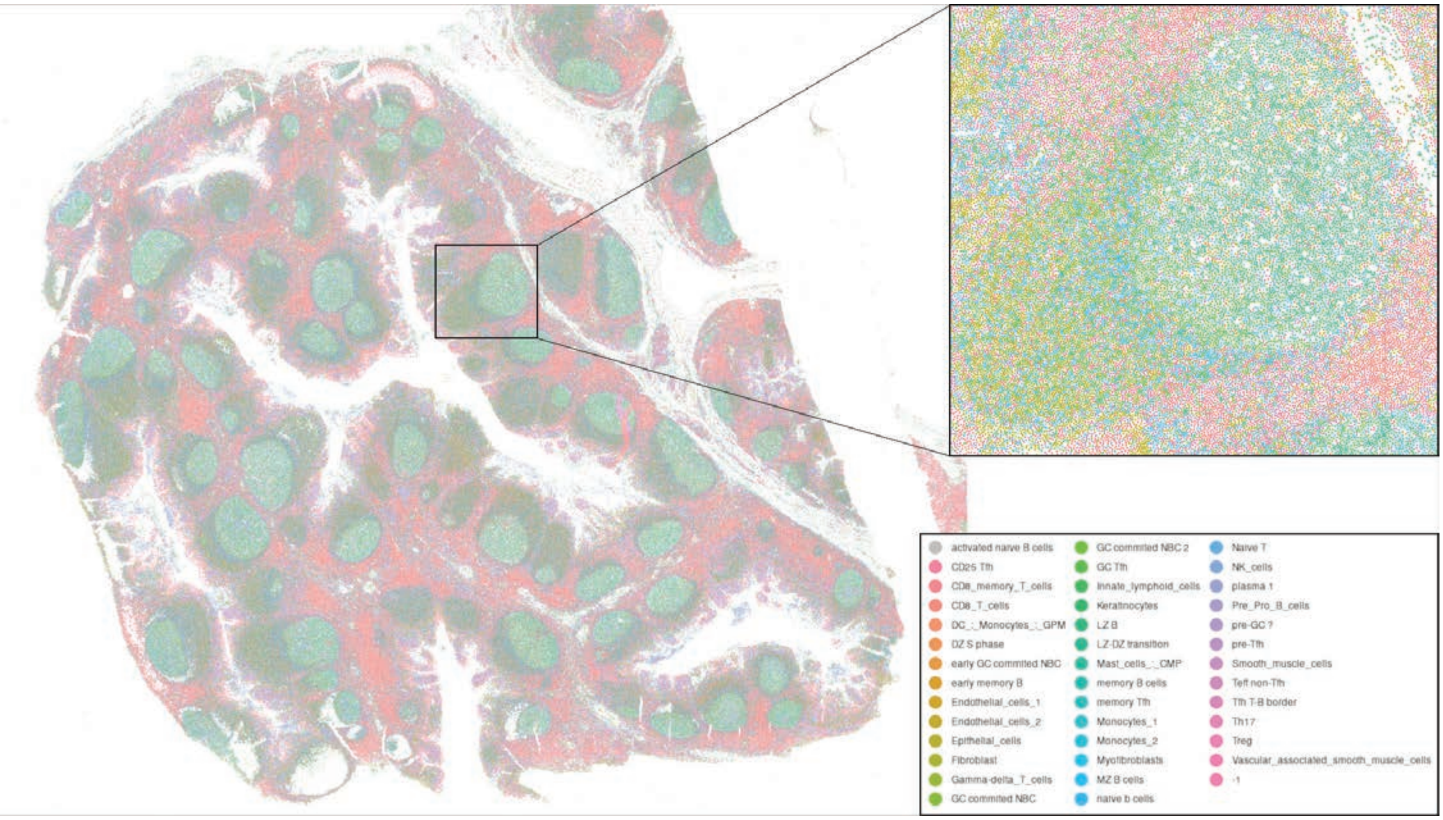
Over 1.6 million patients in the U.S. annually undergo chemo- or radiation- as first-line cancer therapy. After therapy, the most significant challenge for oncologists is identifying non-responders (those with residual or progressive disease), which could allow them to be switched to alternative therapies. Similarly, if those with stable or regressing disease were identified early and reliably, patients could avoid unnecessary and highly morbid surgeries or biopsies for disease confirmation. Unfortunately, expert assessment of post-treatment imaging is challenging, as residual disease is visually confounded with benign treatment-induced changes on imaging. To address these challenges, we will present findings generated through development of RadxTools, a new image informatics toolkit to characterize and quantify treatment response in oncology via routine imaging. This comprises three modules: (a) RadQC to enable quality control of radiomics features across multi-site imaging cohorts, (b) RadTx comprising new radiomics tools which capture local surface morphometric changes and subtle structural deformations unique to tumor response on post-treatment imaging, and (c) RadPathFuse for creating deeply annotated learning sets by spatially mapping post-treatment changes from ex vivo surgically excised histopathology specimens onto pre-operative in vivo imaging. RadxTools has been evaluated in the context of post-treatment characterization for two use cases: distinguishing (a) radiation effects from cancer recurrence for brain tumors; and (b) complete/ partial vs incomplete chemoradiation response for rectal cancers. In addition to showcasing research findings of successfully evaluating RadxTools, we will demonstrate our tool prototypes (integrated into 3D Slicer and CapTk), including Jupyter notebook workflows. We will also present a first-of-its-kind integration of RadxTools for analyzing radiographic data available in the Imaging Data Commons.

Oral Session 2 - Spatial

Data-driven analyses of immunophenotyping data using ImmunoPheno

Authors: Lincoln Wu, Sriya Potluri, Zhangliang Yang, Pablo Gonzalez Camara
Submitter: Pablo Gonzalez Camara
Submitter Email: pcamara@pennmedicine.upenn.edu

Tumor progression, resistance to therapy, and metastasis are closely related to the characteristics of the tumor cell ecosystem. Multiplexed antibody-based cytometry, or immunophenotyping, is the standard method for phenotypic characterization of tissue composition, pathogenesis, and immune infiltration with single-cell (and sometimes spatial) resolution. Traditionally, the identification of cell populations in these data has been facilitated by algorithms that cluster cells according to their antigenic profile and by predefined sets of markers that have historically evolved through trial and error. However, the annotation of these data is a manual, subjective, and laborious process that hinders the reproducibility and accuracy of the results. The design of antibody panels that include specific markers for all cell types and states present in a tissue is usually unfeasible, and the efficiency of commonly used markers is unknown. To overcome these limitations, we are developing informatics technologies that leverage single-cell transcriptomic atlases to assist and automate the design and analysis of multiplexed antibody-based cytometry experiments. ImmunoPheno is a Python library and online resource that automates the identification and annotation of immune cell populations in cancer cytometry data based on harmonized reference single-cell proteo-transcriptomic data. It enables the normalization and harmonization of cytometry data produced by a broad range of technologies, the automated annotation and detection of subtle cell populations in these data, and the design of optimal antibody panels and gating strategies for isolating specific cell populations. The reference data is hosted in an online database and web portal that can be useful in the design of new cytometry experiments. Overall, we expect that ImmunoPheno will boost the phenotypic resolution, accuracy, and reproducibility of multiplexed antibody-based cytometry data analyses in cancer research.



Automated annotation of cell identities in human tonsil tissue section using ImmunoPheno. The tissue section was profiled with PhenoCycler using a panel of 26 antibodies.

Abstracts

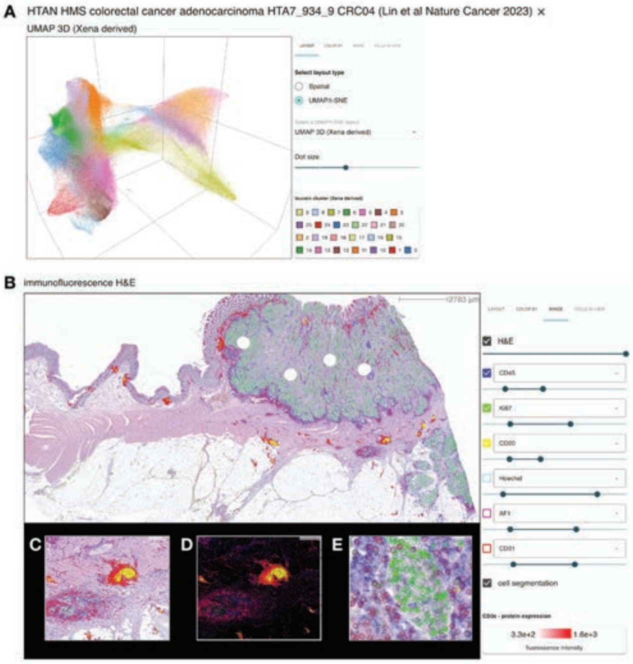
Oral Session 2 - Spatial

Single Cell Visualizations and Analyses on UCSC Xena

Authors: Mary Goldman, Brian Craft, Jing Zhu
Submitter: Jingchun Zhu
Submitter Email: jzhu@soe.ucsc.edu

Advancements in single-cell genomics technologies have revolutionized our understanding of cellular and subcellular behavior, greatly advancing cancer research. Despite its importance, there is a scarcity of user-friendly visualization tools for single-cell data, especially those accessible via the web. Building upon the UCSC Xena platform, we have developed the Xena Single Cell Viewer, an interactive web-browser-based visualization for single cell data. Our viewer supports commonly generated single-cell data types including scRNAseq embedding layouts in either 2D or 3D, as well as large biomedical images, such as miF, t-CyCIF, and CODEX. These images can be up to centimeter-sized specimens imaged at sub-micron resolution across multiplex spectral channels. Our Single Cell Viewer is fast and responsive with interactive zooming and panning even for very large images that are hundreds of gigabytes in size. We allow users to dynamically manipulate individual image layers and channels, including toggling co-registered H&E images as well as adding cell segmentation overlays, adjusting an individual channel's color, saturation and minimum color threshold. Importantly, for both embedding and image views, we enable users to interactively overlay genomics data such as gene-, protein-expression, cell types, and cell scores on top of these images. Users can also view expression from two genes or two proteins, as well as two predicted cell types. Behind these views is a new highly performant Xena data hub specifically for serving these very large single-cell datasets across the web. The HTAN data portal currently links out to our new Single Cell Viewer, enabling researchers to view HTAN data on UCSC Xena.

More recently we have begun to develop preliminary visualizations of the results of data integration and data transfer methods from published computational methods. Integrating spatially resolved data with single-cell profiles presents a significant challenge but offers unparalleled insights into the tumor microenvironment and cell-cell interactions. How to integrate these complimentary multi-modal datasets within the same study is at the frontier of computational biology. We are leveraging existing advanced computational methods, such as MaxFuse and bindSC, to integrate different data modalities. Users can view cell types transferred from one data type to another as well as those same cell type predictions colored by how confident the algorithm was in its prediction. Users can also visually compare cell type predictions for two tools to see how they relate to each other. All views have been developed through user-centered design with user feedback and input.



Screenshots of beta release of Xena single cell views. Publicly available HTAN HMS colorectal cancer atlas data, a 19-plex proteomics dataset generated by Peter Sorger lab for sample HTA7_934_9. There are 1,810,481 cells in view, which were identified by the Sorger lab. (A) 3D UMAP embedding view of the proteomics data colored by louvain cell clusters. (B) Spatial view of CD3E protein expression (red color) overlaid on top of co-registered H&E and multiplex immunofluorescence images. Image size is 87040x41984 pixel / 2.82x1.36 cm. (C-D) Zoomed-in view with and without the co-registered H&E images. (E) Cell segmentation annotations become automatically visible once the user is zoomed in enough. Genomic data rendering is dynamically changed to an open circle to allow viewing of the images underneath. Red to transparent open circle outline indicates CD3E protein expression level from high to low.

Abstracts

Oral Session 2 - Spatial

Visual Analytics for Exploration and Hypothesis Generation Using Highly Multiplexed Spatial Data of Tissues and Tumors

Authors: Jeremy Muhlich, John Hoffer, Simon Warchol, Lukas Herzberger, Ino deBruijn, Johanna Beyer, Peter Sorger, Sandro Santagata, Hanspeter Pfister
Submitter: Jeremy Muhlich
Submitter Email: Jeremy_Muhlich@hms.harvard.edu

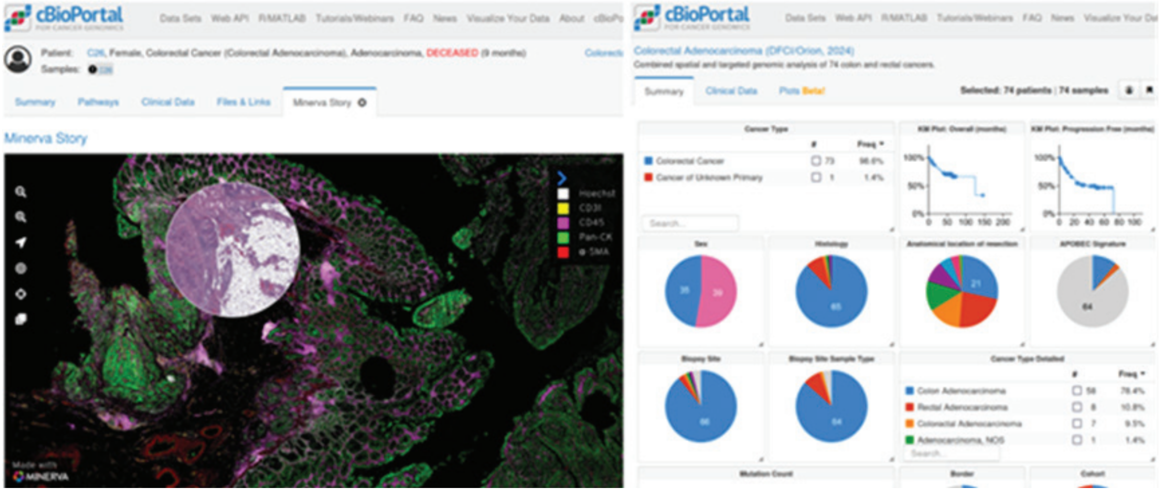
Spatial ’omics has emerged as a breakthrough technology for understanding the types and states of cells in the intact environment of tissues and tumors. Subcellular resolution, highly-multiplexed tissue immunofluorescence (IF) imaging is particularly promising because it builds on two centuries of histopathology and tissue biology, exploits the latest advances in computational microscopy, and links changes in protein levels to changes in morphology. Recent work in the field aims to combine the protein localization information from IF with small-region and single-cell spatial transcriptomics and metabolite imaging for an even more comprehensive view into cancer biology, and even 3-dimensional volumetric imaging of tissues is now on the horizon. We have observed that the most significant barrier to making such data routinely accessible to basic and translational cancer biologists lies not in data collection but rather data visualization and analysis. Existing tools for classical digital pathology and IF imaging of cultured cells simply do not scale to the enormous sizes of gigapixel whole-slide images with dozens of channels.

Previously, our team released the Minerva suite of tools for lightweight visualization and online narrative-driven sharing of highly multiplexed whole-slide images and derived data types. Our recent work has seen improvements to these tools in the areas of general image data exploration, hypothesis generation and testing through visual analytics approaches, and data publication and integration.

In support of image data viewing and exploration, we have integrated a dynamic movable “lens” feature into the Minerva narrative story viewer that allows users to “peek through” one image layer into another containing complementary information, for example viewing both multi-channel IF and classical histopathology hematoxylin & eosin (H&E) staining (Figure 1, left side). We also designed and implemented a novel “residency octree” approach to 3D volume rendering in a standard web browser using the WebGPU API, which will enable Minerva to efficiently render large 3D volumetric images without specialized software.

While working on visual analytics tools for multiplexed IF images, we developed a new method “Psudo” for improved pseudocolor image rendering that takes into account human visual perception and color theory. This method is applicable both to early image review and analysis as well as in choosing color maps for public image sharing. We also built an automated model for optimal contrast adjustment in high dynamic range IF image, replacing a tedious manual process.

Finally, we worked with the cBioPortal team (another ITCR-funded project) to integrate whole-slide imaging data directly inside cBioPortal using Minerva. We have created one dataset with highly-multiplexed IF + H&E images from 74 colorectal cancer patients in which curated clinical and genomic features can be browsed and searched alongside Minerva stories, all within the cBioPortal interface (Figure 1), as well as another such dataset encompassing 44 ovarian cancer patients. These datasets will be publicly released in Fall 2024.



Patient: C26, Female, Colorectal Cancer (Colorectal Adenocarcinoma), Adenocarcinoma, DECEASED (9 months)
Samples: C26

Colorectal Adenocarcinoma (DFCI/Orion, 2024)
Combined spatial and targeted genomic analysis of 74 colon and rectal cancers.

Selected: 74 patients | 74 samples

Cancer Type	#	Freq
🟦 Colorectal Cancer	□ 73	98.6%
🟥 Cancer of Unknown Primary	□ 1	1.4%

Cancer Type Detailed	#	Freq
🟦 Colon Adenocarcinoma	□ 58	78.4%
🟥 Rectal Adenocarcinoma	□ 8	10.8%
🟧 Colorectal Adenocarcinoma	□ 7	9.5%
🟩 Adenocarcinoma, NOS	□ 1	1.4%

Figure 1. Minerva Story integration with cBioPortal. Left: Minerva Story view of a single patient sample with H&E lens over five IF channels. Right: Dataset overview showing summary charts and search / filtering interface for clinical and genomic features.

Abstracts

Oral Session 3 - Omics

Using Large Language Models to Make Galaxy more Useful

Authors: Junhao Qui, Jeremy Goecks
Submitter: Jeremy Goecks
Submitter Email: jgoecks@gmail.com

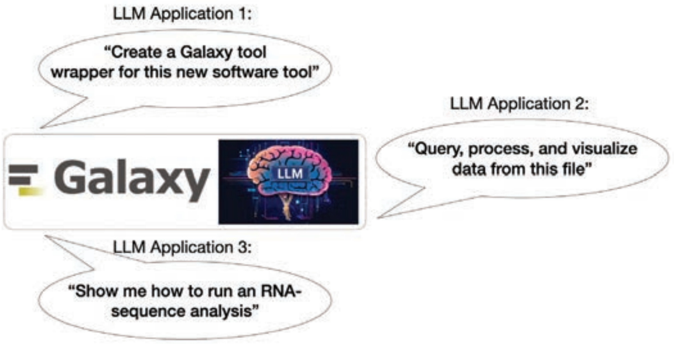
Large language models (LLMs) such as ChatGPT and Llama have proven very useful in supporting programming and data analysis tasks by partially automating them. Given the usefulness of LLMs, we are incorporating LLMs into the Galaxy computational workbench to increase the efficiency and accessibility of Galaxy for both scientific users and software developers.

Galaxy (https://galaxyproject.org/) is an open web-based computational workbench used by thousands of scientists daily throughout the world for many kinds of biomedical data analyses. There are >9,200 analysis tools and visualizations available in Galaxy's Tool Shed that can be used to perform all kinds of biomedical analyses, including genomics, transcriptomics, proteomics, microbiome, image analysis, machine learning, visualizations, and much more. Galaxy users can interactively run analysis tools and visualizations and create automated multi-tool analysis workflows. Galaxy can be accessed either through a web-browser or programmatically via an application programming interface (API), and Galaxy's API can be used to run automated and large-scale analyses.

We are incorporating LLMs into Galaxy in three ways. First, we are using LLMs to partially automate the process of writing "tool wrapper" scripts. Each software tool incorporated into Galaxy must have a tool wrapper script that describe the tool's inputs, parameters, and outputs to Galaxy. Through a combination of prompting with structured examples, LLMs can automatically generate much of the tool wrapper script from a software tool's help text. The second use of LLMs in Galaxy is to support the analysis of datasets using natural language. We have created a Galaxy interactive tool that loads a tabular file into an LLM, and a scientist can then ask the LLM to analyze the data in the file. For instance, a scientist can ask for a summary of the data, including the number and types of rows and columns in the data. A scientist can also ask the LLM to plot a heatmap showing the correlation of the columns. The final application of LLMs that we envision is implementing a natural language user interface (LUI) in Galaxy where scientists can type in actions to take. Example commands are "train a logistic regression model using dataset 1 as input", or "run the single-cell RNA-seq workflow on dataset collection 5." The Galaxy LUI will respond to user text queries with a ranked list of potential actions supplemented with references to relevant Galaxy resource documents. A user can then choose from the list of actions or explore the relevant documents before taking an action.

Implementing these LLM-enabled features in Galaxy is the first of several steps in this line of research. We will evaluate the performance of different LLMs on these tasks. Performance of different prompts and examples, fine-tuning, and structuring output will also be evaluated. Another future direction is to experiment with an agent-based interface that can take actions requested by a user, such as executing a Galaxy tool or workflow. How much autonomy to provide an LLM agent is an open question given there can be substantial costs associated with running analyses on large datasets.

We are implementing three applications of LLMs in Galaxy: (1) automate the process of writing Galaxy tool wrapper scripts; (2) support the analysis of datasets using natural language; and (3) implementing a natural language user interface in Galaxy.


LLM Application 1:
"Create a Galaxy tool wrapper for this new software tool"
LLM Application 2:
"Query, process, and visualize data from this file"
Galaxy
LLM
LLM Application 3:
"Show me how to run an RNA-sequence analysis"

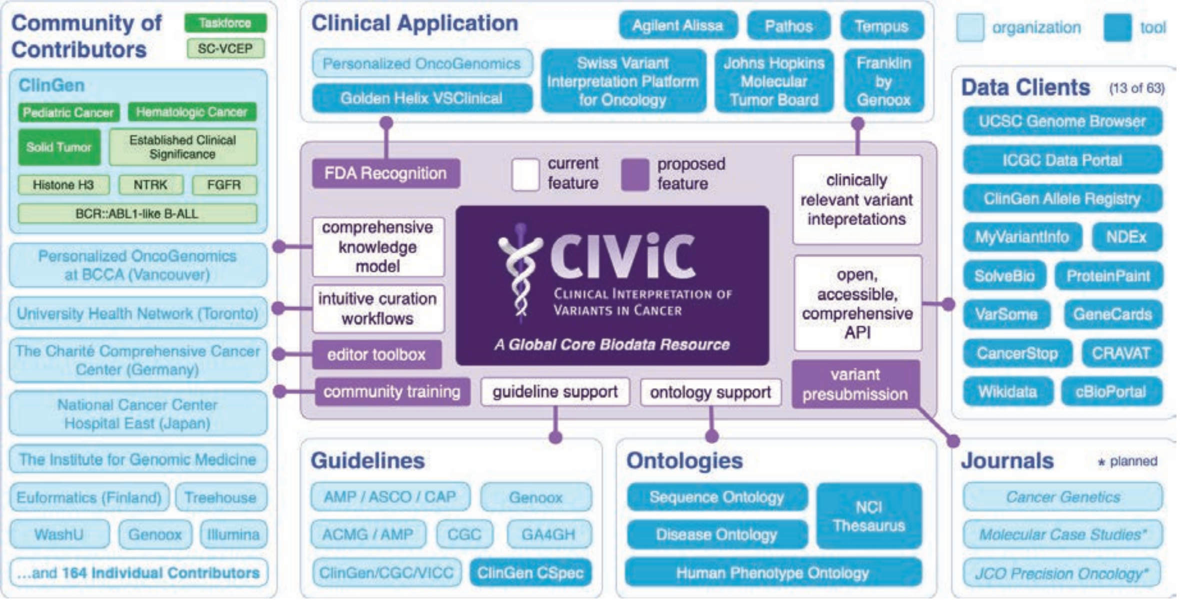
Abstracts

Oral Session 3 - Omics

Accelerating the expert-crowdsourcing of cancer variant interpretation in CIViC

Authors: Obi Griffith, Kilannin Krysiak, Arpad Danos, Jason Saliba, Joshua McMichael, Adam Coffman, Susanna Kiwala, Cameron Grisdale, Caralyn Reisle, Mariam Khanfar, Steven Jones, Alex Wagner, Malachi Griffith
Submitter: Obi Griffith
Submitter Email: obigriffith@gmail.com

Precision oncology involves the use of prevention and treatment strategies tailored to the unique features of each individual cancer patient and their disease. There has been an explosion in the number of molecular alterations or “variants” identified as cancer drivers or linked to cancer prognosis, diagnosis, or drug response. As a result, large numbers of patient-specific variants must be interpreted in the context of a vast and growing biomedical literature describing their clinical significance. These variant interpretations exist largely in private or encumbered databases resulting in extensive repetition of effort. Widespread adoption and standardization of precision medicine requires this knowledge to be centralized, standardized and expert-curated for application in the clinic. To address this need, we created CIViC, a community-driven web resource for Clinical Interpretation of Variants in Cancer, available online at civicdb.org. The CIViC resource is uniquely distinguished from others by its fully open access, rich data model, strong evidence provenance, and large community of volunteer expert curators. CIViC has been widely adopted by the community: with >5,000 individual users and >2,000,000 API requests per month, integrations into more than 60 academic and commercial data clients, as the official variant curation platform for ClinGen Somatic, and as a Global Core Biodata Resource. This widespread adoption has led to a dramatic increase in the numbers of users integrating CIViC into their clinical workflows and submitting content, creating a significant expert moderation and review bottleneck. Furthermore, new types and combinations of clinical biomarkers, standards and guidelines are continually adopted by the precision oncology community. To address these evolving challenges, we have introduced: a new complex molecular profile builder; new disease and drug web views; support for non-gene feature types such as fusions, factors and regions; improved feed to track curation/moderation; support for the new ClinVar somatic variant data model; and more. Using the existing CIViC knowledgebase as training data along with the hypothes.is tool and a custom CIViC browser extension we have also completed construction of a highly unique natural language inference dataset, CIViC-Fact, for developing precision oncology AI models. These models are being used to increase efficiency of biocuration and moderation by automatically “fact-checking” key components of new submissions to the database. We will discuss how these innovations and others allow CIViC to continue to scale to the needs of the cancer variant interpretation community.



CIViC’s role in the cancer variant interpretation ecosystem

Abstracts

Oral Session 3 - Omics

Customize your variant interpretation workflow with OpenCRAVAT

Authors: Jasmine Baker, Kyle Moad, Madison Larsen, Kyle Anderson, Supra Gajjala, Rachel Karchin
Submitter: Rachel Karchin
Submitter Email: karchin@jhu.edu

OpenCRAVAT is an open-source modular variant meta-annotator designed to make variant interpretation accessible to a wide audience. The modular design allows researchers to design customized workflows and utilize a diverse set of analysis methods, fostering a personalized approach to variant interpretation. While valuable information about variants is scattered across hundreds of databases and computational variant effect predictors, OpenCRAVAT centralizes access to these tools and makes them available through an easy-to-use interface. Hundreds of tools can be installed using point-and-click or simple command-line statements, and results are combined and presented in a variety of output formats. These tools cater to a broad range of variant types, encompassing germline, somatic, common, rare, coding and non-coding variants.

We support a wide range of popular input formats for fully customizing your variant annotation workflow, and we also offer tools to create your own input converters if needed. You can access our extensive library of annotation tools and have the option to develop your own annotators. Our system includes an extensive, flexible range of filters to help you reduce candidate lists from millions to just a few key variants. Additionally, we support multiple output formats and provide resources to create custom output reporters. For seamless integration into larger pipelines, all OpenCRAVAT functionalities are accessible via the command line.

Abstracts

Oral Session 4 - Imaging

Advancing Medical Image Visualization: Integrating Polymorph Segmentation in Cornerstone3D for Enhanced OHIF Viewer Capabilities

Authors: Gordon Harris, Alireza Sedghi, James Hanks, Dan Rukas, Rob Lewis, Chris Hafey, Trinity Urban, Erik Ziegler
Submitter: Gordon Harris
Submitter Email: gharris@mgh.harvard.edu

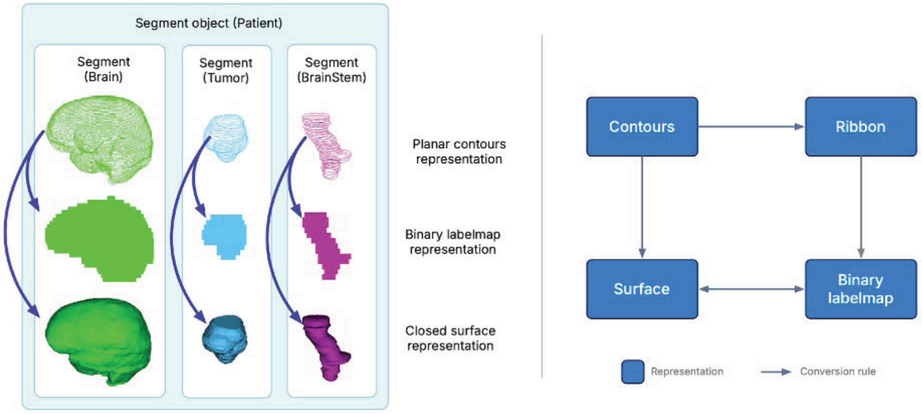
Introduction: The Open Health Imaging Foundation (OHIF) Viewer and its underlying Cornerstone3D lightweight JavaScript libraries are an open-source, web-based medical imaging framework that facilitates the visualization and manipulation of medical images in modern web browsers. OHIF and Cornerstone3D are widely utilized in academic and commercial projects, including NCI's Imaging Data Commons (IDC) and The Cancer Imaging Archive (TCIA), and the NIBIB/ARPA-H Medical Imaging Data Resource Center (MIDRC). OHIF and Cornerstone3D enable a range of imaging tasks, including multi-orientation viewing of volumetric images, interactive tools for image annotation, and segmentation editing and rendering.

One area of the OHIF Framework enhanced in the current year of our ITCR Sustainment grant involves image segmentation and feature representation. In the modern world of medical imaging, many segmentation use cases demand a more sophisticated solution than label maps. Polymorph segmentation, a concept pioneered by the 3D Slicer team (another collaborating ITCR-supported project), allows multiple representations of segmentation objects within a single framework. This approach optimizes storage, analysis, and real-time visualization of image-based objects. In our latest development, we introduced polymorph segmentation to Cornerstone3D, enhancing its capability to handle complex segmentation tasks efficiently.

Methods: We laid the groundwork for supporting multiple representations of the same segmentation in Cornerstone3D. This includes the addition of contour segmentation, which is suitable for rendering contours, such as those found in radiation therapy structure sets (RT STRUCT) in DICOM, the medical imaging format standard. Additionally, we added closed surfaces for rendering the segmentation in 3D. Conversion algorithms were also implemented to facilitate switching between different segmentation representations (see Figure 1). These converters, compiled from C++ to WebAssembly, run in a WebWorker to ensure the main user interface thread is not blocked, maintaining a smooth user experience.

Results: The enhancements to Cornerstone3D have resulted in better support for rendering various segmentation types. The use of contour segmentation significantly reduces memory usage, which is particularly beneficial for handling RT STRUCT contour structure sets. The system is now more versatile in representing segmentation data, providing better performance and memory efficiency.

Conclusion: We have laid the groundwork for representing segmentation in multiple forms within Cornerstone3D. This development allows for flexible and efficient handling of segmentation data, optimizing storage and visualization. This functionality will help support our current grant year set-aside project enabling eCountour, a widely-used open-access radiation therapy treatment planning and training platform, to utilize OHIF and Cornerstone3D for 3D RT mapping. Future efforts will focus on integrating these capabilities into the OHIF Viewer with a well-designed UI and UX, ensuring it remains a leading tool in medical image visualization technology.



Left images show three structures (Brain, Tumor, Brainstem) each depicted with three different segmentation object representations (Planar Contours, Binary Labelmap, Closed Surface). Right diagram shows conversion framework among various segmentation object representations.

Oral Session 4 - Imaging

Preclinical Imaging XNAT-Enabled Informatics (PIXI): An open-source resource to support cloud-based computational workflows for preclinical imaging

Authors: Kooresh Shoghi, Andrew Lassiter, Stephen Moore, James Quirk, Richard Laforest, William Horton, Daniel Marcus
Submitter: Kooresh Shoghi
Submitter Email: shoghilab@gmail.com

Preclinical imaging workflows have been growing in complexity, data size, and analytic requirements. We provide an update on our efforts to develop an open-source preclinical imaging XNAT-enabled informatics (PIXI) platform to manage the workflows of preclinical image data acquisition, to capture imaging-associated experiments including metadata and annotations, and to implement computational pipelines in a unified environment. Our vision for PIXI extends beyond the initial implementation to support a federated network of PIXI instances and a PIXI Center to enable data sharing and collaboration across institutions.

PIXI is based on the widely used XNAT platform as the underlying informatics architecture. PIXI includes: the PIXI Server, which provides core database, visualization, and workflow functionality; PIXI Notebooks for data exploration and analysis; and PIXI Apps to enable automated image processing pipelines through Docker container environment. With the recent release of PIXI 1.0 in February 2024, preclinical positron emission tomography (PET), computed tomography (CT) and magnetic resonance (MR) DICOM images are pushed to the PIXI server for workflow management with metadata captured through the PIXI Web-UI and DICOM image files for search and reporting. Multi-mouse images are supported by form-based data entry and Docker pipeline that splits hotel images into single-mouse datasets. Imaging workflow information can be entered or edited through the Web-UI. In addition, PIXI includes workflows to support upload and management of native Inveon PET and CT images and IVIS bioluminescence (BLI) images. XNAT's search and reporting capabilities have been extended to support PIXI's new data types and metadata. Importantly, we collaborated with the Open Health Imaging Foundation (OHIF) to expand the OHIF viewer's capabilities to support 4-dimensional image visualization and analysis of preclinical images. Finally, Jupyter notebooks has been integrated into the PIXI platform, providing researchers with seamless access to a secure scripting environment directly in the PIXI UI. Additionally, retrieving PIXI database metadata and information from within a Jupyter notebook is made easy with XNATpy. The Jupyter integration utilizes JupyterHub and leverages Docker containers and Docker Swarm for efficient management and scalability of computing resources for Jupyter notebook users. Building upon the Jupyter notebook integration, we have integrated Python dashboards directly into XNAT. Leveraging technologies such as Dash, Panel, Streamlit, and Voilà, this integration provides researchers and other technical users with tools to create and share interactive visualizations and applications that can then be started by general users from within the XNAT UI. This eliminates the need for users to interact directly with Python code.

Overall, the development of the PIXI platform is expected to have a profound impact on the management of preclinical imaging datasets and co-clinical imaging to support cloud-based computational pipelines and integration with multi-scale correlative biology. Since PIXI was released in February 2024, PIXI has over 60 users at various levels of activity, and we anticipate the user base to expand as we continue to disseminate and build the capabilities of PIXI. Additional information, including the free download of PIXI, instructional videos, documentation, and mailing list sign-up, is available at https://www.PIXI.org/.

Abstracts

Oral Session 4 - Imaging

Differential Privacy for Privacy-aware AI in Computational Pathology: Tool or Toy?

Authors: Sarthak Pati, Spyridon Bakas
Submitter: Spyridon Bakas
Submitter Email: spbakas@iu.edu

Background: Differential Privacy (DP) is a software-based technique for increasing privacy during AI model training. DP ensures that the output distribution of a randomized AI algorithm is indistinguishable between two datasets differing by at most a single record. The degree of indistinguishability, represented by E, signifies the privacy loss incurred by the algorithm. A smaller non-zero positive E implies stronger theoretical privacy guarantees, which is achieved when a model is trained with the maximum possible noise added to the gradient, whereas E implies no privacy at all, which is achieved when a model is trained with no noise. Tuning the training process with different noise levels is essential to ensure that we are able to train a private model while preserving the utility of model.

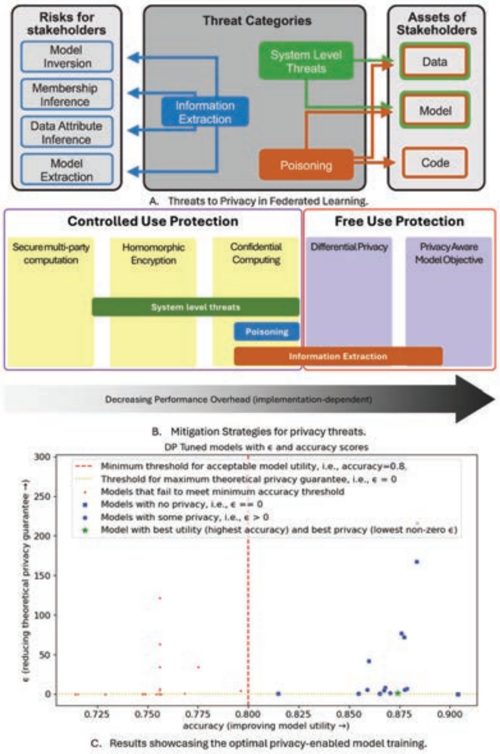
Tumor Infiltrating Lymphocytes (TILs) are a key biomarker in cancer research, across cancer types, associated with patient prognosis and informative for treatment approaches. Increasing clinical research evidence indicates that higher TIL levels correlate with better patient outcomes. Given their widespread occurrence in various cancers, their automated detection and quantification would provide a path towards designing a comprehensive and clinically useful biomarker.

Methods: This study utilizes a dataset to assess the presence of TILs in lung cancer histopathology image patches. The validation dataset consists of 4, 038 samples, with 985 positive and 3, 053 negative instances. We assume the "theoretical" worst classifier to be one which predicts all instances as negative, and thus has an accuracy of 3, 053/4, 038 = 0.75607. We therefore set a minimum accuracy of 0.8 for any AI model to be considered as having some utility.

DP was enabled by integrating Opacus into the Generally Nuanced Deep Learning Framework (GaNDLF), thus enabling zero-code DP model training only using YAML-based configuration files. DP training is performed across a wide array (from 0.0 up to 128.0) of various noise additions (i.e., "noise multipliers") and multiple maximum allowed norm (from 0.015625 up to 1e5) for the model gradients (i.e., "gradient clipping"), which forms the idea of "noise tuning". The overall experimentation yielded a total of 155 models. This is done to maximize the model utility (in this case measured via larger accuracy), and the maximum theoretical privacy guarantee, which is measured via lower non-zero ?.

Results: The "best" candidate AI model was achieved with E = 1.07, a noise multiplier of 1.0, and a gradient clipping of 0.031250, yielding the best validation cross-entropy loss (0.436251) and accuracy (0.873916) on the validation data. This showcased acceptable model performance (i.e., "utility"), while having the highest possible "theoretical" levels of privacy.

Conclusion: The study demonstrates the successful application of DP in AI model training, achieving a balance between privacy and accuracy. However, E provides only a theoretical maximum guarantee of privacy and further research is necessary to translate this into practical healthcare applications



A. Threats to Privacy in Federated Learning.

B. Mitigation Strategies for privacy threats.

C. Results showcasing the optimal privacy-enabled model training.

The validation accuracy and E scores of model training across various noise multipliers and gradient clipping values. The red dashed line represents the accuracy threshold, and red dots are all models that fail to meet it. The yellow dottet line represents the threshold for no privacy (i.e, E == 0), and the blue squares represent the models that meet the accuracy threshold but do not pass the privacy threshold (i.e., no noise added during training). The blue circles represent candidate models that meet the accuracy threshold and have noise added in the training process and hence have privacy guarantees. The green star represents the "best" model candidate with the lowest non-zero E (i.e., best privacy) and highest validation accuracy (i.e, best model utility).

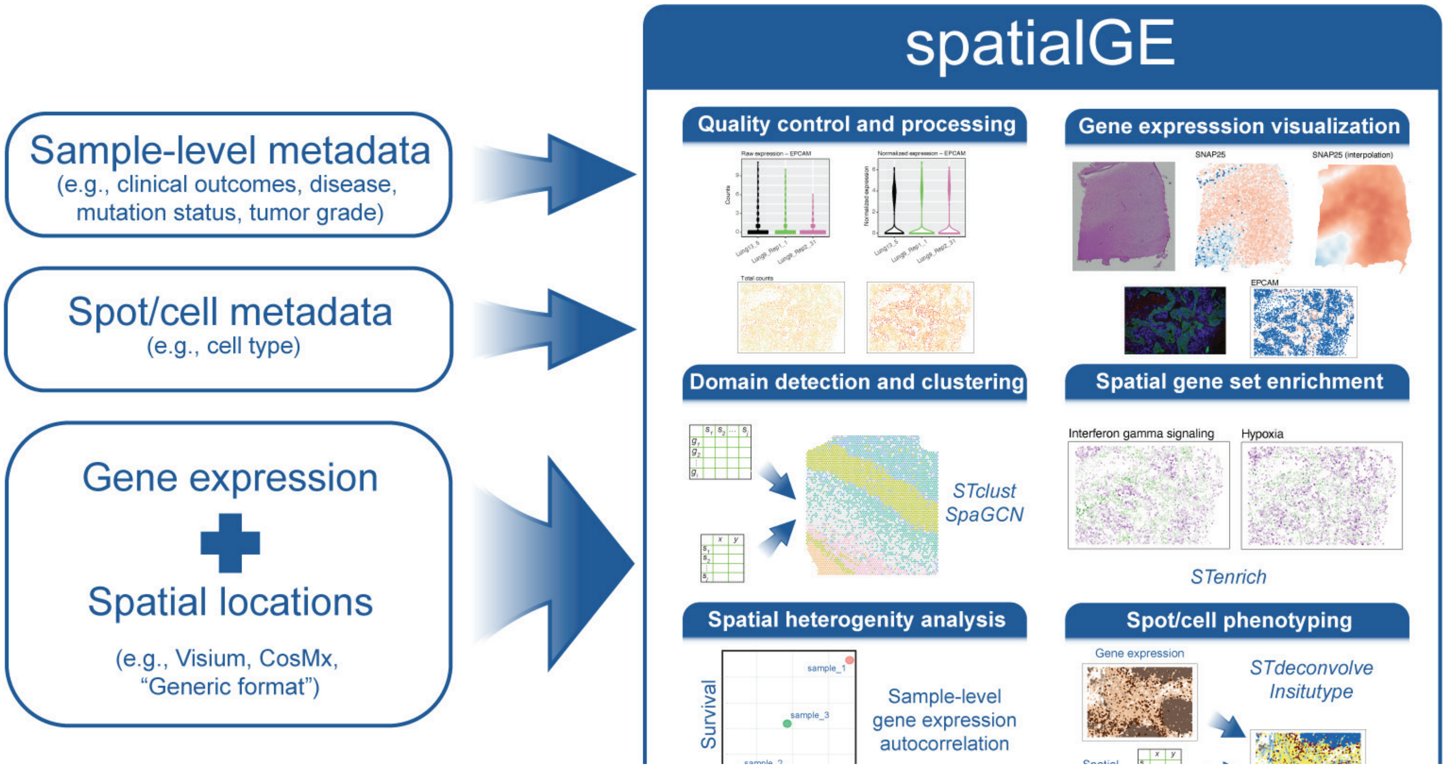
Abstracts

Oral Session 5 - Spatial

Democratizing spatial transcriptomics analysis with spatialGE

Authors: Oscar Ospina, Roberto Manjarres-Betancur, Guillermo Gonzalez-Calderon, Alex Soupir, Inna Smalley, Kenneth Tsai, Xiaoqing Yu, Brooke Fridley
Submitter: Oscar Ospina
Submitter Email: oscar.ospina@moffitt.org

The increasing popularity of spatial transcriptomics (ST) has led to the development of numerous analysis methods, each varying in robustness and user-friendliness. These diverse ST analysis approaches offer endless possibilities for extracting biological insights when studying the tumor microenvironment. However, ST data analysis can be challenging for non-data scientists, which limits their ability to conduct exploratory data analysis and generate new hypotheses for future functional experimentation. To address this challenge, we have previously developed a web application that wraps the functionality of the spatialGE R package to provide a comprehensive, user-friendly, point-and-click platform for the analysis and visualization of ST data. The spatialGE web application guides users through the various steps of analyzing ST data using detailed documentation and vignettes. Recognizing the diversity of tools, we have recently extended the capabilities of spatialGE by adding support to ST data analysis methods outside the spatialGE R package. Algorithms such as SpaGCN (tissue domain detection), STdeconvolve (cell type deconvolution), and InSituType (cell phenotyping) are now available to the cancer research community via the user interface provided by the spatialGE web application. Additionally, we have added support for analyzing single-cell ST data (e.g., CosMx-SMI) and provided test data sets to help users familiarize themselves with the functionality of spatialGE. Finally, we present results derived from the analysis of melanoma brain metastases (Visium) and Merkel cell carcinoma (CosMx-SMI) using the functionality of the spatialGE web application.


Sample-level metadata
(e.g., clinical outcomes, disease, mutation status, tumor grade)
Spot/cell metadata
(e.g., cell type)
Gene expression
Spatial locations
(e.g., Visium, CosMx, "Generic format")
spatialGE
Quality control and processing
Raw expression – EPCAM
Normalized expression – EPCAM
Counts
Normalized expression
Lung13_5
Lung9_Rep1_1
Lung9_Rep2_31
Total counts
Gene expresssion visualization
SNAP25
SNAP25 (interpolation)
EPCAM
Domain detection and clustering
STclust
SpaGCN
Spatial gene set enrichment
Interferon gamma signaling
Hypoxia
STenrich
Spatial heterogenity analysis
Survival
sample_1
sample_3
sample_2
Sample-level gene expression autocorrelation
Spot/cell phenotyping
Gene expression
Spatial
STdeconvolve
Insitutype

Overview of spatialGE functionality

Abstracts

Oral Session 5 - Spatial

Engineering model-based systems to monitor and steer subclonal dynamics

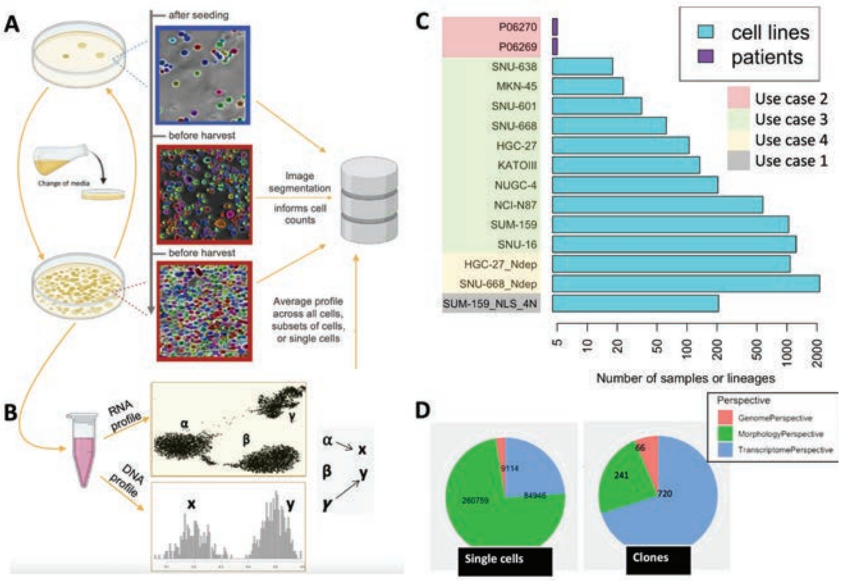
Authors: Thomas Veith, Saeed Alahmari, Vural Tagal, Richard Beck, Issam El Naqa, Noemi Andor
Submitter: Clone Redesign
Submitter Email: clonalredesignlab@gmail.com

Primary tumors as well as cancer cell lines exhibit extensive genetic and transcriptional heterogeneity, with multiple subclones co-existing in the same cancer population. The density under which cells grow has been shown to influence their adaptation to either produce a higher reproductive output or be less susceptible to contact inhibition. Even slight differences in media composition select for different pre-existing clones within a cell line. Periodic oscillations in the nutrient levels of cell lines also contribute to their heterogeneity. While a plethora of algorithms have been developed to infer the clonal composition of cancer specimens, no framework exists to monitor clone-specific measures over time. Here we engineer how in-vitro/in-vivo and in-silico experiments interact into a software solution called CLONEID. An SQL database in the backend, a Java core, and an R user interface come together to form two modules: the lineage tracing module records the pedigree of biological specimens analyzed within or across laboratories and uses computer vision to monitor phenotypic changes, such as variable growth rates. The second module links subclonal multi-omic profiles from different high throughput assays to each other and to the phenotypes from the first module. CLONEID addresses three unmet needs imposed by the spatio-temporal heterogeneity of cancer: i) experimental protocols that offer a high temporal resolution on in-vitro/in-vivo growth dynamics; ii) close monitoring of the temporal proximity between genotypic and phenotypic assay; and iii) reconciling the cost-prohibitive nature of repeated high-throughput multi-omic measurements with ceaseless changes in subclonal composition.

We demonstrate the breadth and depth of the utility of storing genotypic and phenotypic information in CLONEID through four use cases. In the first use case, we screen 11 compounds for their ability to generate a stable polyploid population from a near-diploid breast cancer cell line and use CLONEID to monitor spatio-temporal variation in ploidy level. In the second use case, we recapitulate pseudopalisading necrosis – a pathologic feature that is nearly unique to GBM – using a mathematical model. The model simulates growth and competition for resources by classifying cells as normoxic, hypoxic, or dead, and modeling their temporal evolution along with changes in oxygen. We fit the model to H&E-imaging derived data stored in CLONEID. For the last two use cases, we monitor the growth of gastric cancer cell lines with CLONEID for over 40 passages. We find that adaptation to culture conditions is predictable from transcriptomic data stored in CLONEID for the same cell lines (use case 3). We also compare growth rate adaptations in response to nutrient deprivation to the ploidy of the cell lines to test the influence of nutrient limitation on ploidy drive in cancer (use case 4).

CLONEID brings an infrastructure that scales, forming a foundation upon which a centralized database can be instantiated in the near future, to collect and connect genotypic and phenotypic experimental data worldwide across research laboratories. The benefits such a centralized database could offer include computer vision applications, amelioration of the irreproducibility crisis in biology, and especially streamlined generation of longitudinal datasets.

CLONEID workflow and database content. (A-B) CLONEID monitors phenotypic (A) and genotypic (B) information. (A) Seedings and harvests are physical activities performed by an experimentalist, each of which captured with an image and associated with a new entry in the database. (B) Multi-omic data identifies clones (e.g. scRNA/DNA-seq for quantification of copy number and gene expression). (C-D) Overview of CLONEID database content. (C) Number of lineages in the Passaging table stratified by sample source (cell line or patient) and color coded by use case. (D) Content of multi-omic profiles available as entries in table Perspective. Each entry is linked to a sample/lineage shown in (C). Recursive clonal membership structure allows for analysis of all cells, subsets, or single cells.



Abstracts

Oral Session 5 - Spatial

Quantifying spatial tumor heterogeneity

Authors: Cong Ma, Uthsav Chitra, Benjamin Raphael
Submitter; Ben Raphael
Submitter Email: braphael@princeton.edu

Recent spatial transcriptomics technologies measure RNA expression at thousands of locations in a 2D tumor slice quantifying important features of tumor heterogeneity such as the spatial distribution of cell types and spatial gradients in gene expression. Due to limitations in technology and cost, these measurements are typically sparse with high rates of missing data. We describe two software tools that address these technical limitations and provide a more accurate characterization of tumor heterogeneity by modeling the geometry of individual tumor slices. First, GASTON builds a "topographic map" of spatial transcriptomics data using a deep neural network, modeling continuous and discontinuous spatial variation in gene expression across a tissue slice. Second, CalicoST infers allele-specific copy number profiles for multiple spatially distributed tumor clones. We use these tools to analyze spatial transcriptomics data from multiple cancer types deriving gene expression gradients in the tumor microenvironment and reconstructing spatial tumor evolution.

Oral Session 6 - Omics

Pathway-guided Feature Selection and Integration for Cancer Subtyping

Authors: Ha Nguyen, Dung Pham, Hung Nguyen, Dao Tran, Tin Nguyen
Submitter: Tin Nguyen
Submitter Email: tinn@auburn.edu

Cancer is a complex disease driven by numerous biological pathways activating on multiple molecular levels. With the advancement of multi-omics platforms, cancer subtyping methods have shifted toward the integration of multi-omics data. The goal is to identify subtypes from a holistic perspective, taking into consideration phenomena at various levels. However, current approaches ignore the systems-level knowledge that hold the key characteristics of cancer subtypes. The lack of technologies that can incorporate pathway information, molecular data, and clinical data to cancer subtyping and prognosis presents a true barrier to address cancer health disparities. In this study, we introduce a new approach named Pathway-guided Feature selection and Integration for cancer Subtyping (PFIS), that can determine cancer subtypes using pathway-level information, multi-omics data, and relevant clinical variables. Through an extensive analysis utilizing real patient data encompassing over 11,000 patients across 33 cancers sourced from The Cancer Genome Atlas, we demonstrate the efficiency of PFIS against state-of-the-art methods for cancer subtyping. The proposed approach can identify meaningful subtypes with significantly different survival profiles in 28 out of 33 cancer datasets.

Oral Session 6 - Omics

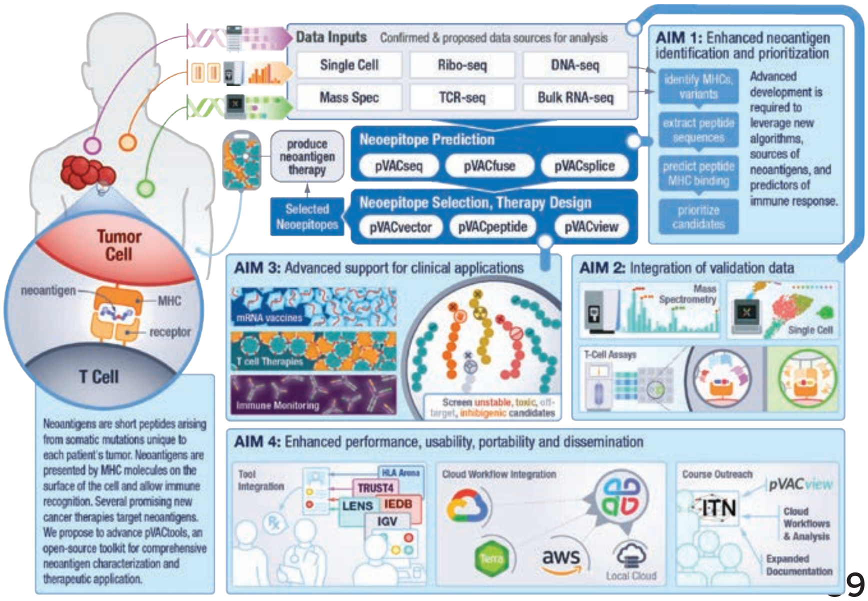
Informatics tools for neoantigen characterization and therapeutic translation

Authors: Susanna Kiwala, Huiming Xia, My Hoang, Kartik Singhal, Evelyn Schmidt, Mariam Khanfar, Joshua McMichael, Jasreet Hundal, Thomas Mooney, Jason Walker, S. Peter Goedegebuure, Christopher Miller, Todd Fehniger, Robert Schreiber, William Gillanders, Obi Griffith, Malachi Griffith
Submitter: Malachi Griffith
Submitter Email: malachig@gmail.com

We have developed pVACtools, an integrated collection of software tools that broadly enable basic science and translational research relating to neoantigens (http://pvactools.org/). Neoantigens are unique peptide sequences generated from mutations acquired somatically in tumor cells. These antigens provide an avenue for tumor-specific immune cell recognition and have been found to be important targets for cancer immunotherapies. Maturing sequencing technologies have allowed researchers to computationally predict potential neoantigens based on tumor-specific mutations. However, neoantigen generation and presentation is complex, and a host of factors must be considered to characterize each potential neoantigen. These include but are not limited to: somatic variant identification, tumor clonality assessment, RNA expression estimation, mRNA isoform selection, inference of translated tumor-specific peptides that arise from various classes of somatic variants, prediction of peptide processing, peptide transportation, peptide-MHC (pMHC) binding, pMHC stability and pMHC recognition by cytotoxic T cells. There has been a rapid development of computational tools that attempt to account for these complexities. However, there are considerable limitations to existing approaches and many outstanding questions related to neoantigen biology and the best approaches to target them therapeutically. pVACtools aims to broadly enable this research.

pVACtools has been widely adopted by the community and has facilitated studies of immune evasion, evolution of the tumor microenvironment, the relationship between neoantigen burden and prognosis, mechanisms of response to immune checkpoint blockade therapy and the neoantigen landscape in numerous tumor types. pVACtools has also become the most widely used tool in personalized neoantigen vaccine clinical trials. At our institute alone, pVACtools has been used to design personalized neoantigen vaccines for at least 184 patients involved in 12 ongoing clinical trials (NCT03422094, NCT03606967, NCT03199040, NCT05111353, NCT03532217, NCT02348320, NCT04397003, NCT03122106, NCT03988283, NCT03956056, NCT05741242), published clinical trials (NCT03121677) and case studies. The immune response data from these patients treated with >2,000 neoantigens collectively, have provided an exceptional opportunity to evaluate and inform improvements to pVACtools.

We will present an update on the capabilities and development of pVACtools and share our experience using pVACtools to directly support neoantigen targeting clinical trials. These experiences have provided valuable perspective on the current state and limitations of neoantigen prediction algorithms and pipelines. We will share our observations and current best practices drawn from successful execution of these trials, but also our view of the highest priority areas for future effort. These include a focus on new classes of neoantigens and integration of maturing data generation approaches including immunopeptidomics, T cell receptor sequencing, T cell immune monitoring and single cell sequencing. Finally, we will discuss how each of these can improve neoantigen clinical trial designs and how neoantigen tools must evolve to improve both the safety and efficacy of neoantigen vaccines and emerging neoantigen targeting T cell therapies.


Data Inputs
Confirmed & proposed data sources for analysis
Single Cell
Ribo-seq
DNA-seq
Mass Spec
TCR-seq
Bulk RNA-seq
AIM 1: Enhanced neoantigen identification and prioritization
identify MHCs, variants
extract peptide sequences
predict peptide MHC binding
prioritize candidates
Advanced development is required to leverage new algorithms, sources of neoantigens, and predictors of immune response.
produce neoantigen therapy
Selected Neoepitopes
Neoepitope Prediction
pVACseq
pVACfuse
pVACsplice
Neoepitope Selection, Therapy Design
pVACvector
pVACpeptide
pVACview
Tumor Cell
neoantigen
MHC
receptor
T Cell
AIM 3: Advanced support for clinical applications
mRNA vaccines
T cell Therapies
Immune Monitoring
Screen unstable, toxic, off-target, inhibigenic candidates
AIM 2: Integration of validation data
Mass Spectrometry
Single Cell
T-Cell Assays
AIM 4: Enhanced performance, usability, portability and dissemination
Tool Integration
HLA Arena
TRUST4
LENS
IEDB
IGV
Cloud Workflow Integration
Terra
aws
Local Cloud
Course Outreach
ITN
pVACview
Cloud Workflows & Analysis
Expanded Documentation

Neoantigens are short peptides arising from somatic mutations unique to each patient's tumor. Neoantigens are presented by MHC molecules on the surface of the cell and allow immune recognition. Several promising new cancer therapies target neoantigens. We propose to advance pVACtools, an open-source toolkit for comprehensive neoantigen characterization and therapeutic application.

Overview of pVACtools development roadmap including strategies for neoantigen discovery, validation, and clinical application.

Abstracts

Oral Session 6 - Omics

Fifteen Years of cBioPortal for Cancer Genomics

Authors: Ino de Bruijn, Tali Mazor, Rima AlHamad, Calla Chennault, Corey Dubin, Jeremy Easton-Marks, Zhaoyuan Fu, Benjamin Gross, Charles Haynes, David M Higgins, Jason Hwee, Jagannathan K Prasanna, Mirella Kalafati, Karthik Kalletla, James Ko, Tim Kuijpers, Sowmiyaa Kumar, Priti Kumari, Ritika Kundra, Bryan Lai, Xiang Li, James Lindsay, Aaron Lisman, Qi-Xuan Lu, Ramyasree Madupuri, Angelica Ochoa, Yusuf Ziya Özgül, Oleguer Plantalech, Matthijs Pon, Baby A Satravada, Jessica Singh, S Onur Sumer, Pim van Nierop, Floris Vleugels, Avery Wang, Manda Wilson, Hongxin Zhang, Gaofei Zhao, Ugur Dogrusoz, Allison Heath, Adam Resnick, Trevor J Pugh, Chris Sander, Ethan Cerami, Jianjiong Gao, Nikolaus Schultz
Submitter: Ino de Bruijn
Submitter Email: debruiji@mskcc.org

The cBioPortal for Cancer Genomics has significantly contributed to the field of cancer research over the past fifteen years by providing an open-access, user-friendly platform for exploring and analyzing large-scale cancer genomics datasets. Launched in 2008, it has empowered researchers worldwide to translate complex genomic data into biological insights and clinical applications, as demonstrated by >30,000 citations to date and the thousands of daily users of the public website (https://cbioportal.org).

One of the key reasons for the success of cBioPortal is its ability to provide access to large and complex genomic data sets through an intuitive interface, without requiring bioinformatics or computational skills. 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, as well as cohort and patient-level visualization.

The public instance of cBioPortal hosts data from >400 cancer genomics studies, encompassing diverse cancer types, molecular profiles, and clinical annotations. Originally developed to visualize data from The Cancer Genome Atlas (TCGA), cBioPortal has, over the years, been used to visualize data from many more consortia, including Stand Up To Cancer, AACR Project GENIE, and the Human Tumor Atlas Network (HTAN). Cancer genomics datasets are also obtained via curation of the literature, and through data submission by the community.

cBioPortal is fully open source (https://github.com/cBioPortal/). Development is a collaborative effort between several academic institutions and commercial partners. We have received code contributions from 110 individuals in the cancer research community and have developed effective processes for collective decision-making. cBioPortal is locally installed by many academic institutions, cancer centers, and pharma companies to visualize their private data. There are at least 93 instances of cBioPortal installed at academic institutions and companies worldwide. We provide detailed installation instructions for Docker Compose to enable users to deploy cBioPortal themselves and refer those who need more support to commercial partners.

cBioPortal has provided training for numerous cancer researchers and developers. We know of well over a hundred students who have partaken in internships and hackathons through internship programs, including the Google Summer of Code. We provide a variety of resources to help cancer researchers learn to use cBioPortal, including tutorial slides, webinars, and videos on YouTube which have been viewed over 100,000 times. The cBioPortal is also widely used in educational programs to teach students about cancer genomics.

cBioPortal continues to evolve and adapt to the rapidly changing landscape of cancer genomics. ngoing efforts focus on incorporating single-cell genomics and imaging data, integrating with clinical decision support systems, and expanding its analytical capabilities to address emerging research questions.

The cBioPortal for Cancer Genomics stands as a testament to the power of open science and collaboration in driving cancer research forward. Its fifteen-year legacy has solidified its position as an indispensable resource for the cancer research community, and its continued evolution promises to further empower researchers to unravel the complexities of cancer and develop more effective treatments.

Abstracts

Oral Session 7 - Clinical

GARDE: Scalable Clinical Decision Support for Individualized Cancer Risk Management

Authors: Guilherme Del Fiol, Caitlin Allen, Ravi Sharaf, Richard Bradshaw, Muhammad Danyal Ahsan, Emerson Borsatto, Elena Elkin, Melissa Frey, Kevin Hughes, Wendy Kohlmann, Polina Kukhareva, Anne Madeo, Che Martin, Chelsey Schlechter, Kimberly Kaphingst, Kensaku Kawamoto
Submitter: Guilherme Del Fiol
Submitter Email: gdelfiol@gmail.com

Evidence-based guidelines recommend risk-stratified cancer screening based on factors such as family history and genetic testing for hereditary cancer syndromes. However, risk stratification is underused due to barriers such as lack of time in primary care, low provider self-efficacy, and limited access to genetic counseling and testing. The Genetic Cancer Risk Detector (GARDE) platform uses (i) population-level algorithms based on National Comprehensive Cancer Network (NCCN) guidelines to identify eligible patients using cancer family history recorded in the electronic health record (EHR); and (ii) automated chatbots for patient outreach, pre-test genetic education, and testing facilitation.

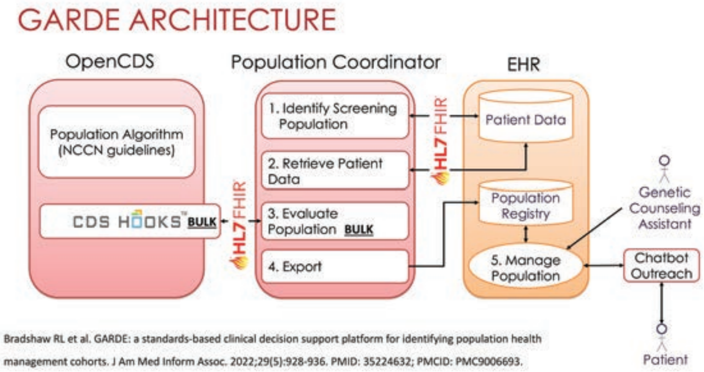
Prior work. The Cancer Moonshot-funded BRIDGE trial compared a GARDE-based approach to usual care on the uptake of genetic testing at University of Utah Health (UHealth) and New York University Langone Health (NYU). GARDE identified 5,155 out of 135,817 (3.8%) primary care patients at UHealth and 14,200 out of 208,071 (6.8%) at NYU who met criteria for genetic testing based on cancer family history documented in the EHR. Of those, 3,073 individuals enrolled in the trial.

Aims. In a recently funded ITCR study, we aim to (i) strengthen the GARDE infrastructure to facilitate deployment (Aim 1); (ii) create a chatbot authoring tool coupled with a content library to facilitate creation and sharing of chatbots among cancer researchers (Aim 1); (iii) develop an implementation toolkit using ITCR Training Network (ITN) tools (Aim 1); (iv) conduct a cost analysis of GARDE implementation (Aim 2); and (v) widely disseminate GARDE (Aim 3).

Progress update. GARDE has been deployed at two new sites: Medical University of South Carolina (MUSC) and Weill Cornell Medicine (WCM). Population analyses using GARDE identified 8,900 out of 178,000 (5.0%) meeting algorithm criteria at MUSC and 1,046 out of 8,420 (12.4%) at WCM. Both sites are conducting pilot studies to investigate different patient engagement and communication approaches supporting hereditary cancer syndrome genetic testing. GARDE is currently available via Docker; it has been successfully deployed on each site's virtual machines, in the cloud using Amazon Web Services (AWS), and Microsoft Azure. A preliminary version of the chatbot authoring platform supports the creation of rule-based chatbots (i.e., scripted content with pre-defined answer options) integrated with large language models via OpenAI's API for open-ended questions. The authoring platform has been shared with collaborators who are using it to create chatbot-based interventions for pragmatic clinical trials in various healthcare areas, such as tobacco cessation, assessment and assistance with social needs, newborn screening, and cascade testing for hereditary cancer syndromes.

Future work. GARDE will support five upcoming pragmatic clinical trials, including (i) augmenting GARDE algorithms with data from the Utah genealogy database linked to a cancer registry, (ii) adapted patient engagement approaches to increase the uptake of genetic testing among individuals from populations experiencing health disparities, and (iii) adapting GARDE to support interventions to increase the uptake of genetic testing for familial hypercholesterolemia.

Conclusion. With ITCR support, GARDE provides a scalable approach to hereditary cancer genetic testing that supports multiple clinical studies to facilitate individualized cancer screening and education



Bradshaw RL et al. GARDE: a standards-based clinical decision support platform for identifying population health management cohorts. J Am Med Inform Assoc. 2022;29(5):928-936. PMID: 35224632; PMCID: PMC9006693.

GARDE architecture

Oral Session 7 - Clinical

Demonstrating the Value of DeepPhe for Translational studies in Breast/Ovarian Cancer and Melanoma

Authors: Alexander van Helene, Harry Hochheiser, Jiarui Yao, Eli Goldner, Sean Finan, John Levander, Dennis Johns, David Harris, Piet de Groen, Elizabeth Buchbinder, Danielle Bitterman, Jeremy Warner, Guergana Savova
Submitter: Harry Hochheiser
Submitter Email: harryh@pitt.edu

Introduction: The goal of the DeepPhe project (U24CA248010) is the development of an open-source natural language processing (NLP) pipeline and visual analytics tool suite designed to unlock the rich information in the Electronic Medical Records (EMR) of patients with cancer. The DeepPhe NLP engine supports extraction of cancer and tumor characteristics from clinical narratives along with clinical genomics, comorbidities, procedures and treatments. This information is then combined with the structured EMR data and displayed through a visualization tool. We are demonstrating the utility of DeepPhe through clinical investigations across multiple cancers – breast cancer, ovarian cancer, colorectal cancer and melanoma – with data streams from three academic centers (Dana Farber Cancer Institute, UPMC and Vanderbilt University Medical Center).

Methods: Our translational studies utilizing DeepPhe combine structured EMR patient data with otherwise unavailable information extracted from EMR clinical narratives across three institutions to answer key questions:

1. What percent of patients with breast and/or ovarian cancer are assessed for tumor markers and/or clinical genomics? This question addresses the critical clinical importance of targeting anticancer therapies that require biomarker information. For example, the drug alpelisib requires knowledge of PIK3CA mutation status. This study, describes whether and when tumor markers and/or clinical genomics were assessed during the patient's journey. As these clinical genomic marker tests and results are often not available in structured data, DeepPhe NLP extracts biomarkers necessary to answer this question.

1. Are patients with metastatic melanoma and a BRAF mutation treated with immunotherapy first, or BRAF inhibition first? This question was partially answered through a phase 3 randomized clinical trial, DreamSEQ (2015-2021), which showed that immunotherapy first is better for BRAF-mutant melanoma patients, in a selected patient population. We replicate the DreamSEQ findings and characterize what actually happens in the real world across a broad spectrum of patients with a range of comorbidities and performance status.

For each question, we developed a study protocol listing the variables and the source for each variable (e.g. the clinical narrative or the structured EMR data). DeepPhe NLP was used to extract variables from the EMR clinical narrative, which were combined with structured data. Data from all three institutions were used in analyses, thus demonstrating generalizability and allowing stress testing the DeepPhe software in a variety of scenarios.

Results: As of May 30, 2024 DeepPhe NLP processed a total of 22,067,062 documents for 915,775 patients across three institutions: 750,685 notes (750,685 patients) at Dana-Farber Cancer Institute, 18,328,243 notes (73,312 patients) at UPMC and 2,988,134 notes (91,778 patients) at Vanderbilt University Medical Center. Each document took less than 0.25 seconds to process. The DeepPhe tool is available at https://deepphe.github.io/ .

Discussion: We have demonstrated the feasibility of deploying DeepPhe across three institutions and extracting key variables from EMR data at scale for translational studies. Difficulties including non-standardized structured data representations have been encountered and are consistent with other experiences working with real-world data across institutions. Results from the analyses will be presented at the ITCR meeting.

Oral Session 7 - Clinical

Matching Genotypes with Personalized Therapies: Development of a Decision Support Infrastructure to Augment the Value of Precision Medicine

Authors: Taxiarchis Botsis, Kory Kreimeyer, Jonathan Spiker, Maria Fatteh, Jamie Wehr, Mimi Najar, Jessica Tao, Nicole Imamovic, Ander Pindzola, Rena Xian, Adrian Dobs, Jenna Canzoniero, Valsamo Anagnostou
Submitter: Taxiarchis Botsis
Submitter Email: tbotsis1@jhmi.edu

Introduction
Despite tremendous progress in precision oncology with the release of several knowledge bases, meta-knowledge bases, and innovative information technologies, there are still major delays in identifying patients who could benefit by these advances. To date, automated solutions to efficiently match clinical-genomic phenotypes to molecularly-driven therapies, as part of clinical trials (CTs), FDA-approved, or off-label treatments, are lacking. Such technologies are paramount in precision oncology, especially for Molecular Tumor Boards (MTBs) that conduct detailed analyses of patient data and existing knowledge. They are also essential in the community setting where resources are limited.

Methods
To bridge this gap and within the NCI ITCR grant U01CA274631, we have been developing the web-based Open Navigator through Precision Oncology INformatics Technology (ONPOINT), a complete platform that integrates standardized clinical-genomic data with external information from several sources, such as clinicaltrials.gov. The current version of ONPOINT implements a sophisticated query that matches a patient's clinical-genomic profile, enriched with annotations provided by the OpenCravat tool, with selected trials pulled from the Aggregate Analysis of ClinicalTrials.gov (AACT) resource containing post-processed records from clinicaltrials.gov. An efficient user interface allows for a detailed review of the automatically retrieved trials using several filtering options, such as targeted variants, conditions, and locations. Users also have the option to view the full records at clinicaltrials.gov. The selected CT records are auto-populated to an editable auto-generated report that incorporates key clinical and pathologic information and MTB recommendations to support submitting physician's decision-making. ONPOINT is currently installed on an IRB-approved web-based environment, it is connected to the JH MTB's REDCap database using the REDCap API, and it actively supports the expert review process in the JH MTB.

Results
We evaluated the ONPOINT's clinical utility in informing decision-making in the JH MTB and the community setting within the JH Clinical Research Network (JHCRN). We first analyzed 16 patient cases where the MTB experts had used ONPOINT to conduct their review and provide genotype-targeted recommendations. The MTB experts selected 107 genotype-matched trials while the ONPOINT's automated query had first identified 98 of them (92%), missing nine trials. A subsequent error analysis informed improvements in the query that are now implemented to improve the performance of our approach. Next, we ingested a set of NGS outputs from 12 cancer patients receiving care at WellSpan Health, which is part of JHCRN, in the ONPOINT and evaluated the fraction of cases that would be suitable for molecularly driven therapies in the context of a CT. Out of 116 variants detected by next-generation sequencing, 27 were deemed actionable (23.3%) and forty genotype-matched CTs were identified nationwide. Of note, none of these patients received genotype-targeted therapies, which highlights the clinical value of our approach in identifying relevant targets and matching them to CTs.

Discussion
Further improvements include incorporating additional external knowledge from resources such as CiVIC, FDA's public repositories, and biomedical literature together with enhancing patient-treatment matching by leveraging previously MTB-recommended genotype-driven therapies. Ultimately a self-learning smart platform will augment the value of precision oncology and increase clinical trial enrollment.

Oral Session 7 - Clinical

Distributed multiple imputation for correlated incomplete data based on federated generalized linear mixed model.

Authors: Yi Lian, Xiaoqian Jiang, Qi Long
Submitter: Yi Lian
Submitter Email: yi.lian@pennmedicine.upenn.edu

It is often desired to jointly analyze healthcare data that are collected and stored at different sites, which can increase the sample sizes for underrepresented groups or rare disease outcomes among other benefits. Multi-site data often feature complex structures such as within-site correlations and between-site heterogeneity. These need to be properly accounted for in any statistical analyses of multi-site data as well as imputation procedures for missing data. Generalized linear mixed models (GLMM) are often used to analyze these data by allowing site-specific random effects and have been used in imputation for missing values. Due to the large sizes of data distributed at each site and privacy concerns, federated/distributed statistical machine learning methods that bring model to data have become necessary, including federated GLMMs. In this study, we propose to use privacy-preserving federated GLMMs to perform multiple imputation for missing data in healthcare data distributed at multiple sites. With considerations given to computation and communication efficiency, we provide a practical multiple imputation algorithm. We use simulation studies and real-world data analysis to demonstrate the performance of our proposed method.

Abstracts

Oral Session 8 - Clinical

GEARBOx: automated, patient-centric clinical trials matching

Authors: Luca Graglia, Brian Furner, Jooho Lee, Lauren Chan, Steve Krasinsky, Tomasz Oliwa, Enal Hindi, Michael Watkins, Kirk Wyatt, Samuel Volchenboum
Submitter: Rolando Palacios
Submitter Email: rpalacios@uchicago.edu

Introduction
Low clinical trial enrollment rates are a major impediment to drug development in pediatrics. This problem is particularly pressing for patients with relapsed and/or refractory disease. Effective treatments are desperately needed, and the rarity of these diseases–along with anticipated low per-site enrollment numbers–precludes many centers from supporting clinical trials.

Identification of trials a patient is eligible for is a cumbersome and manual process, often performed by manually scouring ClinicalTrials.gov using free-text search. Genomic Eligibility AlgoRithm for Better Outcomes (GEARBOx; GEARBOx.pedscommons.org) aims to streamline the clinical trials matching process by comparing patient characteristics with structured inclusion and exclusion criteria.

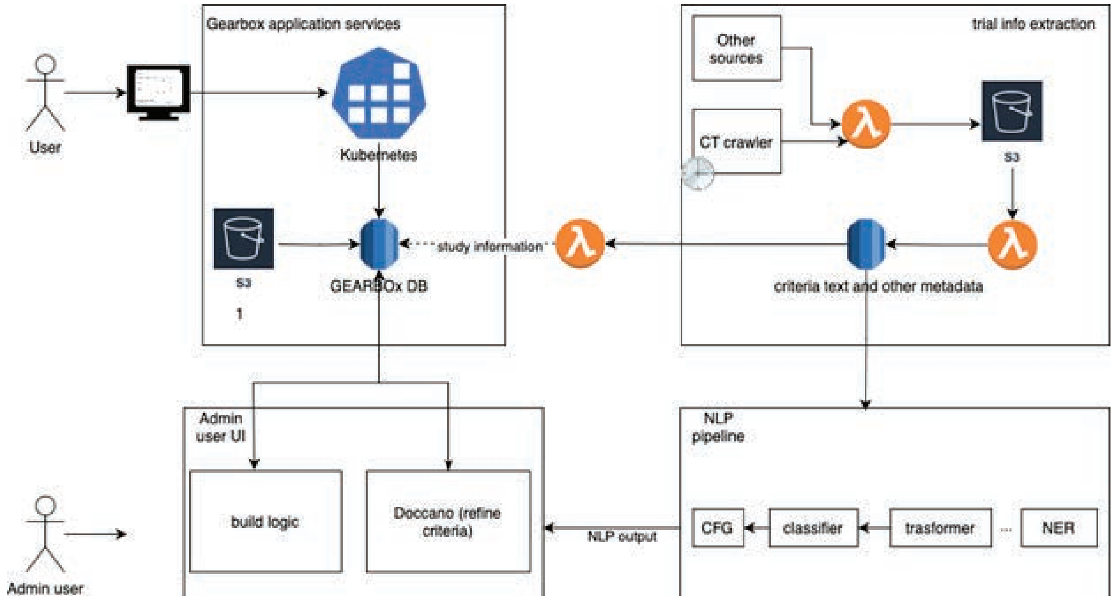
Current Operational State
The GEARBOx platform allows patients or healthcare providers to manually enter demographics and disease characteristics and be matched to relevant clinical trials for pediatric acute leukemia, neuroblastoma, and germ cell tumors. As of May 2024, over 245 unique users have used the tool. Since trial eligibility criteria are published as free text, the process for manual extraction and programming to facilitate automated matching is cumbersome, time-consuming, and error-prone. To streamline this workflow, we have developed a natural language processing (NLP) pipeline to facilitate extraction of eligibility criteria. The pipeline output is paired with a logic builder graphical user interface to streamline the process of adding new clinical trials. The pipeline consists of NLP techniques such as context-free grammar for capturing numerical variables with values, biomedical entity recognition, multi-class classification with transformer models, and the k-nearest neighbors algorithm. Each step of the pipeline maps predefined GEARBOx variables on sentence-, phrase- or entity-level data. All intermediate NLP outputs are filtered with optimal thresholds and presented to human annotators who approve the variable mappings. The models are continuously trained and refined with additional annotation data.

Features under Development
Using the automated NLP-based pipeline, we plan to add clinical trials for osteosarcoma and rhabdomyosarcoma by fall 2024. Current efforts to improve the NLP pipeline are focused on integration of a large language model. To facilitate the addition of a larger number and broader array of clinical trials, we are also ontologizing eligibility criteria concepts and aligning them with relevant standard terminologies. The robust computable and human-readable definitions, synonyms, and cross references found in biomedical ontologies will eliminate the problem of differently-worded but similar or identical criteria being included in duplicate within the tool. Additionally, providing contextual ontology content to users within the GEARBOx user interface allows respondents to provide more accurate data for optimal trial criteria matching. As a final automation step, we are developing methods to pull structured data from the electronic health record via FHIR using a decentralized architecture (Figure). This will reduce the data collection and entry burden for clinicians and is expected to increase acceptability and uptake.

Conclusion

By facilitating clinical trials matching across a large corpus of clinical trials, GEARBOx will assist clinicians in identifying relevant trials, increase choice for patients, and improve clinical trials enrollment rates, thereby expediting the process of drug development.

GEARBOx Infrastructure and Process Flow



Abstracts

Oral Session 8 - Clinical

Evaluating a Novel Algorithm to Process Electronic Adherence Monitoring Device Data

Authors: Meghan McGrady, Kevin Hommel, Constance Mara, Michal Kouril
Submitter: Meghan McGrady
Submitter Email: meghan.mcgrady@cchmc.org

Background: Medication adherence is a current scientific priority of the National Cancer Institute and a top priority in clinical trials. Electronic adherence monitoring devices (EAMDs) are pill bottles or boxes that contain a computer chip that records the dates and times of each bottle/box opening and are increasingly cited as the preferred measure of daily anticancer medication adherence in research. Unfortunately, researchers do not have access to tools that can accurately and efficiently convert raw EAMD actuations into usable adherence data and typically resort to hand re-coding. The lack of relevant tools make it difficult for most labs to utilize this rigorous measurement strategy and introduce concerns regarding human error and reproducibility to those that do. The purpose of our ongoing ITCR-supported R21 is to develop and evaluate the accuracy of a novel algorithm to convert EAMD actuations into adherence data. This abstract reports on the accuracy evaluation process and end-user feedback on the associated user interface (UI) to inform future software development.

Methods: Algorithm accuracy was evaluated by comparing the results of algorithm-produced daily adherence values to the data included in a hand-recoded database from an NIH-funded study including 300 months of EAMD data. Following validation, feedback on the algorithm and associated user interface was obtained from an observational study with 7 likely end-users (n = 6 Research Coordinators, n = 1 Principal Investigator) with 0.25-16 years of EAMD experience. Participants were asked to “think aloud” as they used the algorithm and then complete quantitative and qualitative measures of usability, desired refinements, and satisfaction. Quantitative results were summarized using descriptive statistics and qualitative feedback was analyzed using thematic analysis by the first author.

Results: Results of algorithm validation with 8,986 daily data points indicated that the algorithm demonstrates 99.90% sensitivity, 99.96% specificity, and 99.92% accuracy. Errors were limited to the incorrect handling of 7 missing values on edge cases by the algorithm and the algorithm was refined to address these errors. If selected for an oral presentation, a demonstration of the algorithm will be included. Algorithm pilot testing indicated end-users were enthusiastic about the product, with 100% (n = 7/7) indicating the algorithm would help them produce more reproducible and accurate data and save their teams time and resources. Although all end-users expressed interest in using the algorithm, each identified key features that must be integrated into an associated software package prior to adoption (e.g., REDCap integration, data cleaning and logging features, training tools).

Conclusions: Our novel algorithm accurately transforms data from electronic adherence monitoring devices into the daily adherence data researchers need for analyses. End-users are enthusiastic about the product and their feedback will be used to inform software development.

Oral Session 8 - Clinical

Facilitating androgen deprivation therapy treatment discussions between prostate cancer patients and their physicians via a comprehensive prognosis model that outputs personalized treatment benefit estimates based on cancer-related, genetic, and non-cancer risk factors.

Authors: Jessica Aldous, Matthew Schipper, Ralph Jiang, Robert Dess, Krithika Suresh, Elizabeth Chase, William Jackson
Submitter: Jessica Aldous
Submitter Email: jcaldous@umich.edu

BACKGROUND: NCCN and other national guidelines recommend adding androgen deprivation therapy (ADT) to radiation therapy for many men with localized prostate cancer. Due in part to the side effect profile of ADT, many men opt not to take ADT. In addition, most men with localized prostate cancer will not die of their cancer. There are few resources available that provide estimates of expected ADT efficacy for individual patients based on their prostate cancer and other cause mortality risk. Even fewer also incorporate genomic information, like decipher scores, which impart additional prognosis information on top of current staging systems.

METHOD: By integrating two independently validated prognostic models for prostate specific mortality (STAR-CAP) and other cause mortality (OCCAM), we were able to generate personalized treatment benefit estimates for adding short- and long-term ADT to radiation therapy. The novel two-step approach produces absolute risk estimates for distant metastasis, prostate cancer specific mortality, and all-cause mortality using prostate cancer risk factors, other cause mortality risk factors, and ADT treatment duration.To estimate the risk of prostate cancer specific mortality and distant metastasis, we integrated the validated prognostic STAR-CAP model with ADT and decipher hazard ratios estimated from a meta-analysis of randomized clinical trials. To account for mortality risk from other causes, we integrated the aforementioned risk estimates with the validated OCCAM model to estimate the absolute risk of PCSM, DM and ACM for individual patients based on treatment. Leveraging randomized trial data from the National Research Group (NRG), we will estimate the time varying hazard ratio of ADT treatment on other cause mortality to incorporate into our integrated model. The resulting model will be accessible via an R shiny app designed for physician-patient interactions.

CONCLUSION: Within current standard risk groups, ADT benefits vary widely between patients. A model like ours, which considers cancer, genetic, and non-cancer risk factors, provides personalized treatment benefit estimates. This model's results and presentation via intuitive software facilitates informed treatment decision conversations between patients and their physicians.

Abstracts

Oral Session 9 - Molecular

Inferring Kinase Activity from Tumor Phosphoproteomic Data

Authors: Sam Crowl, Candace Lei, Gabriela Salazar Lopez, Joseph-Levi Custer, Kristen Naegle
Submitter: Kristen Naegle
Submitter Email: kmn4mj@virginia.edu

Phosphoproteomic data, the new era of cancer biopsy profiling, has the potential to unlock precision targets – the kinases that have become aberrantly regulated in that patient's tumor. However, the variable coverage of discovery-based proteomics workflows presents major challenges to the usage and interpretation of the highly sparse data that has resulted from profiling of hundreds of ovarian and breast cancer patients. In our ITCR R21 project, we developed an approach (KSTAR) to infer kinase activities from phosphoproteomic data. Using a novel approach to handling both the predictions of kinase-substrate relationships and approaching the data with a tailored approach to the challenges and opportunities of phosphoproteomics we developed an approach that is significantly more robust, less prone to study bias, and outperforms other approaches (especially for tyrosine kinases). We found that KSTAR can complement clinical standard of care, identifying HER2-positive breast cancer patients that are non-responsive due to low HER2-activity and identifying HER2-negative patients that likely have HER2 activity and are actionable targets for HER2 therapy. In our U01 we are seeking to harden this approach, increasing the speed, lowering the memory requirements and expand to more kinases. We have a rich team of collaborators across a broad range of solid tumors and we are developing deployments for a range of scientists and clinicians.

Lancet2: improved performance and genotyping of somatic variants using localized genome graphs.

Authors: Rajeeva Musunuri, Bryan Zhu, Wayne Clarke, Timothy Chu, Jennifer Shelton, Dickson Chung, Shreya Sundar, Adam Novak, Benedict Paten, Nicolas Robine, Giuseppe Narzisi
Submitter: Giuseppe Narzisi
Submitter Email: gnarzisi@nygenome.org

One of the central challenges in cancer genomics is the ability to accurately detect somatic mutations in heterogeneous tumors, and precisely determine their clonal origin and evolution. This fundamental knowledge is central to the discovery of new cancer therapies. In recent years, reductions in the cost of whole-genome sequencing have enabled researchers to address these questions in unprecedented detail. However, indels and genomic rearrangements can create complex events that defy the traditional linear reference representation and are more challenging to detect and inspect with traditional read alignment tools. Genome graph structures are becoming increasingly popular to encode the genomic sequences from multiple related samples, but to date these developments have been limited to the analysis of germline variants.

Towards addressing these shortcomings, we have developed Lancet2 (https://nygenome.github.io/Lancet2/), the successor of Lancet, a somatic variant caller which leverages local assembly and joint analysis of tumor-normal paired data using region-focused colored de Bruijn graphs. The assembly graphs built by Lancet are small-scale sequence graphs that represent the local genome structures of the tumor and normal samples. Lancet2 is a complete redesign with focus on improved performance, software maintainability, and genotyping accuracy. We performed extensive benchmarking using multiple matched tumor/normal pairs (COLO829, HCC1143, HCC1187, HCC1395) which we sequenced at high-depth using a combination of short (Illumina) and long (ONT) reads. In comparison with other state-of-the-art callers (MuTect2, Strelka2, Octopus, VarNet, DeepSomatic) Lancet2 shows the highest accuracy without significant loss in sensitivity. Also, thanks to a new pull-based reactive multithreading approach, Lancet2 achieves near optimal CPU scaling compared to the other tools, which makes it an ideal choice for cloud deployment.

We will present our recent efforts, including: (1) improved genotyping approach based on the multiple sequence alignment of graph haplotype paths and re-alignment of reads to accurately calculate support; (2) optimized variant filtering with state of the art explainable glassbox machine learning models using InterpretML (https://interpret.ml); (3) use of ONT long-reads to validate variants and expand the truths sets of the cell-lines used for benchmarking with variants supported by the long-reads; (4) integration with SequenceTubeMap (https://github.com/vgteam/sequenceTubeMap) to allow the inspection of somatic variants using elegant tube-map visualizations which show read support aligned to the Lancet graph

Oral Session 9 - Molecular

Integrated metabolic and mechanical modeling for spatio-temporal simulations of tumors in their microenvironment

Authors: Ilija Dukovski, Louis Brezin, Kirill Korolev, Daniel Segrè
Submitter: Ilija Dukovski
Submitter Email: dukovski@bu.edu

Metabolic and biomechanical spatio-temporal heterogeneities are important aspects of tumor development, and their study is crucial for understanding of cancer progression and development of potential therapeutic approaches. The increased availability of spatially-resolved data presents an opportunity for developing predictive methods for translating complex datasets into forecasts of tumor development. To mitigate the gap between data and predictions we are developing spatio-temporal modeling and simulations of cancer metabolism, growth and progression in its microenvironment. In the past, biophysical models based on ordinary differential equations were developed for describing the growth and spread of cell populations. These models, however, seldom considered the details of cancer metabolism, or the metabolic heterogeneity of the tumor environment. Conversely, genome-scale models of human cell metabolism are currently being constructed and successfully used to simulate the complete network of metabolic fluxes in human tissues and cancerous cells. These models however, typically disregard spatial heterogeneity, and treat cancers as collections of identical cells. Through our modeling strategy, we integrate the biophysical and mechanical aspects of tumor growth with the detailed resource allocation dynamics encompassed by metabolism, thus creating a comprehensive platform for simulations of the growth and propagation, as well as metabolic activity in spatially-structured cancerous tumors. With this goal in mind, we repurposed our existing modeling platform COMETS (Computation of Microbial Ecosystems in Time and Space), to enable modeling of non-cancerous and cancerous human cell assemblies. First, we adopted the existing genome-scale model of human metabolism, Human-GEM, available from metabolicatlas.org, to our COMETS methodology. We validated the model using the dynamic Flux Balance Analysis algorithm at the core of COMETS to reproduce cell-lines growth curves on standard growth media. Second, we implemented a model of nonlinear cellular diffusion for the simulations of growth and propagation of cellular populations by mutual pushing and displacement, augmented with a model of demographic noise to simulate the stochastics of birth and death in populations. This model accurately reproduced the instability and branching of the growth front of bacterial colonies, more broadly relevant for growing cellular assemblies producing noncompact morphologies. We showed that the growth front branching enhances the formation and stabilizes the persistence of the homozygous sectors in heterozygous population mix, thus maintaining the genetic diversity of the population. We experimentally confirmed the detailed picture of colony morphology, including some newly identified features. Third, we developed a model of mechanical interactions of populations that are biophysically heterogeneous. By including a friction term between pairs of different cellular populations, we showed that a rich collection of morphologies can be produced by the interplay of differences in resources utilization, growth rates, and mechanical interactions. By integrating the genome-scale metabolic models with the biophysical model of mechanical interactions between heterogeneous cellular populations, we are on our way to creating a comprehensive platform for simulations of both metabolic and mechanical aspects of tumor propagation, and its interaction with a surrounding healthy tissue. The next step in our efforts will be three-dimensional implementation and other tumor-specific modeling capabilities, such as vascularization and therapeutic response.

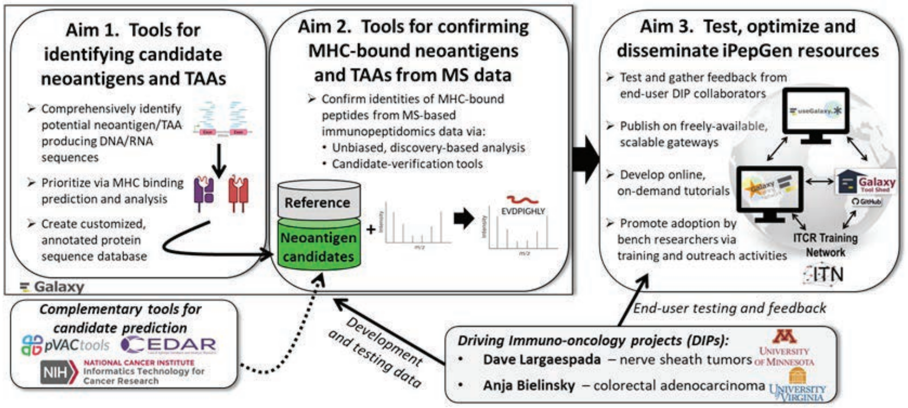
Abstracts

Oral Session 10 - New Award Lightning Talks

The immunoPeptidoGenomic (iPepGen) informatics resource for immuno-oncology research

Authors: Subina Mehta, Reid Wagner, Fengchao Yu, Alexey Nesvizhskii, Pratik Jagtap, Tim Griffin
Submitter: Timothy Griffin
Submitter Email: tgriffin@umn.edu

Recent years have seen a rapid rise in immuno-oncology studies seeking new therapies leveraging immunogenic, non-normal peptide sequences (neoantigens) arising from tumor-specific alterations at the genomic, transcriptomic or proteomic level. Non-normal DNA and RNA sequences that may encode neoantigens can be identified from next-generation sequencing (NGS) data, and further prioritized by predicting their binding to the class I or II major histocompatibility complex (MHC), along with other information on their potential immunogenicity. Immunopeptidomic enrichment of the peptide-MHC complex coupled with liquid chromatography tandem mass spectrometry (LC-MS/MS) confirms the existence of predicted neoantigens as well as other tumor-associated antigens (TAAs) derived from normal protein sequences, including those with post-translational modifications (PTMs). This powerful approach requires 'immunopeptidogenomic' informatics tools that integrate NGS and MS peptidomic data analysis. Despite steadily growing numbers of cancer researchers pursuing these studies, they lack a centralized informatics resource tailored to these requirements. As a solution, we are developing the immunopeptidogenomic (iPepGen) informatics resource. iPepGen will leverage the Galaxy bioinformatics ecosystem, offering cancer researchers accessible workflows to predict neoantigens from NGS data and confirm their presence from MS-based immunopeptidomics data. Our work focuses on achieving these Specific Aims: Aim 1: Optimize and harden modular workflows for identifying, prioritizing and curating neoantigen candidates detected from genomic and/or transcriptomic alterations. The Galaxy ecosystem houses numerous, current tools for assembly and analysis of NGS data, useful for identifying non-normal genomic and transcriptomic sequences that may encode for neoantigen peptides. These tools will be hardened into workflows capable of identifying and predicting candidate neoantigens and integrated with existing tools to predict their binding to the MHC, along with other emerging tools from ITCR-supported groups aimed at prioritizing those candidates with the highest immunogenicity potential. All candidate sequences are merged into a customized protein sequence database, including the host reference sequences, for identifying MHC-bound peptides from MS/MS data. Aim 2: Optimize and harden state-of-the-art MS-based immunopeptidomic analysis modules for identifying and verifying MHC-bound neoantigen and TAA peptides. We are extending Galaxy by implementing the FragPipe suite of tools for efficient matching of MS/MS peptide spectra to sequences contained in customized protein sequence databases generated in Aim 1. FragPipe outperforms other traditional algorithms for MHC-bound peptides in speed and depth of results. Identified peptides are further verified using the PepQuery tool within Galaxy, which rigorously re-evaluates putative peptide spectral matches (PSMs) discovered by FragPipe, providing a second layer of confidence in their identification. These tools identify both novel neoantigen peptides as well as TAAs, including peptides carrying PTMs which may regulate their binding to the MHC. Aim 3: Disseminate tested and optimized workflows and engage in training activities to promote community adoption of the iPepGen resource. Once optimized, online and on-demand training materials will be developed to guide new users through the iPepGen software and workflows, housing these within the freely available Galaxy Training Network resource. This presentation will provide an overview of the aims of our work, and an update on progress towards achieving our goals to date.


Aim 1. Tools for identifying candidate neoantigens and TAAs
➢ Comprehensively identify potential neoantigen/TAA producing DNA/RNA sequences
➢ Prioritize via MHC binding prediction and analysis
➢ Create customized, annotated protein sequence database
Galaxy
Aim 2. Tools for confirming MHC-bound neoantigens and TAAs from MS data
➢ Confirm identities of MHC-bound peptides from MS-based immunopeptidomics data via:
• Unbiased, discovery-based analysis
• Candidate-verification tools
Reference
Neoantigen candidates
+
EVDPIGHLY
Aim 3. Test, optimize and disseminate iPepGen resources
➢ Test and gather feedback from end-user DIP collaborators
➢ Publish on freely-available, scalable gateways
➢ Develop online, on-demand tutorials
➢ Promote adoption by bench researchers via training and outreach activities
useGalaxy
Galaxy Tool Shed
GitHub
ITCR Training Network
ITN
Complementary tools for candidate prediction
pVACtools
CEDAR
NIH NATIONAL CANCER INSTITUTE Informatics Technology for Cancer Research
Development and testing data
End-user testing and feedback
Driving Immuno-oncology projects (DIPs):
• Dave Largaespada – nerve sheath tumors
• Anja Bielinsky – colorectal adenocarcinoma
University of Minnesota
University Virginia

An overview of the aims for developing the iPepGen informatics resource.

Oral Session 10 - New Award Lightning Talks

Computational framework for inference of genetic ancestry from cancer-derived molecular data

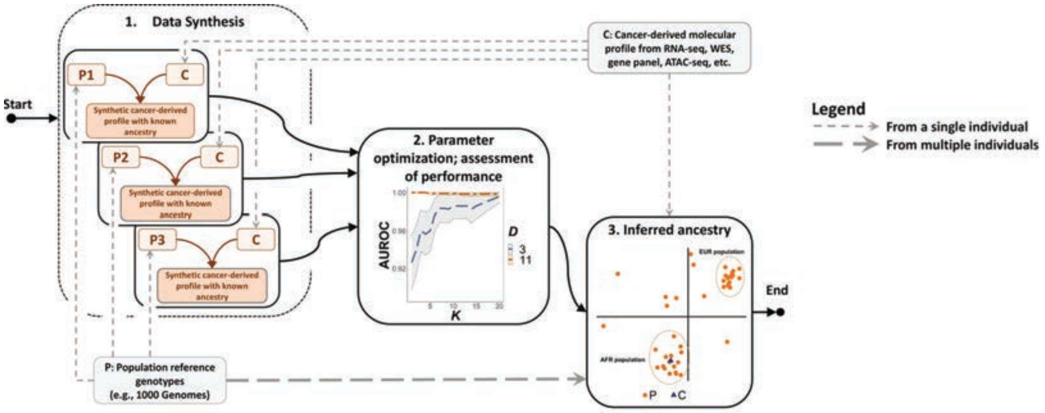
Authors: IPascal Belleau, Astrid Deschênes, Laine Marrah, David Tuveson, Alexander Krasnitz
Submitter: Alexander Krasnitz
Submitter Email: krasnitz@cshl.edu

For multiple cancer types, epidemiological data exhibit strong correlations between, on the one hand, the incidence of the disease, its severity when diagnosed, and its clinical outcome, and, on the other hand, the ancestral background of the patient. This well-documented phenomenon strongly suggests a link between the biology and genetics of cancer in an individual and the individual's genetic ancestry. Indeed, recent research in cancer genomics, both pan-cancer and cancer type-specific, points to genetic and phenotypic differences between tumors occurring in patient populations with differing genetic ancestries, and to the need for more data collection to power further study in this area. It is the purpose of our ITCR project to facilitate such data analysis on a much greater scale, by enabling genetic ancestry inference directly from cancer-derived molecular data, without the need for the patient's cancer-free genotype or self-declared race or ethnicity. Successful completion of this project will unlock vast amounts of such data for ancestry-oriented studies of cancer from two major sources. One is the body of data stored by the Sequence Read Archive (SRA) and similar massive digital repositories, on the order of 106 cancer-derived molecular profiles. The other is the body of archival tumor tissues across multiple medical centers, from millions of which molecular data may be generated.

To date, we have developed software tools for global genetic ancestry inference from multiple types of cancer-derived data, including DNA sequence data from whole genomes and exomes, targeted gene panels, RNA sequences and ATAC-seq. These inferential tools are adaptive, endowed with the ability to optimize their performance for each input cancer-derived molecular profile. This adaptability is achieved using synthetic data, combining the input cancer-derived profile with ancestral backgrounds representing well-defined population groups. As a result, these inference methods perform consistently, and with quantifiable accuracy, across a range of profiling depths and qualities. Our tools are now available as an R package, hosted by Bioconductor and termed RAIDS (Robust Ancestry Inference using Data Synthesis).

Work is currently in progress on expanding and refining these capabilities. The expansion will make our methodology applicable to whole genomes at low coverage and bisulfite-converted sequences. The refinements will enable our tools to deliver inference of global genetic ancestry at a sub-continental level of resolution, of ancestral admixtures and of local ancestry. and targeted sequence panels; RNA sequence data; ATAC-seq and bisulfite-converted sequence data. An open-source, user-friendly and FAIR-compliant software implementation of our methods will be made available to the research community on additional, cloud-enabled, platforms, including Galaxy. Training and community outreach for our software will be provided in collaboration with ITCR Training Network.

We expect our inferential framework to find applications far beyond cancer. Ancestral influences encompass diseases and conditions as diverse as Alzheimer's, lupus erythematosus and immune response to viral infection. Investigation of these influences would be facilitated by ancestral annotation of massive molecular data resources, such as SRA, using our tools.


1. Data Synthesis
C: Cancer-derived molecular profile from RNA-seq, WES, gene panel, ATAC-seq, etc.
Start
P1
C
Synthetic cancer-derived profile with known ancestry
P2
C
Synthetic cancer-derived profile with known ancestry
P3
C
Synthetic cancer-derived profile with known ancestry
2. Parameter optimization; assessment of performance
AUROC
1.00
0.96
0.92
5
10
15
20
K
D
3
11
3. Inferred ancestry
EUR population
AFR population
P
C
End
Legend
From a single individual
From multiple individuals
P: Population reference genotypes (e.g., 1000 Genomes)

Workflow for genetic ancestry inference from cancer-derived molecular data. 1: multiple molecular profiles are synthesized, each combining sequence quality and depth of a given input cancer-derived profile and a population reference genotype. 2: Synthetic profiles are used to optimize inference parameters and assess performance. 3: Genetic ancestry of the donor cancer patient is inferred using the optimal parameters.

Abstracts

Oral Session 10 - New Award Lightning Talks

Estimating the Distribution of Ratio of Paired Event Times in Phase II Oncology Trials

Authors: Li Chen, Mark Burkard, Jianrong Wu, Jill Kolesar, Chi Wang
Submitter: Chi Wang
Submitter Email: chi.wang@uky.edu

With the rapid development of new anti-cancer agents in precision medicine, new endpoints are needed to better measure treatment efficacy in phase II trials. For this purpose, Von Hoff (1998) proposed the growth modulation index (GMI), i.e. the ratio between times to progression or progression-free survival times in two successive treatment lines. An essential task in studies using GMI as an endpoint is to estimate the distribution of GMI. Traditional methods for survival data have been used for estimating the GMI distribution because censoring is common for GMI data. However, we point out that the independent censoring assumption required by traditional survival methods is always violated for GMI, which may lead to severely biased results. We construct both nonparametric and parametric estimators for the distribution of GMI, accounting for the dependent censoring of GMI. Extensive simulation studies show that our estimators perform well in practical situations and outperform existing estimators. A phase II clinical trial using GMI as the primary endpoint is provided for illustration.

Cancer Genomics: Integrative and Scalable Solutions in R/Bioconductor

Authors:
Submitter: Sean Davis
Submitter Email: SEAN.2.DAVIS@CUANSCHUTZ.EDU
Abstract not available

Methods for characterizing mechanobiology of the tumor microenvironment landscape

Authors: Shikhar Uttam
Submitter: Shikhar Uttam
Submitter Email: shf28@pitt.edu
Abstract not available

A Multilevel Data Analytic Solution to Advance Population Cancer Research

Authors: Johnnie Rose
Submitter: Johnnie Rose
Submitter Email: jxr109@case.edu
Abstract not available

A Multilevel Data Analytic Solution to Advance Population Cancer Research

Authors: Johnnie Rose
Submitter: Johnnie Rose
Submitter Email: jxr109@case.edu
Abstract not available

Oral Session 10 - New Award Lightning Talks

Enhancement and further development of informatics methods for long-read cancer sequencing

Authors: Heng Li, CZ Zhang, Katie Campbell
Submitter: Katie Campbell
Submitter Email: campbell@broadinstitute.org

Long-read sequencing is rapidly transforming our knowledge of the human genome as well as the approach to uncovering human genetic variation and alterations. As the cost and the base accuracy of long-read sequencing both approach short-read sequencing, we foresee a transformative potential of long-read sequencing in cancer genome analysis that is hindered by the lack of computational methods, benchmarks, and best practices for long-read cancer genome analysis. None of the current long-read informatic methods accounts for cancer-specific features including loss-of-heterozygosity, aneuploidy, and large chromosomal structural rearrangements. We propose to enhance existing tools and develop new informatic methods for analyzing cancer long-read data. These include minimap2, the de facto standard for long-read alignment, hifiasm, the most cited de novo assembler for accurate long reads, and mLinker, a versatile tool for haplotype inference from long reads and Hi-C data. We will also generate long-read DNA sequencing data on >100 cancer cell lines as a resource for benchmarking and evaluation of long-read informatic methods. Using these data, we will systematically evaluate both the technical performance of the long-read cancer informatic methods developed by us and others. We will also integrate our data and results with other data on the same cell lines to evaluate the potential impact of long-read cancer genomic discovery. Finally, we aim to build and expand an active community of researchers who interact with, generate, analyze, or develop informatic methods for long-read cancer genome data. This community building effort will initially focus on providing tutorials and user examples based on the newly developed informatic methods and newly generated long-read data, and eventually aim to establish a catalog of reference cancer genome assemblies for use by the cancer research community.

Informing mechanistic rules of agent-based models with single-cell multi-omics

Authors: Paul Macklin, Elana Fertig
Submitter: Paul Macklin
Submitter Email: macklinp@iu.edu

Cancer is driven by interactions between diverse cell types and their tissue microenvironment. Emerging single-cell and spatial transcriptomic systems are mapping cancer tissues, in the process capturing the diversity of cell types and states and exposing the importance of spatial cell interactions in determining therapeutic response in individual patients. Multi-omics software developed in ITCR—including CoGAPS, projectR, SpaceMarkers, and Domino developed by our group—can analyze single-cell and spatial multi-omics data to infer cell types and phenotypes in the tumor microenvironment, identify which cells interact, and discover how cell-cell interactions drive molecular changes. However, these analyses yield static snapshots that cannot capture the dynamics of cancer ecosystems. Mathematical modeling can "fill in the gaps" between these snapshots, allowing teams to form hypotheses on how and why cells interact, "encode" their hypotheses as simulation rules, and perform "virtual experiments." However, simulation rules and their parameters are difficult to match to genomic data. This proposal bridges the gap between bioinformatics and mathematical biology by merging our bio-informatics software for single-cell and spatial multi-omics data with PhysiCell, an agent-based modeling framework developed by our group to simulate the movement and interaction of many individual cell agents in virtual tissue environments. The "glue" between these packages is a novel cell behavior grammar that "encodes" cell rules learned from high-throughput data as intuitive, interpretable hypothesis statements that can be automatically transformed into simulation code. In this proposal, we refine the cell behavior grammar while analyzing previously published cancer data to create digital "templates" for key cell types in cancer ecosystems, refine PhysiCell to import the templates, and create PhysiCell Cloud: a free, "zero-install" cloud resource to build, execute, and visualize cancer models without writing computer code. We refine CoGAPS, SpaceMarkers, projectR, and Domino to learn cell behavior rules from spatial transcriptomics data and format them with the grammar, and extend PhysiCell to read cell types, positions, and rules stored in standard single-cell, spatial, and multi-omics classes. We develop sophisticated pipelines for PhysiCell models that can quantify model uncertainty, automatically fit models to transcriptomics data, and validate models on real world tumor datasets. We extend PhysiCell Cloud to a full-fledged science gateway that includes secure and searchable user storage, data structures and code (APIs) to connect PhysiCell Cloud to Python, R, and Bioconductor pipelines in ITCR, and a cost-free high-performance computing backend to seamlessly run large-scale model exploration and uncertainty quantification pipelines. Educational expertise and community feedback—including from an advisory board, annual training workshops, and daily classroom use—will drive usability refinements. Altogether, this approach will bridge bioinformatics and mathematical modeling to provide a comprehensive platform for patient-specific mechanistic tumor modeling, to enable future computationally-driven experimental design, virtual clinical trials, and digital twins research.

Oral Session 10 - New Award Lightning Talks

Structure-guided cancer immunotherapy design with HLA-Arena and CrossDome

Authors: Martiela Vaz de Freitas
Submitter: Martiela Vaz de Freitas
Submitter Email: mvazdefr@Central.UH.EDU

The broader use of T-cell-based therapies is still hindered by challenges related to the identification of peptide targets that are both immunogenic (capable of activating T-cells) and safe (do not trigger on-target/off-tumor or off-target toxicities). This is in part due to persistent dependency on biased sequence-based methods, despite recent breakthroughs in structural modeling and machine learning that could be leveraged to support new workflows for the identification of tumor-associated antigens (TAAs). To address this issue, and foster the design of better T-cell-based immunotherapies, we propose a new computational environment (HLA-arena 2.0) that will integrate existing ITCR resources, with new bioinformatics methods for structural modeling and analysis of key cellular immunity receptors; namely T-cell receptors (TCRs) and Human Leukocyte Antigen (HLA) receptors. Our working hypothesis is that the combination of multi-omics data with large-scale structure-based analysis can overcome most of the limitations of existing pipelines for TAA discovery, therefore enabling the design of better and safer T-cell-based immunotherapies. To test this hypothesis, we will implement a new workflow for structure-guided TAA discovery, integrating HLA-Arena with pVACtools (ITCR-funded package for sequence-based neoantigen discovery) and CrossDome (an R package for off-target toxicity prediction). In collaboration with researchers from MD Anderson Cancer Center, we will develop and test workflows to address existing needs in T-cell-based immunotherapy. We will focus on two different cancer types, that represent different challenges for cancer immunotherapy. We will benchmark our structure-guided TAA discovery workflow using immunopeptidomics data on melanoma. We will also run off-target toxicity predictions to identify the safest among 10 potentially therapeutic T-cell clones targeting two melanoma-derived TAAs from SLC45A2. Melanoma is a type of solid tumor for which greater success has been observed with immunotherapy treatments. On the other hand, acute myeloid leukemia (AML) is a type of blood cancer in which severe reactions to immunotherapy have been observed. In this context, we examine transcriptomic datasets (bulk and single-cell data) from AML patients, aiming at uncovering TAAs and TCRs that are associated with effective immune responses to AML. Finally, we will use CrossDome and existing data on known TAAs to develop The Cancer off-target Toxicity Atlas (TCTA). For each known TAA, this new database will contain a list of potential off-targets that should be tested when targeting these TAAs with immunotherapies. Predicted off-targets will be annotated with additional data (e.g., tissue expression, HLA-binding, immunogenicity, etc). All methods will be made available to the community through user-friendly workflows, facilitating the design of better and safer T-cell-based immunotherapies for numerous types of cancer. The proposed methods will be deeply integrated into the ITCR network, creating many opportunities for future collaborations. In addition, the long-term goals of the proposed research are well aligned with NCI's mission to achieve more effective and less toxic cancer treatments, therefore helping people live longer and healthier lives.

Integrative Analysis and Visualization Platform for Cancer Regulatory Genomics

Authors: Zhiping Weng
Submitter: Zhiping Weng
Submitter Email: zhipingweng@gmail.com
Abstract not available

Oral Session 11 - Trainee Lightning Talks session 1

The three-dimensional structure of extrachromosomal DNA reveals novel conformation changes

Authors: Biswanath Chowdhury, Kaiyuan Zhu, Chaohui Li, Jens Luebeck, Katerina Kraft, Shu Zhang, Lukas Chavez, Paul S. Mischel, Howard Y. Chang, Vineet Bafna
Submitter: Chaohui Li
Submitter Email: chl221@ucsd.edu

Somatic copy number amplifications of oncogenes are major drivers of cancer pathogenicity, particularly when the oncogenes are amplified by extrachromosomal DNA (ecDNA). EcDNAs are large, acentric, circular molecules that occur in nearly 20% of cancers. Random segregation of ecDNA promotes copy number heterogeneity leading to resistance and poor outcomes for patients. Moreover, their circular topology and conformational changes disrupt topological domains and rewire regulatory circuitry. A deeper understanding of the three-dimensional shape of ecDNA could improve our understanding of gene regulation on ecDNA.

Here we develop a method, ec3D, for reconstructing the 3D structures of ecDNA from Hi-C. ec3D models ecDNA genome as a (circular) chain of beads. It takes a candidate ecDNA sequence as a list of segments on the reference genome as input, extracts the ecDNA regions and reorganizes the Hi-C interactions to account for the rearranged genome, and infers the 3D structure of ecDNA through maximizing the Poisson likelihood of interactions (PASTIS, Varoquaux, 2014) given the distances between individual beads. Among key methodological innovations, ec3D allows representation of circular genomes. Second it utilizes constraint optimization to resolve ecDNA structures that contain multiple copies of large segments, by separating their chromatin configuration information. No current method for analyzing Hi-C data supports these tasks. Finally, ec3D outputs visualizations of the structure it computes which allows users to view, manipulate, and analyze the 3D structure of ecDNA. Such visualizations are essential for understanding the molecular architecture, function, and interactions of genes and regulatory elements on ecDNA.

To test the method, we first simulated 300 ecDNA structures with varying conformations and Hi-C data. Our method reconstructed the ecDNA 3D structures with high fidelity (Pearson Correlation coefficient (PCC)= 0.93) between true and reconstructed structures. By taking 300 pairs of random structures as negative samples, the method achieved the maximum F1 score of 0.88 at the cutoff PCC of 0.87.

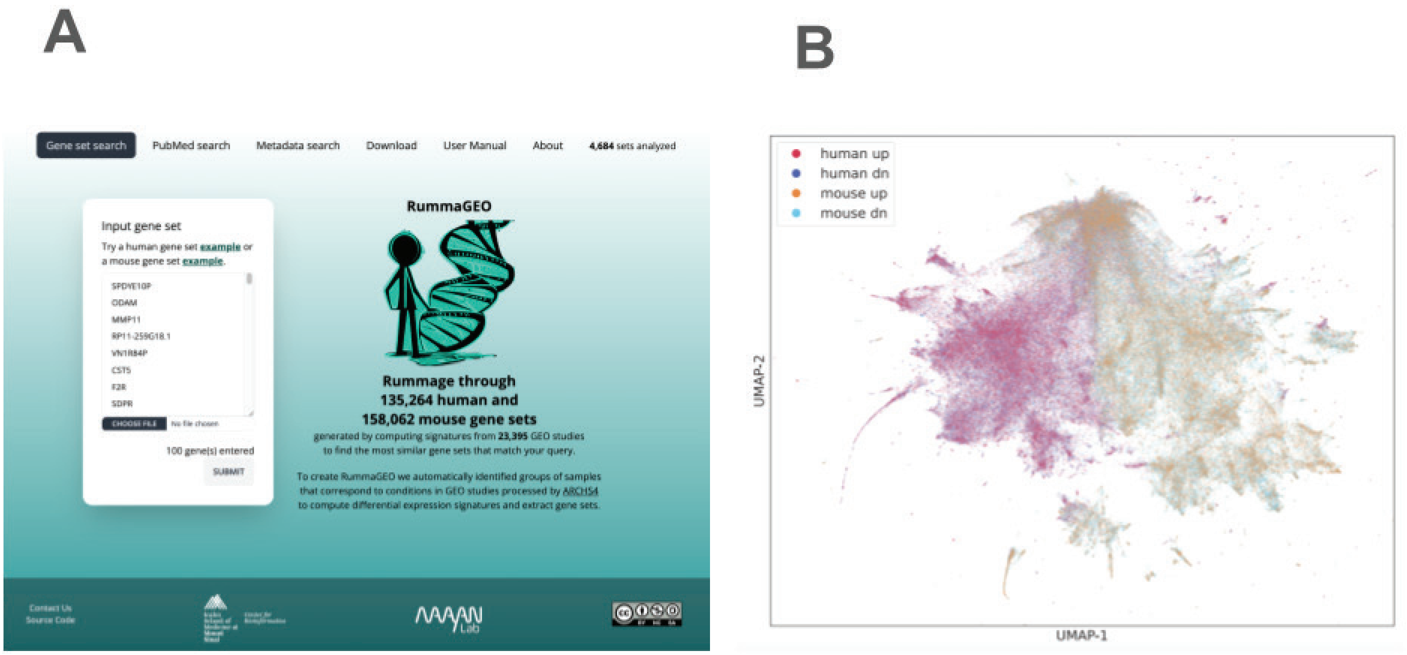
We then reconstructed 3D structures of ecDNAs from GBM39 (~1.25 Mb) and RCMB56 (~3.2 Mb) cell lines. In GBM39 where the ecDNA involves circularization of the genomic segments in close proximity, the topological domains were very similar to the original chromosomal structures, but the circularization created new proximities resulting in 3 distinct sub-loops. For RCMB56 ecDNA, which involves multiple genomic segments, we observed the disruption of normal topological domains and the creation of new domains due to conformational changes. Utilizing a Louvain community clustering on the structure, we found 4 pairs of genomic regions that were spatially close but far apart in genomic distance (separation >6.4Mbp). As one example, the structure suggested spatial proximity between chr1:93.77-93.91 Mbp containing the oncogene DNTTIP2 and chr1:87.26-87.34Mbp, over a genomic distance >6 Mbp. In summary, our method provides the first description of the 3-dimensional structure of ecDNA and hints at the role of the spatial configurations in gene regulation on ecDNA. A movie of the reconstructed structure is at this link.

Oral Session 11 - Trainee Lightning Talks session 1

RummaGEO: Automatic Mining of Human and Mouse Gene Sets from GEO

Authors: Giacomo Marino, Daniel Clarke, Eden Deng, Avi Ma'ayan
Submitter: Avi Ma'ayan
Submitter Email: avi.maayan@mssm.edu

The Gene Expression Omnibus (GEO) is a major open biomedical research repository for transcriptomics and other omics datasets. It currently contains millions of gene expression samples from tens of thousands of studies collected by many biomedical research laboratories from around the world. While users of the GEO repository can search the metadata describing studies for locating relevant datasets, there are currently no methods or resources that facilitate global search of GEO at the data level. To address this shortcoming, we developed RummaGEO, a webserver application that enables gene expression signature search of a large collection of human and mouse RNA-seq perturbation studies deposited into GEO. To develop the search engine, we performed offline automatic identification of sample conditions from the uniformly aligned GEO studies available from ARCHS4. We then computed differential expression signatures to extract gene sets from these studies. In total, RummaGEO currently contains 171,441 human and 195,265 mouse gene sets extracted from 29,294 GEO studies. Next, we analyzed the contents of the RummaGEO database to identify statistical patterns and perform various global analyses. The contents of the RummaGEO database are provided as a web-server search engine with signature search, PubMed search, and metadata search functionalities. Overall, RummaGEO provides an unprecedented resource for the biomedical research community enabling users to identify relevant GEO studies based on expression data. Users of RummaGEO can incorporate the RummaGEO results into their analysis workflows for comparative and integrative analyses as well as for hypothesis generation. The RummaGEO search engine is available from: https://rummageo.com/.



Try a human gene set example or a mouse gene set example.

100 gene(s) entered

RummaGEO

Rummage through 135,264 human and 158,062 mouse gene sets generated by computing signatures from 23,395 GEO studies to find the most similar gene sets that match your query.

To create RummaGEO we automatically identified groups of samples that correspond to conditions in GEO studies processed by ARCHS4 to compute differential expression signatures and extract gene sets.

(A) The RummaGEO user interface enables users to upload their gene sets to query against the gene sets contained within the RummaGEO database. (B) UMAP visualization of all the gene sets within the RumamGEO database. Each point in the plot represents a gene set and sets are arranged by their gene contents similarity.

Oral Session 11 - Trainee Lightning Talks session 1

Strategic Patch-based Deep Learning Workflow for Accurate Breast Cancer Cell Classification from Microscopy Images

Authors: Harrison Yee, Kailie Matteson, Joshua Goldwag, John Lamar, Margarida Barroso, Xavier Intes, Uwe Kruger
Submitter: Harrison Yee
Submitter Email: hajayee@gmail.com

Cancer cell identification within microscopy images is vital in diagnosing and guiding patient treatment following biopsies or tissue resections. Traditionally, such classifications rely on expert analysis of key cellular features such as cell and nuclear morphology. While automated solutions utilizing image analysis or deep learning have emerged, extracting different, or alternative, distinctive features allowing for cell line classification remains challenging due to physical and technical constraints. An example comes from prior work that has focused on organelle morphological and topological features for cancer cell line classification. Herein, we developed a workflow that employs fine-tuned patch-based image preprocessing and deep learning for automated classification of cancer cells within confocal microscopy images. Image preprocessing first consists of partitioning and filtering the entire 3D microscopy image into non-sparse 2D image patches. As optimal parameters for patch extraction depend upon the method of image acquisition, we also detail upon the grid search method our methodology employs for parameter tuning. Following the formation of the image patch dataset, a convolutional neural network (CNN) with channel-wise intermediate data fusion was used to perform breast cancer cell classification based on individual organelle features. Our methodology demonstrably outperforms conventional deep learning approaches when evaluated on a dataset encompassing six distinct cancer cell lines. Furthermore, it was determined that mitochondria emerged as the most informative organelle for classification as shown through single-organelle classifiers and Grad-CAM analysis. Overall, findings from this study demonstrate the practicality and robustness of this approach for cancer cell classification based on organelle features.

OmicsMLRepo: Ontology-leveraged metadata harmonization to improve AI/ML-readiness of omics data in Bioconductor

Authors: Sehyun Oh, Kaelyn Long, Kai Gravel-Pucillo, Levi Waldron, Sean Davis
Submitter: Sehyun Oh
Submitter Email: shbrief@gmail.com

Efforts to establish comprehensive biological data repositories have been significant at national and institutional levels. Despite the large volume of data collected from diverse studies, the cross-study analysis across those repositories and joint modeling between omics and non-omics data remains largely limited due to the nature of non-omics metadata, such as lack of standardization, high complexity, and heterogeneity. This lack of metadata harmonization also hinders the application and development of machine learning tools, which can serve a pivotal role in managing and analyzing complex and high-dimensional multi-omics data.

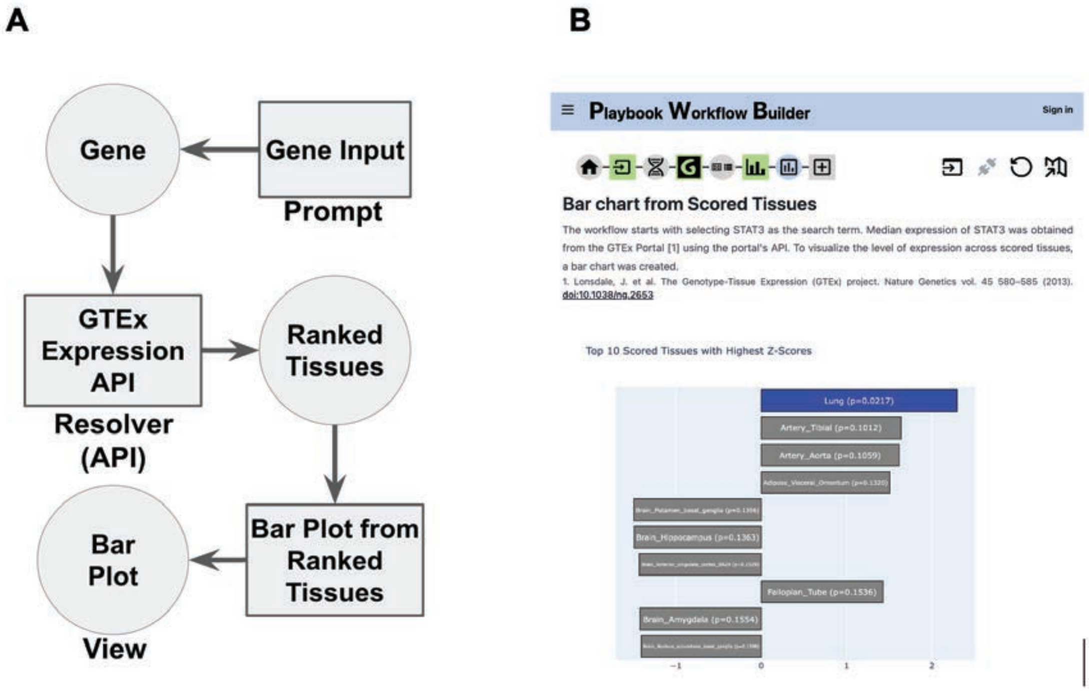
To address this issue, we initiated the OmicsMLRepo project, harmonizing and standardizing metadata from Omics data resources. This process involved the manual review of metadata schema, the consolidation of similar or identical information, and the incorporation of ontologies. As a result, we have harmonized hundreds of studies on metagenomics and cancer genomics data, accessible through two R/bioconductor packages - curatedMetagenomicData and cBioPortalData. Furthermore, we developed a software package, OmicsMLRepoR, allowing users to leverage the ontologies in metadata search. Using manually harmonized metadata as a gold standard, we are developing an automated metadata harmonization tool. This ongoing project leverages various Natural Language Processing (NLP) techniques and will be applied to other Omics data resources. In summary, the OmicsMLRepo project facilitates cross-study, multi-faceted data analyses through metadata harmonization and standardization, making omics data more AI/ML-ready.

Oral Session 11 - Trainee Lightning Talks session 1

Playbook Workflow Builder: Interactive Construction of Bioinformatics Workflows from a Network of Microservices

Authors: Daniel Clarke, Avi Ma'ayan
Submitter: Avi Ma'ayan
Submitter Email: avi.maayan@mssm.edu

Many biomedical research projects produce large-scale datasets that may serve as resources for the research community for hypothesis generation, facilitating diverse use cases. Towards the goal of developing infrastructure to support the findability, accessibility, interoperability, and reusability (FAIR) of biomedical digital objects and maximally extracting knowledge from data, complex queries that span across data and tools from multiple resources are currently not easily possible. By utilizing existing FAIR application programming interfaces (APIs) that serve knowledge from many repositories and bioinformatics tools, different types of complex queries and workflows can be created by using these APIs together. The Playbook Workflow Builder (PWB) is a web-based platform that facilitates interactive construction of workflows by enabling users to utilize an ever-growing network of input datasets, semantically annotated API endpoints, and data visualization tools contributed by an ecosystem. Via a user-friendly web-based user interface (UI), workflows can be constructed from contributed building-blocks without technical expertise. The output of each step of the workflows are provided in reports containing textual descriptions, as well as interactive and downloadable figures and tables. To demonstrate the ability of the PWB to generate meaningful hypotheses that draw knowledge from across multiple resources, we present several use cases. For example, one of these use cases sieves novel targets for individual cancer patients using data from the GTEx, LINCS, Metabolomics, GlyGen, and the ExRNA Communication Consortium (ERCC) Common Fund (CF) Data Coordination Centers (DCCs). The workflows created with the PWB can be published and repurposed to tackle similar use cases using different inputs. The PWB platform is available from: https://playbook-workflow-builder.cloud/.


A
Gene
Gene Input
Prompt
GTEx Expression API
Resolver (API)
Ranked Tissues
Bar Plot
View
Bar Plot from Ranked Tissues
B
≡ Playbook Workflow Builder
Sign in
Top 10 Scored Tissues with Highest Z-Scores
Lung (p=0.0217)
Artery_Tibial (p=0.1012)
Artery_Aorta (p=0.1059)
Adipose_Visceral_Omentum (p=0.1320)
Brain_Hippocampus (p=0.1363)
Fallopian_Tube (p=0.1536)
Brain_Amygdala (p=0.1554)
−1
0
1
2

Bar chart from Scored Tissues

The workflow starts with selecting STAT3 as the search term. Median expression of STAT3 was obtained from the GTEx Portal [1] using the portal's API. To visualize the level of expression across scored tissues, a bar chart was created.

1. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics vol. 45 580–585 (2013). doi:10.1038/ng.2653

Different PWB metanode types and how they are strung together to form workflows. A) In this example, the prompt type of metanode takes a gene as the input; then the resolver metanode uses the GTEx API to obtain the expression of the input gene from across human tissues. Finally, a view metanode visualizes the contents returned from the API as a bar chart. B) Screenshot from the executed workflow in the PWB platform.

Abstracts

Oral Session 11 - Trainee Lightning Talks session 1

Immunotherapy Treatment Outcome Prediction in Small Cell Lung Cancer (SCLC) through Topo-Geometric Characterization of Pulmonary Artery Vasculature: A proof-of-concept study

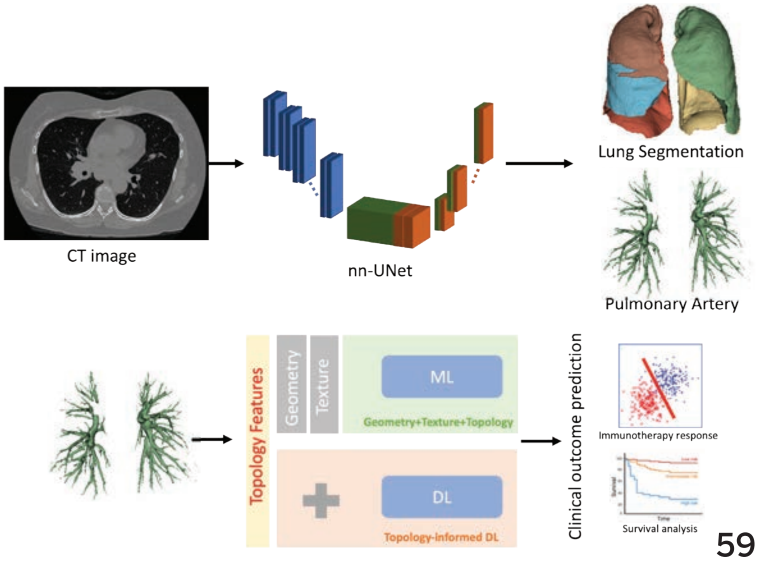
Authors: Moinak Bhattacharya, Jiachen Yao, Shirish M Gadgeel, Chao Chen, Prateek Prasanna
Submitter: Moinak Bhattacharya
Submitter Email: moinak.bhattacharya@stonybrook.edu

Introduction
Immunotherapy (IO) has emerged as a promising treatment for various cancers, yet its efficacy varies significantly among patients. We hypothesize that the morphology and texture of pulmonary artery vasculature plays a crucial role in IO treatment refractoriness in SCLC. Our goal is to characterize pre-treatment vessel morphology via topology and geometry-based attributes and use these CT-derived features to enhance RECIST-based IO response prediction. Furthermore, we employ a topology-informed deep learning model to integrate and analyze these features, providing a comprehensive approach to understanding the relationship between pulmonary artery characteristics and IO outcomes.

Materials and Methods
We conducted a retrospective study on 67 CT images of patients diagnosed with extensive-stage SCLC at Henry Ford Medical Center, focusing on those who received platinum-etoposide chemotherapy combined with immunotherapy. Using nn-UNet, we developed an automated pulmonary artery segmentation model. We extracted texture, geometric, and topological radiomics features from the segmented vessels, utilizing PyRadiomics, for vascular texture radiomics and image-based distance transforms for topological features. In another analysis, we developed a novel topology-guided deep learning prediction framework that incorporates the vessel mask as an auxiliary input, highlighting the spatial interplay between vessel regions and surrounding image patches. The segmentation performance was evaluated using metrics such as Dice Similarity Coefficient, and response/no-response classification and survival using AUC and C-index, respectively. Feature selection and hyperparameter optimization were conducted using ANOVA and Optuna, respectively, with statistical significance assessed through t-tests.

Results
Our segmentation model achieved a DSC of 0.87±0.01. For IO response prediction, combining texture, geometric, topological, and clinical features yielded an AUC of 0.78±0.13. The combined features achieved a c-index of 0.71±0.07 for the ipsilateral lung and 0.66±0.09 for tumor-adjacent lobes, outperforming individual feature sets. For the topology-guided deep learning framework, we achieved an AUC of 0.80 ± 0.04. Interestingly, our models significantly outperform a clinical feature-only model (age, sex, smoking status) suggesting the predictive value of vessel texture and morphology in IO response prediction.

Conclusion
Our study demonstrates that morphology and texture of lung vasculature is implicated in IO response and overall survival in patients with extensive-stage SCLC; we proposed two frameworks to incorporate such features into a response prediction model. Our approaches highlight the potential of advanced topo-geometrical radiomics and deep learning in personalizing immunotherapy regimens.



Workflow of the proof-of-concept study illustrating immunotherapy treatment outcome prediction in small cell lung cancer. The process integrates topo-geometric radiomics and deep learning to analyze the pulmonary artery vasculature and predict IP treatment response.

Oral Session 11 - Trainee Lightning Talks session 1

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

Authors: Debolina Chatterjee, Justin Couetil, Tianhan Dong, Jie Zhang, Kun Huang, Chao Chen, Travis Johnson
Submitter: Debolina Chatterjee
Submitter Email: dchatter@iu.edu

Spatially resolved transcriptomics (SRT) 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 processes. 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 tumor samples. DEGAS employs latent representations of gene expression data and domain adaptation to transfer disease attributes from 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 onto publicly available breast cancer 10X Genomics Visium data and validated these high-risk signatures in a novel dataset of Triple Negative Breast Cancer data generated by us. Additionally, we applied DEGAS to publicly available Nanostring CosMx FFPE samples from normal and Hepatocellular Carcinoma samples, revealing high-risk cells and high-risk cell topologies 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.

Oral Session 12 - Trainee Lightning Talks session 2

AmpliconSuite: Analyzing focal amplifications in cancer genomes

Authors: Jens Luebeck, Edwin Huang, Forrest Kim, Ted Liefeld, Bhargavi Dameracharla, Rohil Ahuja, Kaiyuan Zhu, Soyeon Kim, Hoon Kim, Roel G.W. Verhaak, Michael Reich, Paul S. Mischel, Jill Mesirov, Vineet Bafna
Submitter: Jens-christian Luebeck
Submitter Email: jluebeck@ucsd.edu

Focal amplifications in the cancer genome, particularly extrachromosomal DNA (ecDNA) amplifications, are a pivotal event in cancer progression across diverse cancer contexts. However, identifying and delineating different kinds of amplification events from whole-genome sequencing (WGS) data remains a challenge due to their complex profiles of copy number and structural variation. We present AmpliconSuite, a collection of tools that enables robust identification of focal amplifications from WGS data.

At the core of AmpliconSuite is the AmpliconArchitect (AA) method. AA jointly analyzes both structural variants (SVs) and copy numbers (CNs) within WGS data to identify and characterize focal amplifications. To create robust predictions of focal amplification status from AA outputs, we created AmpliconClassifier (AC), which classifies amplifications into distinct categories, including ecDNA, breakage-fusion-bridge (BFB) cycles among others. Combining these tools into a single workflow, we created AmpliconSuite-pipeline, available through GenePattern, Bioconda, Nextflow and other options.

To foster collaboration and data sharing, AmpliconSuite integrates with a platform we created called AmpliconRepository.org. This community-editable platform allows researchers to share focal amplification calls generated by AmpliconSuite publicly or privately. Notably, AmpliconRepository.org harbors ecDNA predictions on over 2,525 tumor samples from TCGA, PCAWG, and CCLE.

AmpliconSuite makes identification of focal amplifications reproducible and simple to use, and empowers users to share analyses publicly. Additionally, it introduces novel methods we recently developed, including ecContext within AC for categorizing types of ecDNA based on matching patterns of structural variation to the mechanisms of formation. Together, AmpliconSuite establishes itself as a valuable resource for researchers investigating focal amplifications in cancer.

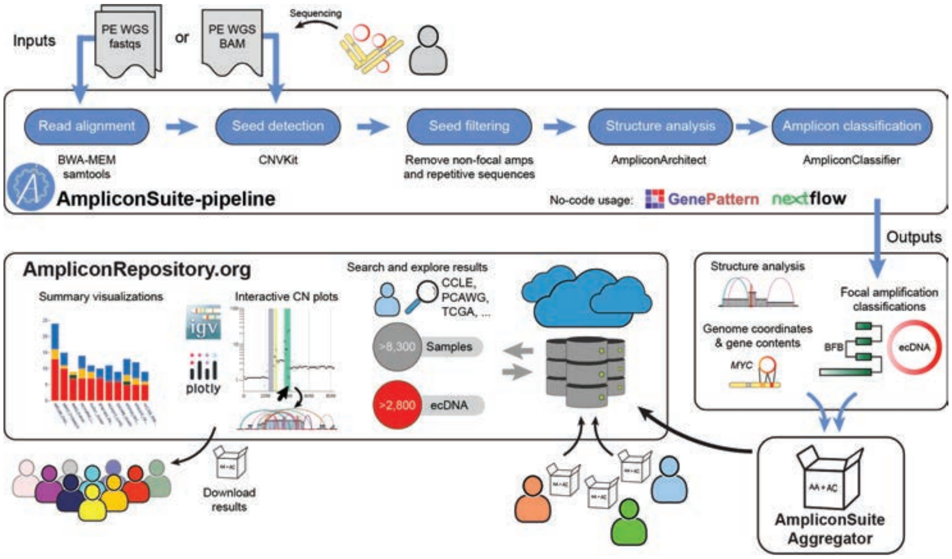


Illustration of the workflow wrapped by AmpliconSuite-pipeline, as well as its connection to AmpliconRepository.org.

Oral Session 12 - Trainee Lightning Talks session 2

Advancing Precision Oncology: Collaborative Efforts in the Curation and Classification of Somatic Cancer Variants by ClinGen Somatic and CIViC

Authors: Mariam Khanfar, Jason Saliba, Arpad Danos, Kilannin Krysiak, Adam Coffman, Susanna Kiwala, Joshua McMichael, Cameron J Grisdale, Ian King, Shamini Selvarajah, Rashmi Kanagal-Shamanna, Laveniya Satgunaseelan, David Meredith, Madina Sukhanova, Charles G Mullighan, Mark G Evans, Yassmine Akkari, Gordana Raca, Angshumoy Roy, Ramaswamy Govindan, Jake Lever, Alex H Wagner, Obi L Griffith, Malachi Griffith
Submitter: Mariam Khanfar
Submitter Email: k.mariam@wustl.edu

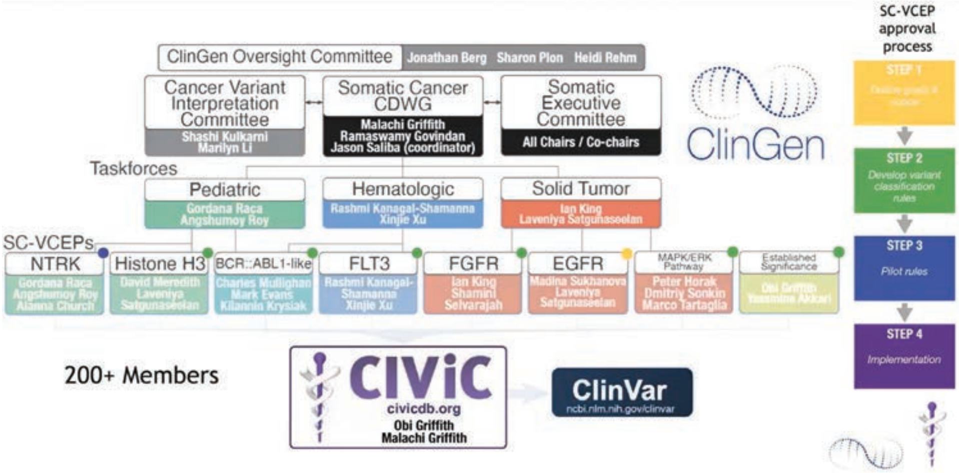
The accurate interpretation of somatic variants is essential for informed decisions in cancer diagnosis, prognosis, and treatment strategies. The Clinical Genome Resource (ClinGen) Somatic Cancer Clinical Domain Working Group (CDWG) has over 200 multi-disciplinary experts forming a community that develops data curation guidelines and standards that determine the clinical significance and oncogenicity of somatic variants in cancer.

The Somatic CDWG has established subspecialty-focused taskforces, including Pediatric Cancer, Hematological Cancer, and Solid Tumor. These taskforces lead curation projects working on developing Somatic Cancer Variant Curation Expert Panels (SC-VCEPs) through a detailed 4-step approval process. Currently, seven SC-VCEPs (NTRK-fusions, FGFR variants, FLT3 variants, Histone H3 variants, BCR::ABL1-like B-ALL alterations, variants of established clinical significance, and SNVs in MAPK/ERK-Pathway) are working through the approval process, while one SC-VCEP, EGFR variants, is in the planning stage.

These expert panels adapt existing AMP/ASCO/CAP and ClinGen/CGC/VICC Oncogenicity guidelines and develop new criteria as needed to assess the oncogenicity of specific genomic alterations. For instance, specific guidelines have been formulated for the classification of fusion proteins and complex mutations; such standards have led to the development of the first NTRK fusion-specific oncogenicity classification guidelines and the modification of the ClinGen/CGC/VICC Oncogenicity guidelines [PMID:35101336] for FGFR3, FLT3, and Histone H3 variants.

ClinGen collaborates with the Clinical Interpretations of Variants in Cancer (CIViC) knowledgebase to enhance the clinical utility of these curations. CIViC plays a crucial role in this ecosystem by providing a free and publicly accessible platform for the curation and dissemination of high-quality, expert-reviewed somatic cancer variant interpretations. It also supports the dissemination of ClinGen-approved classification codes and the creation of oncogenic assertions.

To date, the ClinGen Somatic groups have generated more than 793 CIViC Evidence Items and 43 Assertions from over 400 published papers. This ongoing collaboration is instrumental in advancing the field of precision oncology; as it not only enhances the interpretation of somatic variants but also ensures their practical applicability in clinical settings, ultimately leading to improved patient outcomes.



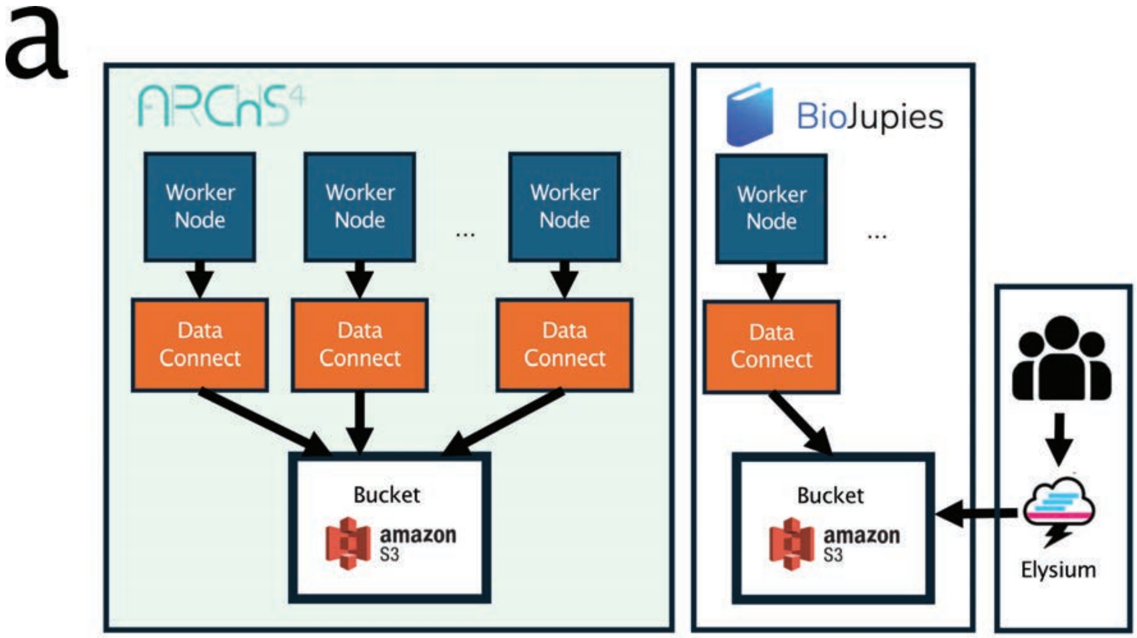
ClinGen Somatic Cancer Clinical Domain Working Group Partnered with CIViC to Accelerate Curation by through Somatic Cancer Variant Curation Expert Panels

Oral Session 12 - Trainee Lightning Talks session 2

DataCrossways: A Unified, Scalable Data Management Layer Applied to Enhance the ARCHS4 Resource

Authors: Alexander Lachmann, Avi Ma'ayan
Submitter: Avi Ma'ayan
Submitter Email: avi.maayan@mssm.edu

All RNA-seq and ChIP-seq Sample and Signature Search (ARCHS4) is a comprehensive gene expression data resource that provides uniformly aligned RNA-seq data from all human and mouse studies deposited into the Gene Expression Omnibus (GEO). The current data management layer of the ARCHS4 resource implements two separate cloud-based data storage systems to enable system administration and user access. To enhance the ARCHS4 service, we developed a new unified data management solution called DataCrossways. This upgrade addresses several limitations of the original implementation of ARCHS4 and enhances overall functionality. The initial version of ARCHS4 relied on an on-the-fly data loading, and the utilization of separate AWS S3 storage and pipelines for aligning user-uploaded data through BioJupies. The redesigned system introduces a unified pipeline that leverages scalable data management. This new architecture enables seamless integration of user-submitted FASTQ files together with the files processed to serve the ARCHS4 resource. The DataCrossways platform also enables users to have programmatic access to the ARCHS4 data management layer. This facilitates users to interact with the ARCHS4 pipeline programmatically. The always-on RNA-seq alignment pipeline, combined with the DataCrossways file management layer, facilitates efficient processing and management of both user-generated and system data. By implementing multipart upload functionality to S3, the new data management layer removes the limits of maximum file size. By consolidating the infrastructure of the ARCHS4 resource with DataCrossways, we improved maintainability, streamlined data management, and enhanced scalability. The DataCrossways platform, available from https://github.com/MaayanLab/datacrossways, was developed as a generic Data Commons template that can be utilized for the rapid deployment of other similar projects.



a) Current ARCHS4 data management layers including separate data management logic for BioJupies and direct user interactions. b) Unified ARCHS4 data management using DataCrossways.

Oral Session 12 - Trainee Lightning Talks session 2

pVACview: Visualization Tool for Neoantigen Prioritization

Authors: Evelyn Schmidt, Huiming Xia, My Hoang, Susanna Kiwala, Joshua McMichael, Zachary L. Skidmore, Bryan Fisk, Jonathan J. Song, Jasreet Hundal, Thomas Mooney, Jason R. Walker, S. Peter Goedegebuure, Christopher A. Miller, William E. Gillanders, Obi L. Griffith, Malachi Griffith
Submitter: Evelyn Schmidt
Submitter Email: evelyn@wustl.edu

Somatic variants in tumors give rise to unique peptide sequences called neoantigens that can activate CD8+ and CD4+ T cells. Neoantigens are being actively pursued as targets for cancer therapy. DNA and RNA sequencing advancements have allowed researchers and clinicians to computationally predict neoantigens and create personalized neoantigen therapies based on each patient's tumor-specific mutations. Currently, over 100 clinical trials are utilizing these therapies globally.

The growing interest in neoantigen cancer therapies makes understanding the details of the identification and prioritization of neoantigens crucial. Complexities can include alternative transcript annotations, algorithms that produce different binding measurements, presentation, and immunogenicity measurements, and variable peptide lengths/registers each with distinct characteristics. Recent studies have shown that factors such as mutation position, allele-specific anchor location, and the variation of T-cell response to long versus short peptides should also be considered when selecting neoantigens. Numerous computational tools have been developed to account for these complexities. The outputs of these tools are often prediction scores that require careful interpretation and understanding of the algorithms to conclude which neoantigens should be prioritized. Therefore, the users of these tools frequently consider only a simplified subset of predicted neoantigens, for example, limiting prediction to a single RNA isoform or only considering the top-ranked candidates, while other potentially important neoantigens are never evaluated. Currently, no tools successfully address the challenge of presenting all relevant information to facilitate candidate selection.

We have developed pVACview, an interactive visualization tool to aid in the efficient and accurate prioritization and selection of neoantigens for personalized neoantigen therapies. pVACview has an intuitive interface that succinctly summarizes neoantigen predictions while providing detailed variant, transcript, and peptide information, allowing for a thorough exploration of each candidate. pVACview supports the visualization of any neoantigen prediction tool's output such as vaxrank and annotation tools such as NeoFox. Users can upload, explore, and export evaluations and comments on their neoantigens which increases accuracy and effectiveness in basic and translational settings. The application is available as part of the pVACtools suite at pvactools.org and as an online server at pvacview.org.

Oral Session 12 - Trainee Lightning Talks session 2

Immune validation of neoantigen vaccines through clonal TCR analysis in patients with pancreas cancer

Authors: Kartik Singhal, Felicia Zhang, Xiuli Zhang, S. Peter Goedegebuure, Christopher A. Miller, Gue Su Chang, Jasreet Hundal, John Garza, Mike D. McLellan, William E. Gillanders, Obi L. Griffith, Malachi Griffith
Submitter: Kartik Singhal
Submitter Email: kartiksinghal2814@gmail.com

Neoantigen vaccines represent a promising strategy for cancer immunotherapy, providing personalized treatment options based on individual tumor-specific antigens. Somatic mutations in tumor cells generate neoantigens that may bind to major histocompatibility complex (MHC) molecules and get presented on the cell surface. These neoantigens can be recognized by T cells to initiate a tumor-specific response and are important targets of cancer immunotherapies. The pVACtools software suite, developed by the Griffith Lab at Washington University in St. Louis, provides a comprehensive computational platform for the prediction and prioritization of cancer neoantigens. As the field of neoantigen-based immunotherapies progresses, there is a growing need to validate these neoantigens to determine if they are generating tumor-specific immune responses in patients, and incorporate this information into neoantigen prediction algorithms. The ability of T cells to specifically recognize neoantigens is mediated by their T cell receptors (TCRs). TCRs are highly diverse and are generated through a process of genetic rearrangement called V(D)J recombination during T cell development. This diversity enables T cells to adaptively recognize many antigens, including neoantigens as foreign. The binding of such TCRs to MHC-presented neoantigens triggers a cascade of immune responses, leading to the killing of tumor cells.

In this study, we aimed to explore immune responses in patients receiving personalized cancer vaccines by monitoring clonally expanded TCR sequences in peripheral blood mononuclear cells (PBMCs) before and after neoantigen vaccine treatment. To assess the immune response to neoantigen vaccines, blood samples were collected from patients prior to, and during treatment with neoantigen vaccines. PBMCs were stimulated in-vitro in separate pools using the peptides included in the patient's neoantigen vaccine. Bulk TCR sequencing was used to identify clonally expanded TCRs in each peptide stimulated pool. Subsequently, single cell RNA sequencing (scRNAseq) and single cell TCR sequencing (scTCRseq) enabled determination of the full-length TCR sequences for the expanded clones, and their corresponding cell types, offering further insights into the CD4 and CD8 T cell mediated responses. Using this approach, we identified clonally expanded T cells from neoantigen vaccine-treated patient PBMCs for multiple peptides used in the vaccine. By applying this approach at scale, we aim to compare peptides that give rise to expanded clones against those that do not, to identify characteristics that may contribute to making a candidate neoantigen immunogenic and incorporate this information into pVACtools.

Oral Session 12 - Trainee Lightning Talks session 2

pVACsplice: predicting and prioritizing tumor-specific splicing antigens

Authors: My Hoang, Miller Richters, Susanna Kiwala, Jeffrey P Ward, Ramaswamy Govindan, Obi L Griffith, Malachi Griffith
Submitter: My Hoang
Submitter Email: hmy@wustl.edu

Personalized anti-cancer treatments such as cancer vaccines often target tumor-specific neoantigens originating from SNPs and indels. Alternative splicing of mRNAs may also create highly valuable neoantigens that, to date, have been less frequently exploited.

Most existing tools that perform splicing-derived neoantigen discovery rely exclusively on RNA-seq data and identify a myriad of novel aberrant splicing events. Given that alternative splicing is often tissue-restricted, it is difficult to verify that most of these events are truly tumor-specific and thus, good targets for immunotherapy. To address this, we developed pVACsplice, a open-source, open-license Python package that predicts and prioritizes neoantigens derived from aberrant splicing events specifically caused by somatic mutations.

This work leverages our experience developing RegTools, a computational method for identifying novel junctions emerging from cis-splicing mutations from tumor DNA and RNA. pVACsplice takes RegTools tumor-specific transcripts as input, translates these transcripts to mutated proteins, and extracts mutated epitopes of a user-defined length. It then predicts the binding affinity of these epitopes to MHC alleles, along with a host of other annotations including: their immunogenicity, eluted ligand likelihood, probability to be processed by the proteasome, similarity to reference proteome, and whether they fall within manufacturing constraints. pVACsplice is highly configurable to meet a multitude of use cases, providing settings for novel splice junction coverage cutoff, type of junction to use for predicting epitopes, and thresholds for VAF, coverage, and gene expression. Results are prioritized to select the best neoantigen candidate for each splice site, taking into consideration criteria such as the binding affinity, transcript support level, and biotype. Each selected neoantigen candidate is then tiered into a category depending on its suitability for vaccine manufacturing.

As a proof-of-concept, we performed epitope discovery with pVACsplice on a small cell lung cancer (SCLC) cohort. The analysis of 60+ patients revealed multiple epitopes associated with genes known to be involved in SCLC pathogenesis such as TP53, RB1. This exercise suggests that cis-splicing mutations represent a valuable source of neoantigen candidates for cancer immunotherapies.

Significance: pVACsplice predicts neoantigens from a selective, high-quality splicing source. pVACsplice is a part of the pVACtools suite - a collection of bioinformatic tools facilitating target discovery and production of cancer immunotherapy (pvactools.org). pVACsplice integrates tightly with other methods in pVACtools suite,creating a comprehensive pipeline for identifying, prioritizing, and manufacturing neoantigen-targeting immunotherapies.

Oral Session 12 - Trainee Lightning Talks session 2

Building Contextual m6A Knowledge Graph in Cancer through Literature Analysis with reguloGPT

Authors: Xidong Wu1, Sumin Jo1,2, Yiming Zeng2, Arun Das2,3, Ting-He Zhang2,3, Parth Patel4, Yuanjing Wei5, Lei Li5, Shou-Jiang Gao2,6, Jianqiu Zhang4, Dexter Pratt7, Yu-Chiao Chiu2,3, Yufei Huang1,2,3
Submitter: Yufei Huang
Submitter Email: sumin.Jo@pitt.edu

N6-methyladenosine (m6A) is the most common mRNA modification in humans, significantly influencing cellular processes such as mRNA stability, splicing, and translation. Its role extends to normal physiological functions and complex disease states, particularly cancer. Recent studies emphasize m6A's critical involvement in cancer progression, including tumor initiation, growth, and resistance to treatments, as well as its effects on the tumor microenvironment and immune response. Dysregulation of m6A also contributes to cancer metastasis by facilitating tumor invasion and spread. Understanding the Molecular Regulatory Pathways (MRPs) associated with m6A in cancer is essential for identifying potential therapeutic targets and developing effective treatments. It is crucial to consider how m6A modifications influence various developmental processes like embryonic development, tissue maturation, and organogenesis. Despite these insights, the dynamic and context-dependent nature of m6A's molecular regulatory mechanisms in cancer is not yet fully understood. Knowledge Graphs (KGs) are instrumental in organizing and analyzing MRPs, offering structured representations of complex biological interactions. Current tools for extracting KGs from biomedical literature struggle with capturing complex, hierarchical relationships and contextual information about MRPs. While recent advances in Natural Language Processing (NLP) have automated some aspects of this process, NLP techniques still face challenges with the complexity and diversity of MRPs and often lack critical contextual details. Large Language Models (LLMs) like GPT-4 present a promising solution due to their advanced language processing capabilities. However, their potential for end-to-end KG construction, especially for MRPs, remains largely untapped. We introduce reguloGPT, a novel GPT-4 based in-context learning prompt designed for extracting regulatory graphs and context from sentences describing regulatory interactions. reguloGPT creates a context-aware relational graph that effectively represents the hierarchical structure of MRPs and resolves semantic inconsistencies by embedding context directly within relational edges.We began by developing an annotated dataset of 400 titles related to m6A's MRPs to evaluate initial performance. Rigorous testing of reguloGPT on this benchmark dataset demonstrated significant improvements over existing algorithms and other LLMs. We also developed a novel G-Eval scheme using GPT-4 for annotation-free performance evaluation, which aligned with benchmark evaluations. Applying our approach to 969 additional titles resulted in the m6A Molecular Regulatory Pathway Knowledge Graph (m6A-MRP-KG), which elucidates m6A's regulatory mechanisms across various contexts. This KG enhances our understanding of m6A's roles in gene expression regulation, cancer, and immunity, offering new insights and potential therapeutic strategies. The results highlight reguloGPT's transformative potential for extracting valuable biological knowledge from literature.

Poster Presentation

Expanding Genetic Test Result Delivery with a Hybrid Rule-Based/Large Language Model Chatbot for Return of Positive Results

Authors: Emma Coen, Guilherme Del Fiol, Kim Kaphingst, Caitlin Allen
Submitter: Emma Coen
Submitter Email: ecoen@g.clemson.edu

Background: Delivery of positive genetic testing results poses emotional challenges for patients and resource burdens for healthcare providers. Chatbots could be helpful to return results because they can provide timely, consistent, and clear communication while potentially alleviating some of the workload on healthcare providers

Objective: The objective of this presentation is to explain how we plan to leverage the ITCR-funded Genetic and Risk-stratified Detection Evaluation (GARDE) chatbot authoring platform to provide educational and support for individuals identified with as having positive genetic testing results

Methods: We explore patient perspectives on receiving positive genetic testing results through a chatbot through semi-structured qualitative interviews. In addition, we completed prompt engineering to develop a hybrid rule-based/large language model (LLM) chatbot using the GARDE platform that is capable of delivering positive results in a scientifically accurate and empathetic manner. Prompt engineering was conducted within OpenAI's GPT4 Playground environment and evaluated by testing the LLM components with hypothetical cases, where experts rated responses across several domains to ensure chatbot's effectiveness

Results: Through semi-structured interviews with 22 participants from HBOC and Lynch Syndrome communities, the research sheds light on the potential of chatbots to improve genetic test result delivery. Key findings highlight enhanced patient understanding and emotional support, as well as a positive impact on decision-making processes. Prompt engineering results indicate improved response accuracy and empathy, with quantitative findings demonstrating increased patient comprehension and satisfaction as evidence by higher ratings in clarity, tone, and quality

Conclusion: Recommendations are provided for scaling chatbot implementation and integrating patient feedback to ensure continuous improvement. By offering insights into the roles of prompt engineering and patient feedback in return of positive genetic testing results, we leverage the existing GARDE platform to provide valuable guidance for integration into practice. Future work includes using the GARDE chatbot authoring tool to integrate the resulting LLM-based chatbot within a hybrid chatbot that uses a pre-defined script to cover foundational content about positive genetic test results and allows patients to ask questions that are redirected to the LLM described above. This hybrid chatbot will be used to support the delivery of positive test results to patients receiving whole genome sequencing as a part of the In Our DNA SC program led by the Medical University of South Carolina (MUSC).

Poster Presentation

Developing a Statistical and Visualization Tool for Cancer Registries to Detect Cancer Hot Spots for Small Geographic Areas

Authors: Jacob Oleson
Submitter: Jacob Oleson
Submitter Email: jacob-oleson@uiowa.edu

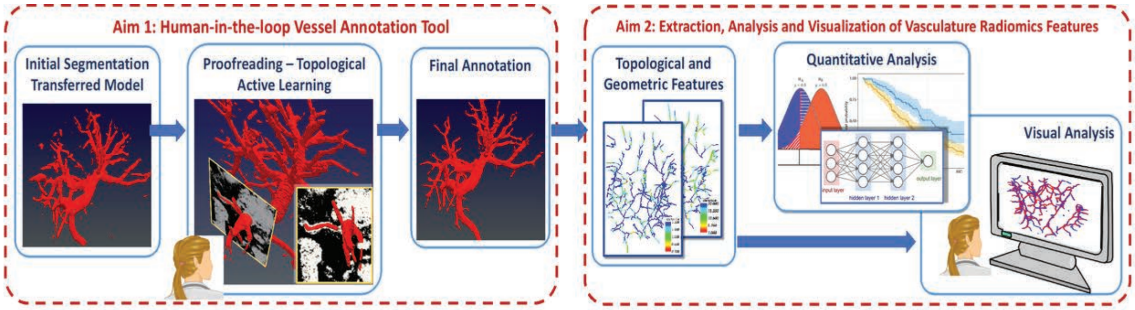
Many central cancer registries are asked to present cancer data (rates, counts) for sub-state geographic areas, and to identify hot spots for cancer risk in their states. Yet, it is challenging to calculate reliable cancer rates for geographic areas with small case counts (e.g., townships, rural counties) because computing rate and risk estimates with small counts traditionally results in rates that are unstable and have very large standard errors. We have developed a Bayesian hierarchical model and visualization tool that allows us to estimate and show cancer rates and cancer risk in geographic regions with small case counts (e.g., Zip-Code Tabulation Area). The cancer rates and cancer risk include both incidence and mortality for eight selected cancers. Our methodology also calculates and quickly shows where significant clusters of high and low cancer prevalence are located. We started this project examining data for Iowa and have extended this work through collaboration with the New Mexico Tumor Registry and the Kentucky Cancer Registry. The age-adjusted cancer rates generated by the model, as well as probability of risk, are displayed in interactive maps that will be accessible broadly and for free on the web. Our interactive visualization tools allow the user to switch between these various maps depending on the type of data that they are interested in viewing. In this presentation, we will explore the mapping tool, demonstrate results from the different states, and highlight its potential use for central cancer registries and beyond. Ultimately, these maps allow users to compare risk for cancer where they live to other cities in the state and identify areas with high risk in need of interventions. We are hoping to make this tool more broadly accessible to central cancer registries, especially those who struggle visualizing data with small case counts.

Topological uncertainty for vascular segmentation

Authors: Saumya Gupta, Prateek Prasanna, Chao Chen
Submitter: Saumya Gupta
Submitter Email: saumgupta@cs.stonybrook.edu

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.


Aim 1: Human-in-the-loop Vessel Annotation Tool
Initial Segmentation Transferred Model
Proofreading – Topological Active Learning
Final Annotation
Aim 2: Extraction, Analysis and Visualization of Vasculature Radiomics Features
Topological and Geometric Features
Quantitative Analysis
input layer
hidden layer 1
hidden layer 2
output layer
Visual Analysis

Overview of our project: Characterization of the vessel structures using novel topological and geometry-informed radiomics. Development of a risk scoring system using topology and geometry radiomics to predict response to immunotherapy in lung cancer.

Abstracts

Poster Presentation

EMERSE (the text processing/searching tool) updates for 2024

Authors: David Hanauer, Lisa Ferguson, Kellen McClain, Guan Wang
Submitter: David Hanauer
Submitter Email: hanauer@umich.edu

EMERSE (https://project-emerse.org) is a text processing and searching system designed for analyzing free text clinical notes from electronic health records (EHRs). EHR notes contain vital data necessary for multiple research tasks, ranging from patient cohort identification to data abstraction for case report forms. EMERSE was designed to be user-friendly and easily accessible by non-technical researchers.

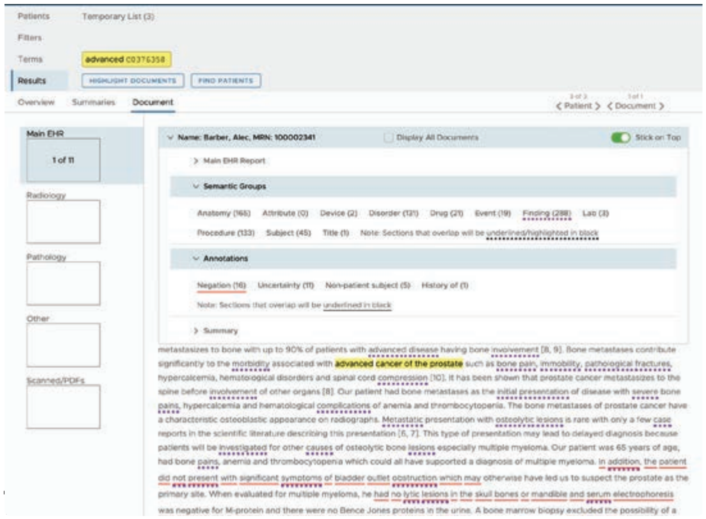
Our upcoming release (expected in summer 2024) will have built-in natural language processing (NLP) capabilities including annotations for negation, uncertainty, subject (patient versus non-patient) and prior history. With these annotations and an updated user interface, users can search for detailed information such as instances where it is specifically mentioned that a family member did not have breast cancer, or notes where a patient had prostate cancer, excluding mentions that are negated, uncertain, or mentioned in the context of a prior history. EMERSE also will support labeling of (and searching for) terms based on the Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs). Incorporating this functionality while maintaining ease-of-use, as well as fine-grained-control for users, has required substantial changes to the user interface.

We have also recently completed an application programming interface (API) that allows for audited, back-end access to clinical notes in a HIPAA-compliant manner (e.g., audit trails of all queries are maintained). This has been beneficial to research teams building their own data analytic pipelines, including those training artificial intelligence (AI) models.

Our team focuses our work on two primary groups that we interact with: (1) end-user researchers that want to use the software and (2) information technology/informatics teams that must implement the enterprise software locally at their institutions. These two groups have very different needs and expectations. In addition to a user guide, we have multiple technical implementation and ongoing operational guides. We have also developed a general security document for distribution that contains answers to many questions asked during software security reviews, which are now almost universally required by implementing sites.

EMERSE is now live at 12 sites and is undergoing implementation at 7 more (see https://project-emerse.org/community.html). Additionally, EMERSE has supported over 670 peer-reviewed papers and abstracts. In addition to completing the NLP work for our next release, we are also working to (1) enable the incorporation of customized NLP labeling in lieu of, or in addition to, what EMERSE provides out-of-the-box; (2) better understanding of security concerns about our network feature so that we can address these in a future release; (3) surveying end users to better understand how they have been using EMERSE and the kind of work it has supported; and (4) a process for extracting data from templated notes. We are also beginning to explore how large language models might be incorporated to provide novel capabilities for our user base.

Acknowledgement: This was work supported by the National Cancer Institute under Award U24CA269315.



metastasizes to bone with up to 90% of patients with advanced disease having bone involvement [8, 9]. Bone metastases contribute significantly to the morbidity associated with advanced cancer of the prostate such as bone pain, immobility, pathological fractures, hypercalcemia, hematological disorders and spinal cord compression [10]. It has been shown that prostate cancer metastasizes to the spine before involvement of other organs [8]. Our patient had bone metastases as the initial presentation of disease with severe bone pains, hypercalcemia and hematological complications of anemia and thrombocytopenia. The bone metastases of prostate cancer have a characteristic osteoblastic appearance on radiographs. Metastatic presentation with osteolytic lesions is rare with only a few case reports in the scientific literature describing this presentation [6, 7]. This type of presentation may lead to delayed diagnosis because patients will be investigated for other causes of osteolytic bone lesions especially multiple myeloma. Our patient was 65 years of age, had bone pains, anemia and thrombocytopenia which could all have supported a diagnosis of multiple myeloma. In addition, the patient did not present with significant symptoms of bladder outlet obstruction which may otherwise have led us to suspect the prostate as the primary site. When evaluated for multiple myeloma, he had no lytic lesions in the skull bones or mandible and serum electrophoresis was negative for M-protein and there were no Bence Jones proteins in the urine. A bone marrow biopsy excluded the possibility of a

Screen shot showing one term in yellow entered as a normal term ("advanced") combined with a CUI for prostate cancer (C0376358). Terms identified with a dotted purple underline are UMLS Findings (semantic type), and terms underlined in red are considered negated.

Poster Presentation

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

Authors: Zinaida Perova, Mauricio Martinez, Tushar Mandloi, Marcelo Rios Almanza, Steven Neuhauser, Dale Begley, Debra Krupke, Carol Bult, Helen Parkinson
Submitter: Zinaida Perova
Submitter Email: zina@ebi.ac.uk

CancerModels.Org (www.cancermodels.org) is a research platform that standardizes, harmonizes, and integrates the complex and diverse data associated with Patient-Derived Cancer Models (PDCMs) for the cancer community - including clinical, genomic, and functional data from patient-derived xenografts (PDX), organoids and cell lines.

Users can search for models of interest via a web interface or the REST API and explore molecular data summaries for models of specific cancer types. The underpinning data model was augmented with additional dimensions and covers gene expression, gene mutation, biomarkers, images, immune markers (TMB, MSI, HLA), patient treatment, and drug dosing studies. Moreover, the knowledge is enriched with links to external resources - publication platforms, cancer-specific annotation tools (COSMIC, CIViC, OncoMX, OpenCRAVAT, ClinGen), and raw data archives (EGA, dbGAP, GEO). In addition, all our data is available via cBioPortal instance and can be accessed via cloud analysis platforms, such as Terra and Cancer Genomics Cloud.

We provide expertise and software components to support several worldwide consortia, including PDXNet (https://www.pdxnetwork.org/) and EurOPDX (https://europdx.eu/). In collaboration with OICR (https://oicr.on.ca/) we built a user-facing Metadata dictionary and Validator (https://www.cancermodels.org/validation/dictionary) to facilitate model submission to CancerModelsOrg. Our software is freely available under an Apache 2.0 licence (https://github.com/PDCMFinder).

Future work will focus on the addition of rare models with rich accessible metadata and data, particularly drug-resistant cohorts, new data visualizations, integration of new data types and resources, as well as devising a model quality rating using user feedback and model availability information. In collaboration with DeepPhe project (https://deepphe.github.io/), we will assess the capabilities of Large Language Models to extract PDCM-relevant knowledge from publications. This could enable a more streamlined and efficient content acquisition, thus expanding resource with well-described and richly annotated PDCMs. These developments will maximize utility and improve reusability of models and data, and reduce barriers to model and data sharing.

In conclusion, CancerModels.Org aggregates over 8400 models across 13 cancer types, including rare pediatric PDX models (13%), and models from minority ethnic backgrounds (47%), totaling over 252B data points across a variety of data types: clinical metadata, molecular data, and treatment-based information. CancerModels.Org is the largest free-to-consumer and open-access resource of this kind.

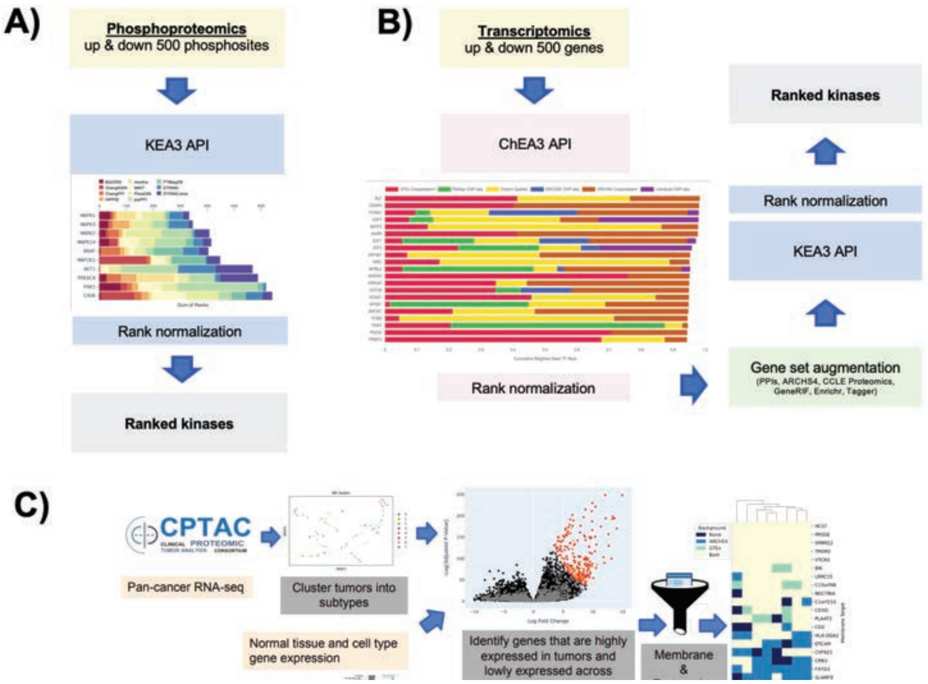
Abstracts

Poster Presentation

Multiomics2Targets: Computational Workflow to Identify Targets for Cancer Cohorts Profiled with Transcriptomics, Proteomics, and Phosphoproteomics

Authors: Giacomo Marino, Eden Deng, Daniel Clarke, Ido Diamant, Avi Ma'ayan
Submitter: Avi Ma'ayan
Submitter Email: avi.maayan@mssm.edu

The availability of data from the profiling of cancer patients with multiomics technologies is rapidly increasing. However, integrative analysis of such data for knowledge extraction and practical hypotheses generation for clinical applications is not trivial. Recently, we developed Multiomics2Targets, a bioinformatics workflow that enables users to upload transcriptomics, proteomics, and phosphoproteomics data matrices as well as accompanying clinical data collected from the same cohorts of cancer patients. After uploading the data, Multiomics2Targets produces a report that resembles a research publication. The uploaded data matrices are processed, analyzed, and visualized using several tools related to the ARCHS4 resource to produce ~80 figures and ~30 tables. Figure and table legends, as well as descriptions of the methods and results are provided. The reports include an abstract, an introduction, methods, results, discussion, conclusions, and references sections. Multiomics2Targets reports can be exported as PDF or Jupyter Notebooks, and can be cited. Additionally, since the pipeline is implemented as a Jupyter Notebook, the source code used to perform the analysis and produce the report is embedded within the report and can be easily viewed, modified, and run locally. Multiomics2Targets can be used to perform alternative analyses when only one or two omics datasets are uploaded. Importantly, we applied Multiomics2Targets to analyze the multomics dataset from the CPTAC3 pan-cancer cohort. Multiomics2Targets is available at: https://multiomics2targets.maayanlab.cloud/.


A)
Phosphoproteomics
up & down 500 phosphosites
KEA3 API
Rank normalization
Ranked kinases
B)
Transcriptomics
up & down 500 genes
ChEA3 API
Rank normalization
Gene set augmentation
(PPIs, ARCHS4, CCLE Proteomics, GeneRIF, Enrichr, Tagger)
KEA3 API
Rank normalization
Ranked kinases
C)
CPTAC
CLINICAL PROTEOMIC TUMOR ANALYSIS CONSORTIUM
Pan-cancer RNA-seq
Cluster tumors into subtypes
Normal tissue and cell type gene expression
Log Fold Change
Identify genes that are highly expressed in tumors and lowly expressed across
Membrane &

Workflows for eXpression2Kinases pathway recovery and the target identification process
A) In the transcriptomics component of the X2K workflow, sets of 500 up- and down- genes are submitted to ChEA3 for transcription factor enrichment analysis, followed by gene set augmentation using one of 6 methods. The expanded gene set is then submitted to KEA3 to rank the most likely kinases upstream of the observed transcriptional changes. B) In the phosphoproteomics component of the workflow, sets of 500 up- and down- phosphosites are submitted to KEA3 for kinase enrichment analysis to rank the most likely kinases responsible for the observed phosphoproteomic changes. The top-ranked kinases from the transcriptomic and phosphoproteomic pipelines are then compared for agreement to recover likely signaling pathways. C) For target identification, CPTAC3 pan-cancer transcriptomics data is first analyzed by the Leiden algorithm to identify clusters for each cancer type. At the same time, gene expression from normal tissues and cell types is compiled from GTEx and ARCHS4. Each cluster of samples from each cancer subtype is compared to the normal tissues and cell types from the two background atlases to identify genes that are highly expressed in the tumors. The results are visualized at box plots and volcano plots. Top ranked genes are then filtered for including only genes that give rise to transmembrane proteins using proteomics resources: COMPARTMENTS, CPTAC3 proteomics, HPA and HPM. The results across clusters are visualized as heatmaps.

Abstracts

Poster Presentation

Data Explorer 2.0: a more powerful tool to interrogate the Cancer Dependency Map

Authors: Joshua Dempster, Randy Creasi, Yvonne Blanco, Barbara De Kegel, Mustafa Kocak, Francisca Vazquez, Philip Montgomery, Caterina Campbell
Submitter: Joshua Dempster
Submitter Email: dempster@broadinstitute.org

The mission of the Cancer Dependency Map is to accelerate cancer discoveries across the research community. To advance this goal, the Cancer Data Science team works not just to provide data but also public tools to analyze it through the DepMap Portal. These data include genome-wide CRISPR screens, multimodal genomics profiling, and drug screens. To help users develop insights from these datasets, we recently released Data Explorer 2.0, a general purpose and flexible plotting tool for making visual comparisons in DepMap data. Today, Data Explorer 2.0 allows users to define, save, and aggregate custom model cohorts and gene sets, identify the defining features and vulnerabilities of a cohort, and look deeper into results using additional plot types and customization. Current work is focused on streamlining workflows while introducing even more powerful features. For example, users will be able to compare the same model in different growth conditions, examine sgRNA-level data, and analyze combinatorial screens. Through these efforts, we will enable users to rapidly identify relationships in DepMap at any scale.

Updates from The Cancer Proteome Atlas: a new platform for animal model data and the continued chatbot development

Authors: Jun Li, Wei Liu, Yitao Tang, Yiling Lu, Han Liang
Submitter: Han Liang
Submitter Email: hliang1@mdanderson.org

Reverse-phase protein arrays (RPPAs) represent a powerful functional proteomic approach to elucidate cancer-related molecular mechanisms and develop novel cancer therapies. To facilitate community-based investigation of the large-scale protein expression data generated by this platform, we have developed a user-friendly, open-access bioinformatic resource, The Cancer Proteome Atlas (TCPA, http://tcpaportal.org), which contains several web applications. In addition to these web applicaitons/datasets focusing on human samples (patients or cell lines), we are developing a new web application focusing on animal model samples, which includes >10000 samples (including PDX, patient-derived xenograft, CDX, cell line-derived xenograft, and genetically manipulated mice GEMM). This resource would greatly facilitate the translation of large functional proteomic datasets to testable hypotheses for experimental and clinical investigations. To further address the informatic challenges of analyzing diverse datasets in TCPA, we continue to develop the LLM-based chatbot (TCPAplus) through which users can analyze the RPPA data through human nature languages and obtain the results and related analytic reports without a learning curve. TCPAplus empowers a broad research community to explore high-quality RPPA datasets and generate testable hypotheses in an effective and intuitive manner, representing the direction of next-generation data analytics.

Uncovering Druggable Vulnerabilities in Cancer with the AVERON Notebook

Authors: Hongyue (Nicole) Chen, Brian Revennaugh, Haian Fu, Andrey Ivanov
Submitter: Andrey Ivanov
Submitter Email: andy.a.ivanov@gmail.com

Genomic alterations, such as missense mutations, often lead to the activation of oncogenic pathways and cell transformation by rewiring protein-protein interaction (PPI) networks, leading to the acquisition of cancer hallmarks. The discovery of mutant-directed PPIs may uncover new mechanisms of oncogenic signaling and provide new targets for personalized therapeutic interventions in cancer. The advances in high-throughput screening technologies and computational approaches enabled comprehensive profiling of mutant-dependent PPIs in cancer cells. However, elucidation of functional consequences of mutant-induced changes in PPI networks and their impact on clinical outcomes of cancer patients remains highly challenging. To address this challenge, we develop a new 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 mutant-induced neomorph PPIs (neoPPIs). Based on experimentally determined or computationally predicted networks of mutant-directed neoPPIs, AVERON employs specially designed algorithms and statistical techniques to assess the levels of PPIs in cancer patients in terms of PPI scores. It examines and visualizes neoPPI impact on clinical outcomes and identifies neoPPI-regulated distinctive sets of signature genes and oncogenic pathways. Furthermore, the AVERON can uncover clinically significant and druggable neoPPI-regulated genes to target cancer dependency on mutant-directed PPIs. 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 new target and therapeutic development in cancer.

Abstracts

Poster Presentation

Inclusion of new clinically actionable variant types into the CIViC data model

Authors: Arpad Danos, Kilannin Krysiak, Jason Saliba, Adam Coffman, Susanna Kiwala, Joshua McMichael, Mariam Khanfar, Cameron Grisdale, Malachi Griffith, Obi Griffith
Submitter: Arpad Danos
Submitter Email: arpaddanos@gmail.com

The integration of next generation sequencing into clinical practice for cancer patients has resulted in the emergence of a treatment bottleneck, where interpretation of large numbers of tumor variants is required. One approach to this problem is to offer private or siloed datasets accessible behind paywalls or exclusive to institutions. A different approach, utilized by the CIViC knowledgebase (www.civicdb.org) is to leverage public curation and expert moderation of cancer variant data, housed in a public web resource, with no restrictions on the data use, no signin required for data access, and a public API. In CIViC, data is curated from publications and meeting abstracts into a cancer variant data model consisting of free text and structured fields, and comprised of six evidence types (Predictive, Diagnostic, Prognostic, Predisposing, Oncogenic, and Functional). Evidence is also curated into summary Assertion statements, which reflect the state of what is known for a given variant, and also incorporate field-wide guidelines such as those from AMP/ASCO/CAP (Li et al, 2017) for tiering of clinically actionable somatic variants, or those from ClinGen/CGC/VICC (Horak et al, 2022) for evaluation of somatic variant oncogenicity.

Since the inception of CIViC, there have been multiple updates to the data model, including addition of new evidence types, incorporation of published guidelines, and a large update adding Molecuar Profiles (MPs), which enable evidence to be associated to complex combinations of variants from different genes. The data model for CIViC variants is highly flexible, and curators can specify variants for a gene at the genomic DNA level, using mRNA, or specificed at the level of changes to amino acids, such as BRAF V600E. Variants are also curated using categorical notation (e.g. "bucket" variants), which collect together groups of variants associated to a gene, such as EGFR Exon 20 Insertions. Despite this variant flexibility, there is a need to curate variants which are not directly associated to a specific gene.

One class of variants of this type are a set of genomic phenomena associated with clinical impact, such as Tumor Mutational Burden (TMB), which can have predictive value for checkpoint inhibition in some cancers. Other examples include Microsatellite instability (MSI), which has prognostic value in cancers including colorectal cancer, and the presence of HPV infection, which has clinical importance in multiple cancer types. In order to incorporte these types of evidence, CIViC has introduced the Factor data type, with instances of Factors being TMB, MSI, HPV, and others such as kataegis, complex karyotype, and chromothripsis. Evidence Items and Assertions may be associated to variants of these new types. Each Factor is associated to a corresponding term from the NCI Thesaurus. Curators may create complex MPs consisting of Gene and Factor type variants. Genes and Factors are together grouped as different instances of Features in the new CIViC data model, and additional types of Features, such as Regions, and an updated model for Fusions, are currently under development.

Integrating multi-omics analyses into agent-based models with the Bioinformatics Walkthrough to predict evolution of pancreatic ductal adenocarcinoma

Authors: Daniel Bergman, Paul Macklin, Elana Fertig
Submitter: Daniel Bergman
Submitter Email: dbergma5@jh.edu

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.

Poster Presentation

Extracting Social Determinants of Health from Pediatric Patient Notes Using Large Language Models: Novel Corpus and Methods

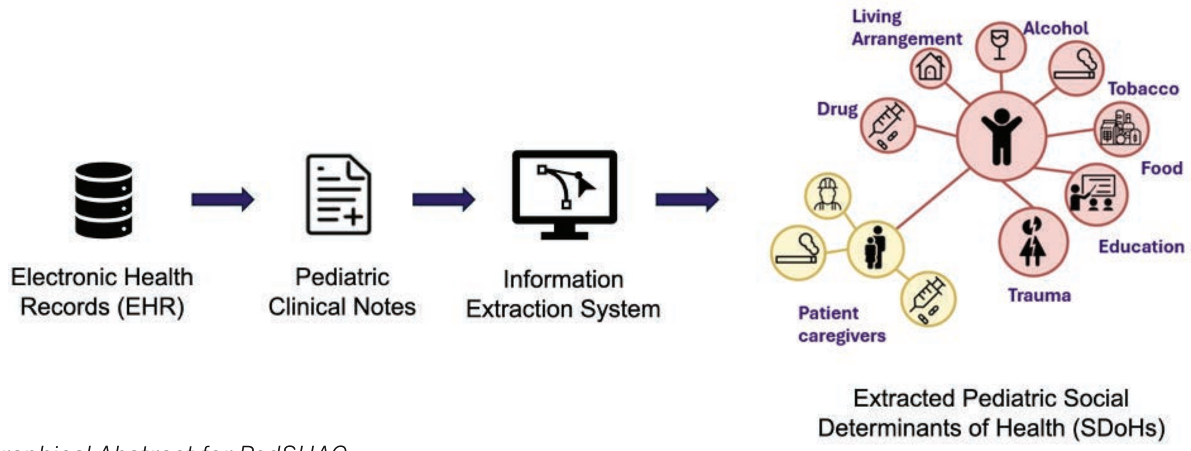
Authors: Yujuan Fu, Giridhar Ramachandran, Nicholas Dobbins, Namu Park, Michael Leu, Abby Rosenberg, Kevin Lybarger, Fei Xia, Ozlem Uzuner, Meliha Yetisgen
Submitter: Meliha Yetisgen
Submitter Email: melihay@uw.edu

Objective: In the United States, around 9.620 children under the age of 15 will be diagnosed with cancer in 20224. Although cure-rates for children with some types of cancer have advanced from 20% in the 1970s to 80% in the 2000s, we have seen little improvement in the past decade. Furthermore, the benefits of these advancements are inequitably distributed, with well-documented differences in outcomes across racial and ethnic groups that may be attributable to the enduring effects of social inequities. For these reasons, a current goal of pediatric oncology clinical research is to identify and target previously understudied factors affecting the outcomes of children with cancer, including "nonclinical" factors such as Social Determinants of Health (SDOH). SDOH covers conditions in which people work and live, which significantly influence health outcomes, especially in pediatric populations where early interventions can yield long-term benefits. However, the variability, rarity, and inconsistent documentation practices of SDoH, primarily recorded as unstructured text in Electronic Health Records (EHRs), hinder their effective application in health interventions. Therefore, this study focuses on the annotation and automatic extraction of structured, fine-grained, and comprehensive representations of pediatric SDoH events from unstructured clinical notes.

Materials and Methods: In this work, we present a novel annotated corpus, the Pediatric Social History Annotation Corpus (PedSHAC), and evaluate the automatic extraction of detailed SDoH representations using Large Language Models (LLMs). We explored LLM-based information extraction (IE) across multiple dimensions, including pretrained architectures – mSpERT, Flan-T5, and GPT-4; learning strategies – fine-tuning and in-context methods; and prompting approaches – one-step text-to-event and two-step question-answering(QA).

Results: PedSHAC comprises annotated social history sections from 1,260 clinical notes obtained from pediatric patients within the University of Washington (UW) hospital system. Employing an event-based annotation scheme, PedSHAC captures ten distinct health determinants to encompass living and economic stability, prior trauma, education access, substance use history, and mental health with an overall annotator agreement of 81.9 F1. Our proposed fine-tuning LLM-based extractors achieve high performance at 78.4 F1 for event arguments. In-context learning approaches with GPT-4 demonstrate promise for reliable SDoH extraction with limited annotated examples, with extraction performance at 82.3 F1 for event triggers.

Conclusions: We introduce PedSHAC, a novel corpus with fine-grained annotations for 10 Pediatric SDoH events from social history sections of unstructured clinical notes. Our experiments with state-of-the-art transformer models demonstrate that detailed SDoH representations can be extracted from pediatric narratives with performance comparable to that of human annotators, providing an automatic approach for incorporating valuable SDoH information into clinical and research applications. We envision such fine-grained information across multiple critical SDoH types can help the research community study the impact of SDOH on other child health outcomes.


Electronic Health Records (EHR)
Pediatric Clinical Notes
Information Extraction System
Living Arrangement
Alcohol
Tobacco
Drug
Food
Education
Trauma
Patient caregivers
Extracted Pediatric Social Determinants of Health (SDoHs)

Graphical Abstract for PedSHAC

Abstracts

Poster Presentation

Developing a Rule-Based Algorithm to Identify Recurrent Non-Hodgkin Lymphoma in Electronic Health Data

Authors: Mara Epstein, Feifan Liu, Laura Susick, Yanhua Zhou, Shane Bole, Lydia Goldthwait, Wendy Haykus, Muthalagu Ramanathan, George Divine, Christine Johnson
Submitter: Mara Epstein
Submitter Email: mara.epstein@umassmed.edu

Background: Recurrent cancers are not captured in a standardized way by US tumor registries. We aim to develop rule-based and machine learning-based algorithms to identify recurrent cases of two common subtypes of non-Hodgkin lymphoma (NHL), diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL) in electronic health record (EHR) data. Here we present findings from the rule-based algorithm. Our goal is to create a tool with a positive predictive value (PPV) greater than 75% to identify recurrent cases for population-based research.

Methods: Incident DLBCL and FL cases diagnosed between 2000-2018 were identified in tumor registry data from the Virtual Data Warehouse (VDW), a distributed standardized repository of EHR and claims-based datasets housed at two study sites participating in the Health Care Systems Research Network (HCSRN; Site A: claims only; Site B: claims and EHR data). We compiled a comprehensive list of pharmacy and procedure codes to indicate when each patient started first-line treatment, capturing 15 dispensed drugs, the administration of those drugs, and related processes. We defined recurrent cases as those who completed first-line treatment followed by a gap of ?6 months with no treatment-related codes, but who later restarted treatment. The algorithm was built using data from Site A and tested at Site B. Results were validated by chart review of 166 randomly selected EHRs from site B, which were entered into a REDCap-based abstraction form. Measures of validity were calculated for site B. The algorithm was then revised specifically to reduce the false positive rate.

Results: Site A identified 225 patients (137 DLBCL and 88 FL). The mean age at diagnosis was 70.7 years for DLBCL and 66.5 for FL. Mean follow-up time post-diagnosis was 3.8 years for DLBCL and 5.2 years for FL. Twenty-three DLBCL and 19 FL patients met criteria for recurrent disease, with a mean of 3.4 years from diagnosis to first recurrence. At site B, 392 patients were identified (246 DLBCL and 146 FL). The mean age at diagnosis was 65.7 for DLBCL and 63.6 for FL. Mean post-diagnosis follow-up time was 5.8 years for DLBCL and 8.4 years for FL. Forty-nine DLBCL and 48 FL patients met criteria for recurrent disease. Chart review at Site B determined a sensitivity of 94%, specificity of 45%, positive predictive value (PPV) of 52% and negative predictive value (NPV) of 92% for the first algorithm. Following revisions to reduce the false positive rate, the algorithm had reduced sensitivity (66%) and NPV (75%), but higher specificity (88%) and PPV (82%).

Conclusions: The number of recurrent cases identified by the algorithm are in line with clinical expectations. We have developed two algorithms that may be applied to EHR data for population-based research into recurrence of two common subtypes of NHL depending on the research question. Our revised algorithm met our a priori goal for PPV with the intent of application to large population-based databases. As a next step, the algorithms will be applied to EHR data at a third HCSRN site for further assessment.

Abstracts

Poster Presentation

Preclinical Imaging XNAT-Enabled Informatics (PIXI): An open-source resource to support cloud-based computational workflows for preclinical imaging

Authors: Andrew W. Lassiter, Stephen M. Moore, James D. Quirk, Richard Laforest, William Horton, Daniel S. Marcus, Kooresh I. Shoghi
Submitter: Andrew Lassiter
Submitter Email: andrewl@wustl.edu

Preclinical imaging workflows have been growing in complexity, data size, and analytic requirements. We provide an update on our efforts to develop an open-source preclinical imaging XNAT-enabled informatics (PIXI) platform to manage the workflows of preclinical image data acquisition, to capture imaging-associated experiments including metadata and annotations, and to implement computational pipelines in a unified environment. Our vision for PIXI extends beyond the initial implementation to support a federated network of PIXI instances and a PIXI Center to enable data sharing and collaboration across institutions.

PIXI is based on the widely used XNAT platform as the underlying informatics architecture. PIXI includes: the PIXI Server, which provides core database, visualization, and workflow functionality; PIXI Notebooks for data exploration and analysis; and PIXI Apps to enable automated image processing pipelines through Docker container environment. With the recent release of PIXI 1.0 in February 2024, preclinical positron emission tomography (PET), computed tomography (CT) and magnetic resonance (MR) DICOM images are pushed to the PIXI server for workflow management with metadata captured through the PIXI Web-UI and DICOM image files for search and reporting. Multi-mouse images are supported by form-based data entry and Docker pipeline that splits hotel images into single-mouse datasets. Imaging workflow information can be entered or edited through the Web-UI. In addition, PIXI includes workflows to support upload and management of native Inveon PET and CT images and IVIS bioluminescence (BLI) images. XNAT's search and reporting capabilities have been extended to support PIXI's new data types and metadata. Importantly, we collaborated with the Open Health Imaging Foundation (OHIF) to expand the OHIF viewer's capabilities to support 4-dimensional image visualization and analysis of preclinical images. Finally, Jupyter notebooks has been integrated into the PIXI platform, providing researchers with seamless access to a secure scripting environment directly in the PIXI UI. Additionally, retrieving PIXI database metadata and information from within a Jupyter notebook is made easy with XNATpy. The Jupyter integration utilizes JupyterHub and leverages Docker containers and Docker Swarm for efficient management and scalability of computing resources for Jupyter notebook users. Building upon the Jupyter notebook integration, we have integrated Python dashboards directly into XNAT. Leveraging technologies such as Dash, Panel, Streamlit, and Voilà, this integration provides researchers and other technical users with tools to create and share interactive visualizations and applications that can then be started by general users from within the XNAT UI. This eliminates the need for users to interact directly with Python code.

Overall, the development of the PIXI platform is expected to have a profound impact on the management of preclinical imaging datasets and co-clinical imaging to support cloud-based computational pipelines and integration with multi-scale correlative biology. Since PIXI was released in February 2024, PIXI has over 60 users at various levels of activity, and we anticipate the user base to expand as we continue to disseminate and build the capabilities of PIXI. Additional information, including the free download of PIXI, instructional videos, documentation, and mailing list sign-up, is available at https://www.PIXI.org/
study the impact of SDOH on other child health outcomes.

Abstracts

Poster Presentation

Enhanced MSBooster for Sensitive HLA Peptide Identification

Authors: Fengchao Yu, Kevin Yang, Pratik Jagtap, Reid Wagner, Timothy Griffin, Alexey Nesvizhskii
Submitter: Fengchao Yu
Submitter Email: yufe@umich.edu

The detection and analysis of human leukocyte antigen (HLA) peptides are one of the common approaches used to study cancer immunopeptidomics. In mass spectrometry-based proteomics, database searching is the primary approach for identifying peptides from human cell or tissue samples. During the search, tandem mass spectra are compared with theoretical spectra generated in silico from a proteome database. The peptide with the highest-scored theoretical spectrum is reported as the identification for each tandem mass spectrum. However, traditional theoretical spectra lack fragment intensity and retention time information. Recently, advances in deep-learning prediction have enabled the accurate prediction of the fragment intensity and retention time. We developed and published a peptide-spectrum match (PSM) rescoring tool, MSBooster, to perform the prediction and calculate two additional scores after MSFragger database search. These scores are the entropy score, which measures spectral similarity, and the normalized retention time difference. These two scores, along with other traditional database search scores, are used together to re-rank PSMs and boost sensitivity. Since the publication, MSBooster has become one of the most popular rescoring tools, and is used in almost all tryptic peptide identifications performed by FragPipe. However, HLA peptides have different fragmentation patterns because they are different from most tryptic peptides that end with Lysine or Arginine. Traditional in silico fragmentation rules, which generates b- and y-ions, do not very suitable for HLA peptides. There are more ion types, such as a-ion, immonium, and internal ions, other than b- and y-ions. The fragment intensities are also different from those of the tryptic peptides.

To tackle these challenges, we have expanded the capabilities of MSBooster to accommodate additional ion types beyond the b- and y-ions. Furthermore, we have integrated Koina, a fragment intensity and retention time prediction server, into MSBooster to support a broader range of models for HLA, tandem mass tags (TMT), and phosphorylated peptides. Depending on the user's selection, MSBooster employs specific models to perform the predictions, and calculate scores. If the models are not the built-in DIA-NN models, MSBooster sends the peptides to the Koina server to obtain prediction results. Our experiments demonstrated that using data-type-specific models results in higher sensitivity without significantly increasing processing time. The difference is particularly notable for the HLA peptides compared to others. The latest version of MSBooster has already been released and included in the latest version of FragPipe, a proteomics data analysis suite with both graphical user interface and command line interface.

Discovery of Novel CDK Inhibitors with ARCHS4/RummaGEO and the LINCS L1000 Datasets

Authors: John Erol Evangelista, Alexander Lachmann, Daniel Clarke, Avi Ma'ayan
Submitter: Avi Ma'ayan
Submitter Email: avi.maayan@mssm.edu

CDK inhibitors are emerging as one of the most successful family of drug targets for cancer. In this project we aimed at identifying novel CDK inhibitors by mining data from the Library of Integrated Network-Based Cellular Signatures (LINCS) program (https://lincsproject.org/) and ARCHS4 (https://maayanlab.cloud/archs4/). The NIH Common Fund LINCS program has profiled the effect of over 30,000 drugs and small molecules on human cancer cell lines to produce approximately one million gene expression signatures. At the same time, the ARCHS4 resource provides uniformly aligned RNA-seq samples collected by thousands of studies listed on the NCBI GEO database. We recently processed the data from the ARCHS4 resource to automatically generate over 170,000 gene expression signatures. These signatures are served on the RummaGEO platform (https://rummageo.com/). These large collections of gene expression signatures enable the mining for novel drugs and targets. Here we report on the development of a novel computational workflow that can be used to identify CDK inhibitors by mining the data from these resources. Using differentially expressed gene sets from CDK inhibitors extracted from RummaGEO, we benchmarked several strategies to perform drug and compound prioritization by querying the LINCS L1000 dataset. As a result of this benchmark, we identified over 100 potential preclinical compounds that previously were not known to have CDK inhibitor activity. In addition, the benchmark suggests the best strategy to query the LINCS L1000 resource. The new methods to query the L1000 dataset, based on the results of the benchmark, are provided as a Python package with documentation.

Poster Presentation

Introduction to Bioinfor-omics: Online Course about the Application of Bioinformatics Methods to Multi-Omics Datasets

Authors: Heesu Kim, Daniel Clarke, Giacomo Marino, Zhuorui Xie, Alexander Lachmann, Ido Diamant, John Erol Evangelista, Eden Deng, Stephanie Olaiya, Sherry Jenkins, Avi Ma'ayan
Submitter: Avi Ma'ayan
Submitter Email: avi.maayan@mssm.edu

To provide an engaging training and outreach activity that would promote the use of the ARCHS4 resource, we have developed an interactive comprehensive online course called "Introduction to Bioinfor-Omics". The course consists of 18 lectures, each lecture has a slide deck, video recordings, quizzes, and code examples. Overall, the course offers over 20 hours of video content, 180 quiz questions, and 18 Jupyter notebooks for hands-on learning. The course starts with an introduction to computer programming and biomedical research concepts, and then proceeds to cover practical bioinformatics applications. It embarks on teaching essential skills in Python programming, UNIX shell commands, and version control using Git and GitHub. The course then explains technical concepts focusing on methods from Systems Biology crucial for multi-omics data analysis in cancer research. This covers sequence alignment, clustering algorithms, and data normalization methods. Following those, dimensionality reduction techniques such as PCA, t-SNE, and UMAP are discussed, followed by statistical concepts such as multiple hypothesis correction, and differential expression analysis. Then, enrichment analysis techniques are presented including the application of various workflows to analyze multi-omics dataset collected from cancer patients. The course also introduces Knowledge Graph databases and the Cypher query language, providing a different point of view on how biological data from different sources can be integrated. The curriculum incorporates state-of-the-art concepts from Text Mining including in-depth discussion of large language models (LLMs), utilizing APIs to construct bioinformatics workflows, and foundational concepts of AI such as Machine Learning and Deep Learning. Overall, the Introduction to Bioinfor-Omics course provides an engaging opportunity for interested students to explore concepts from multi-omics bioinformatics analysis, equipping them with the essential skills to advance cancer research.

Topological Features for Histopathology Modeling

Authors: Meilong Xu, Prateek Prasanna, Chao Chen
Submitter: Prateek Prasanna
Submitter Email: prateek.prasanna@stonybrook.edu

It is well established that tumor morphology (e.g., formations of glands, cell clusters, cell mixture) plays an essential role in tumor progression. Different pathology imaging techniques can now visualize the rich spatial information, enabling more structure-focused computational approaches aimed at examining these structural cues and how they are related to outcomes. We propose to develop and study topological and spatial features for characterizing tumor microenvironment an tissue architecture. Using the mathematic theory of persistent homology, we have developed a topological approach that provides a quantitative assessment of the architecture of tissue structure including clusters, shapes, and voids that capture and quantitate the architectural features intrinsic to glandular and vascular structures and varying patterns of stromal cellular infiltration.

We will present different results including outcome prediction, analysis of novel pathology imaging (FIBI) modality, as well as novel generative models. These results provide proof of concept that such topological features have the potential of predicting tumor progression and unveiling tumor evolution biology.

Poster Presentation

CARS-TEA: Utilizing ARCHS4 to Enable Enrichment Analysis at the Transcript Level

Authors: Anna Byrd, Giacomo Marino, Avi Ma'ayan
Submitter: Avi Ma'ayan
Submitter Email: avi.maayan@mssm.edu

Gene set enrichment analysis is a popular computational method that provides functional context to gene sets produced by omics Gene set enrichment analysis is a popular computational method that provides functional context to gene sets produced by omics profiling assays. However, most enrichment analysis tools, algorithms, and databases do not consider the fact each human gene gives rise to multiple transcripts that may have different functional roles, participate in different macro-molecular complexes and other protein interactions, and differ in their pattern of expression. Here we introduce Correlation-based Automated RNA Sequencing Transcript Enrichment Analysis (CARS-TEA), an enrichment analysis tool that provides enrichment analysis at the transcript level. To develop CARS-TEA, we created transcript set libraries by computing gene expression signatures from the ARCHS4 and RummaGEO resources. We also used the ARCHS4 gene-gene co-expression correlation matrix to convert gene set libraries from Enrichr to transcript set libraries. Overall, the CARS-TEA database contains over 650,000 annotated transcript sets available for search from the CARS-TEA web-server application. To demonstrate the utility of CARS-TEA, we applied it to analyze the CPTAC3 pan-cancer cohort. By comparing enrichment analysis at the proteomics, phosphoproteomics, and transcriptomics levels, we demonstrate that enrichment analysis at the transcript level more accurately recovers the correct pathways and biological processes in tumors from individual patients.

Using ARCSH4 for scRNA-seq Imputation and Cell Type Identification

Authors: Nasheath Ahmed, Giacomo Marino, Billal Ali, Sophie Gideon, Avi Ma'ayan
Submitter: Avi Ma'ayan
Submitter Email: avi.maayan@mssm.edu

Single-cell RNA sequencing (scRNA-seq) facilitates the profiling of gene expression of tumor samples at the single cell level. However, due to the current technical limitations, the expression vectors from each single cell contain many genes with no expression value. While there are many imputation algorithms to fill in the expression of the missing values, most methods only use data from the measurements of the other genes in single cells from the same experimental sample. We have developed a scRNA-seq imputation method that leverages the rich RNA-seq gene expression data provided by ARCHS4. The imputation strategy first identifies a collection of ARCHS4 samples most similar to the incomplete single-cell profiles. It then learns coefficients through an optimization algorithm to impute the single cell data. We demonstrate that the imputation model improves clustering and subsequent cell type identification. In addition to developing an imputation model with the ARCHS4 resource, we also utilize ARCHS4 to improve scRNA-seq algorithms for cell type identification. ARCHS4 data is utilized for developing reliable cell markers for hundreds of cell types, as well as augmenting cell markers with gene-gene co-expression correlations computed from the ARCHS4 resource. We applied the ARCSH4-assisted imputation and cell type identification methods to melanoma tumors profiled by the CPTAC consortium to identify and prioritize small molecules and immuno-therapeutic targets for individual tumors with the assistance of the tools TargetRanger and SigCom LINCS. Overall, the project demonstrates how the ARCHS4 resource can enable enhanced knowledge extraction from scRNA-seq data. The imputation and cell type identification algorithms applied to target indemnification given tumor samples will be made as a website with an interactive graphical user interface at: https://scrnaseq2targets.maayanlab.cloud/.

Poster Presentation

A novel transformer-based deep learning model for predicting binding interactions between HLA class I molecules and peptides

Authors: Kun Hee Kim, Xianli Jiang, Yukun Tan, Jae Jun Ku, Shaoheng Liang, Maura Gillison, Ken Chen
Submitter: Kun Hee Kim
Submitter Email: kkim14@mdanderson.org

Predicting the binding affinities between human leukocyte antigen class I molecules (HLA-I) and peptides is crucial for understanding immune responses and developing immunotherapies. Although computational models have been developed to facilitate efficient and accurate screening of ligands with high binding probabilities, they often fail to generalize on peptide sequences of varying lengths. Deep learning models have been proposed, but they tend to bias toward common peptide lengths found in training data and often lack interpretability to reveal underlying mechanisms driving binding.

Methods
We introduce a novel deep learning model designed to predict the bindings between HLA-I molecules and peptides (pHLA-I). Our model, based on transformer architecture, encodes the amino acid sequences of HLA-I molecules and peptides using a co-attention mechanism that iteratively updates the representations of both sequences to capture their crucial interactions driving binding. The final representations of the peptides' classification tokens (CLS) are then fed into a classification head to predict binding probabilities.

Results
Our model successfully competes against established pHLA-I prediction models, including TransPHLA, another transformer-based deep learning model using a self-attention mechanism, achieving an F1 score of 92.8% for independent test data and 87.8% for external data. Notably, our model is more robust on longer peptides with more than 12 residues, which constitute less than 5% of the training data. On the external data, our model showed F1-scores for binding predictions with peptides of lengths 12, 13, and 14 that were 7.2%, 6.9%, and 21.4% higher, respectively, compared to TransPHLA (F1-scores: 0.74, 0.62, and 0.51). Additionally, we facilitated the inference of binding motifs for each HLA-I allele with different ligand lengths from the attention scores at co-attention layers. We observed that the reconstructed binding motifs from attention scores aligned with binding motifs of well-studied HLA-I alleles.

Conclusions
Our model not only provides robust and accurate pHLA-I binding predictions with enhanced interpretability but also facilitates the inference of binding motifs for rare and under-examined HLA alleles. Furthermore, the updated representations of HLA and peptide sequences generated by our model could potentially extend to modeling other immunological tasks, such as T-cell receptor immunogenicity prediction, thereby broadening the model's application in immunotherapy and vaccine development.

The DepMap Portal: Enabling explorations of cancer cell lines and dependencies

Authors: Ali Mourey, Jessica Cheng, Nayeem Aquib, Philip Montgomery, Randy Creasi, Sarah Whitaker, Josh Dempster, Lauren Golden, Yvonne Blanco, Katie Campbell, Francisca Vazquez
Submitter: Sarah Wessel Whitaker
Submitter Email: swessel@broadinstitute.org

Global efforts are underway to characterize cancer cell lines and identify their genetic and pharmacological vulnerabilities. However, interpretation is dependent on the accessibility of many datasets and their integration across several modalities. To empower researchers to make use of this growing base of knowledge, we have harmonized and published dozens of datasets alongside software and visualization tools which enable researchers to explore and analyze data. In our most recent update of the DepMap portal, we have introduced Context Manager, a new tool, allowing users to define cohorts to use in analysis throughout the portal. In addition, we have improved the organization of datasets and redesigned our downloads area along with Data Explorer to incorporate these new conventions. Now in its fifth year, the public DepMap portal (depmap.org) receives 10k unique visitors per week and has an active community forum (forum.depmap.org).

Poster Presentation

Metabolomics Outcomes as a Function of Biological Complexity from Colorectal Cancer-Related Microbiome Samples

Authors: Jennifer Nguyen, Joseph Krampen, Jungmoo Huh, Kathryn McBride, Patrick Schloss, Marcy Balunas
Submitter: Marcy Balunas
Submitter Email: mbalunas@umich.edu

With colorectal cancer (CRC) cases rising worldwide, there is an increasing need for non-invasive CRC screening and detection of alternative biomarkers from complex biological samples. Current methods fail to provide accurate detection and abundances of compounds from complex clinical samples including fecal samples analyzed by mass spectrometry (MS). This work aims to investigate how increasing complexity of MS samples impacts the characterization of metabolites and the results of untargeted metabolomics, with the overall goal of including algorithms for more accurate quantification of MS features into our mums2 R package designed to facilitate advanced analyses of untargeted metabolomics data. Herein, we performed untargeted MS on CRC-relevant samples, strains, and metabolites, including generating additional complexity at all levels by creating standardized mixtures. As expected, results indicated substantially higher number of metabolites in the complex samples as compared to the standardized metabolite and monoculture samples. Our analytical methods detected shared metabolites between simplified and complex samples, yet metabolite abundances were more variable in complex samples. Overall, we show that sample complexity can obscure ion detection and accurate abundance measurements when using untargeted MS. Future studies will move beyond first-ordered approaches using chromatographic features and instead apply multinomial mixture models to obtain more accurate quantification of metabolites. Continued improvements in accuracy of untargeted MS analyses of complex biological samples will increase the metabolic information gathered and facilitate identification of novel biomarkers for diseases such as CRC.

Investigating epithelial-mesenchymal transition and endothelial-mesenchymal transition in the tumor microenvironment of HPV+ head and neck cancer

Authors: Catherine Zhou
Submitter: Catherine Zhou
Submitter Email: cz68@rice.edu

Epithelial-mesenchymal transition (EMT) and endothelial-mesenchymal transition (EndMT) have been associated with increased metastatic potential and poor treatment response in head and neck cancer. However, the tumor-intrinsic factors driving EMT and EndMT in HPV-positive head and neck cancer remain poorly understood. This study investigates the transformation of malignant epithelial cells and endothelial cells into cancer-associated fibroblasts (CAFs) through EMT and EndMT processes and their role in tumor progression. Using a multi-faceted approach including single-cell RNA-sequencing, lineage tracing, copy number variation analysis, cell-cell interactions, and pathway enrichment analysis, we examined transformed fibroblasts in HPV-positive head and neck cancer samples. Our findings reveal distinct CAF populations with diverse origins, including those derived from epithelial and endothelial cells. We propose an activation trajectory for these CAF types, each demonstrating unique interactions within the tumor microenvironment. Notably, we observed the presence of EndMT-derived CAFs in several samples from non-major histologic responders, suggesting a potential link to treatment resistance. Furthermore, we uncovered specific signaling pathways and transcription factors driving the EMT and EndMT processes in HPV-positive head and neck cancer, providing potential targets for therapeutic intervention.

Abstracts

Poster Presentation

Overture: An Open-Source Genomics Data Platform

Authors: Christina Yung, Mitchell Shiell, Jon Eubank, Justin Richardsson, Brandon Chan, Robin Haw, Lincoln Stein, Melanie Courtot, Overture Team
Submitter: Mitchell Shiell
Submitter Email: mshiell@oicr.on.ca

Overture is an open-source software suite designed to overcome challenges in storing, managing, and sharing genome-scale datasets. Our mission is to provide informaticians with accessible infrastructure that enables data reuse and accelerates scientific discovery. Our modular and scalable microservice architecture, comprising Ego, Song, Score, Maestro, and Arranger, powers projects such as VirusSeq, ICGC ARGO, and the IHCC Cohort Atlas.

Overture's core capabilities cater to three primary user groups: data consumers retrieving data, data providers submitting datasets, and administrators managing users, configuring data models, and customizing portal interfaces. This poster highlights how users interact with our platform. It broadly covers data submission, exploration, download, and configuration of the search UI and data model. We will also include updates on our work in integrating visualization tools such as jBrowse and IOBIO.

We are particularly excited to share how we are overcoming barriers to adoption. While microservice architectures offer numerous advantages (scalability, flexibility, resilience, etc.), ease of setup and deployment is not one of them. To reduce the friction of adopting our platform, we have followed a framework that focuses on answering three core questions: How do people see what Overture does? How can they try it out? And how can they own it?

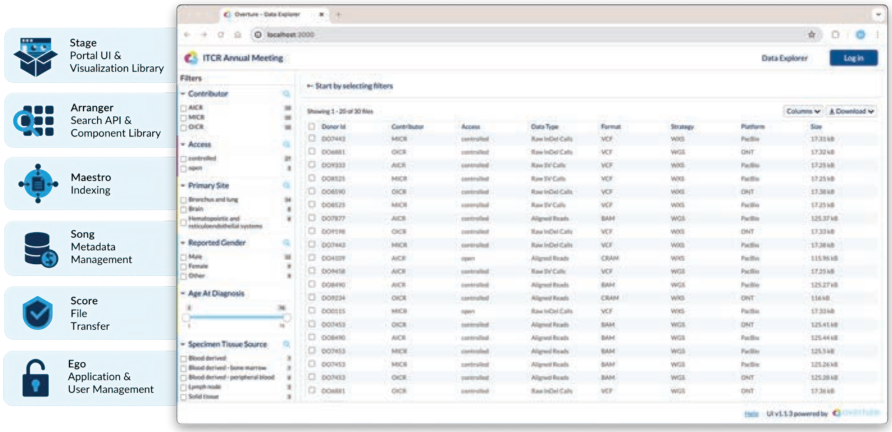
For the "see it" question, we created an Overture demo portal (https://www.demo.overture.bio/). This read-only environment is one click away from our homepage, providing new and prospective users with a first impression of Overture. From the demo, users can interact with our representative mock dataset on the exploration page and browse supplemental content built into the portal's Stage UI. These guides summarize how the Overture platform works and link to more in-depth resources that fulfill the next step of our user journey: trying it out.

For the "try it" question, we now have platform guides in addition to our individual product documentation. Our foundational guides cover data submission, download, and core administrative tasks required to configure an operational Overture platform. These guides also leverage our new Overture Quickstart, a Docker Compose that allows users to spin up our entire platform in minutes, complete with prepopulated mock data and a pre-configured admin user. This enables users to follow each guide step-by-step on their local platform, providing a truly hands-on experience.

To answer to the ownership and adoption of our platform, we have containerized and standardized the installation of our microservices. Each service can be installed with Docker and an environment variable file, ensuring broad portability across different environments. This is reflected in our product documentation and our first end-to-end deployment guide, which explains each stage, service, and environment variable used to set up an Overture platform.

We encourage you to explore our demo portal (https://www.demo.overture.bio/), quickstart, and updated platform guides (https://www.overture.bio/getting-started). If you wish to contact us remotely, our Slack channel, like our software, is always open and available (https://www.overture.bio).

Overture's Microservice Architecture: The diagram illustrates the integration of Overture's six core microservices: Ego, Song, Score, Maestro, Arranger, and Stage. These services work in concert to create a comprehensive genomic data management platform. Ego, or alternatively Keycloak, handles identity and permission management, securing all user and application interactions. Song and Score manage data submission, validation, and retrieval, providing data tracking and quality control. Maestro indexes metadata from Song into Elasticsearch, which Arranger uses to generate a GraphQL search API and the user interface for data exploration seen in this image. Finally, Stage integrates these services into a React-based front-end portal. This modular approach, allows for flexible deployment and customization to meet diverse genomic research needs.



Showing 1 - 20 of 30 files

Donor Id	Contributor	Access	Data Type	Format	Strategy	Platform	Size
DO7443	MICR	controlled	Raw InDel Calls	VCF	WXS	PacBio	17.33 kB
DO6881	OICR	controlled	Raw InDel Calls	VCF	WGS	ONT	17.32 kB
DO9333	AICR	controlled	Raw SV Calls	VCF	WXS	PacBio	17.25 kB
DO8525	MICR	controlled	Raw SV Calls	VCF	WXS	PacBio	17.25 kB
DO6590	OICR	controlled	Raw InDel Calls	VCF	WXS	ONT	17.38 kB
DO8525	MICR	controlled	Raw SV Calls	VCF	WXS	PacBio	17.25 kB
DO7877	AICR	controlled	Aligned Reads	BAM	WGS	PacBio	125.37 kB
DO9198	OICR	controlled	Raw InDel Calls	VCF	WXS	ONT	17.33 kB
DO7443	MICR	controlled	Raw InDel Calls	VCF	WXS	PacBio	17.38 kB
DO4109	AICR	open	Aligned Reads	CRAM	WXS	PacBio	115.96 kB
DO9458	AICR	controlled	Raw SV Calls	VCF	WGS	PacBio	17.25 kB
DO8490	AICR	controlled	Aligned Reads	BAM	WGS	PacBio	125.27 kB
DO9224	OICR	controlled	Aligned Reads	CRAM	WXS	ONT	116 kB
DO0115	MICR	open	Raw InDel Calls	VCF	WXS	PacBio	17.33 kB
DO7453	OICR	controlled	Aligned Reads	BAM	WGS	ONT	125.41 kB
DO8490	AICR	controlled	Aligned Reads	BAM	WGS	PacBio	125.44 kB
DO7453	MICR	controlled	Aligned Reads	BAM	WGS	PacBio	125.5 kB
DO7453	MICR	controlled	Aligned Reads	BAM	WGS	PacBio	125.26 kB
DO7453	OICR	controlled	Aligned Reads	BAM	WGS	ONT	125.28 kB
DO6881	OICR	controlled	Raw InDel Calls	VCF	WGS	ONT	17.36 kB

Poster Presentation

Towards an automated expert knowledge base reviewer using natural language processing

Authors: Caralyn Reisle, Cameron J. Grisdale, Kilannin Krysiak, Arpad M. Danos, Mariam Khanfar, Erin Pleasance, Jason Saliba, Melika Bonakdar, Malichi Griffith, Obi L. Griffith, Steven J.M. Jones
Submitter: Caralyn Reisle
Submitter Email: caralynreisle@gmail.com

Interpretation of genomic findings remains one of the largest barriers to automation in processing precision oncology patient data due to the high level of expertise in cancer biology, genomics, and bioinformatics required. Efforts to streamline this process include creating cancer knowledge bases (KB) to store annotations of individual genes and variants. Cancer KBs are essential to interpreting genomic findings by facilitating matching patient variants to their known relevance in literature, but the creation of such resources is time-consuming. The open-data cancer KB CIViC (https://civicdb.org) adopted a crowd-sourcing method to curate structured content efficiently. However, these submissions still require expert review, leading to a new bottleneck in expanding content.

Recent advances in natural language processing (NLP) and computer infrastructure have given rise to large generative text models, such as GPT-4, commonly referred to as large language models (LLMs). General purpose LLMs show impressive performance but are prone to errors, emphasizing the importance of fact-checking, especially in high-stakes applications such as precision oncology. Adapting these general-purpose models to be suitable to fact-check cancer KBs requires fine-tuning using a task-specific dataset. Several fact-checking models and datasets exist but are out of domain (political facts, celebrity news, etc.) or limited to scientific abstracts. By contrast, the claims that will need to be verified in the precision oncology domain are complex and require scientific expertise as well as the integration of highly heterogeneous knowledge ranging from basic functional experiments to advanced clinical trials.

To address this deficiency, we have curated a CIViC companion dataset of cancer facts and their corresponding evidence in the literature using the web-based tool hypothesis (https://web.hypothes.is). Cancer facts represent individual accepted evidence items in CIViC and are composed of therapeutic, prognostic, diagnostic, and biological associations of variants. The cancer fact dataset contains annotated data manually curated by 11 experts, covering 1165 CIViC entries, and 707 publications. Using this bespoke dataset, we have fine-tuned several different LLMs (ex. LLama3) to act as an expert KB reviewer. Thus far, our reviewer model has achieved promising results in verifying cancer facts, with up to 89% accuracy, and continues to improve as further data is collected. The reviewer model will be used in tandem with human experts to streamline the review process and prioritize content for review.

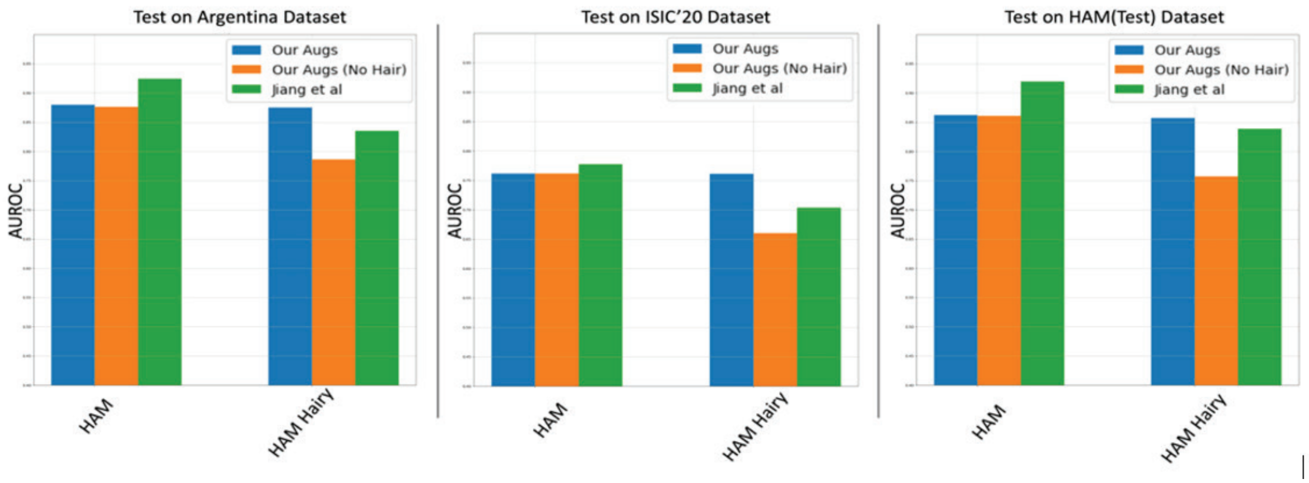
Future work includes open-access publication of this dataset so that the scientific community may use it to further the field of fact-checking and ultimately improve KBs used for precision oncology. Artificial intelligence and natural language processing will play a key role in the future of evidence-based medicine enabling precision oncology programs to efficiently and accurately interpret genomic data. Increased capacity and reduced turn-around times are critical in preparing precision oncology programs to transition from clinical trials to the standard of care.

Poster Presentation

Image Embeddings to Reduce Image Feature Variance

Authors: Torop Max, Jennifer Dy, Veronica Rotemberg, Kivanc Kose
Submitter: Veronica Rotemberg
Submitter Email: vrotemberg@gmail.com

Trained naively, deep learning (DL)-based models (e.g., self-supervised representation or classification models) are known to encode clinically irrelevant image features. The biases introduced by these confounding factors will then affect the generalizability of the learned model, often making it less reliable on new data. This issue is particularly relevant in domains such as dermatology, where clinicians use a large number of different imaging devices (cameras, dermatoscopes, and optical attachments that can be used on any smartphone) available through various providers. The lack of standardization in image acquisition devices and associated image collection procedures leads to unwanted and frequently non-diagnostic but correlated-with-variables-of-interest image features. Under such circumstances, model performance can vary dramatically based on data source. For instance, each of the top 25 performing models from the International Skin Imaging Collaboration (ISIC) 2019 challenge correctly identified all melanoma images from one source while failing to identify any melanoma from another. One of the main aims of the U24 grant is to develop representation models that are capable of learning diagnostic features and that are robust against these confounding image variations. Towards this end, we have developed novel augmentation methods, such as generating hairy versions of the lesions, inserting rulers or skin tag markers commonly encountered in dermoscopy images, adding vignetting (white or dark) or light reflections to the edges of the images, as well as introducing color variations (e.g., tint, brightness, contrast) reflecting the differences between different dermatoscopes. We trained three different self-supervised models using three different augmentation strategies: (i) ISIC 2020 winner, (ii) Jiang et al., and (iii) our proposed augmentations. We first tested the source-invariance quality of the models. We used these three trained models to calculate the source invariance ratio, which is defined as the mean of pair-wise similarity between embeddings within the same source divided by the mean pair-wise similarity between embeddings from the source and embeddings from all other sources. Our results show that our augmentations have the best source invariance (as a metric, lower is better). We also calculated the capability of estimating the source from image embeddings obtained using these models. We trained a 6-class classifier (one class for each source) using logistic regression and Multilayer perceptron methods to find out if we can predict the source of the image from their embedding representations. Our results show it is harder to predict the source of the embeddings from our augmentations. We also tested the diagnostic performance of the embeddings generated by these three models. First, we compared the binary classifiers trained using the embeddings generated by the three different self-supervised models. The results indicate that diagnostic classification models trained using our augmentations result in ~2-5% lower AUC, which is a sign of lower diagnostic performance. However, naively checking only the diagnostic performance is unreliable as it does not account for the potential biasing effect of the artifact on the diagnostic results. In order to check the robustness of our model against biases due to confounding artifacts.



We trained 2 binary classification models, one with the original HAM training dataset and another after augmenting the melanoma class with hair augmentations. Testing on images from 3 different sources, we observed that models trained using embeddings generated using our augmentations are robust against confounding artifacts as the performance did not degrade.

Abstracts

Poster Presentation

The Integrative Genomics Viewer (IGV) for Cancer Research

Authors: James Robinson, Helga Thorvaldsdottir, Jill Mesirov
Submitter: Caralyn Reisle
Submitter Email: helga@broadinstitute.org

Cancer genomic studies produce a flood of diverse data from patient tumors, circulating free DNA (cfDNA), animal models, and cell lines. They include whole exome, whole genome, copy number, and RNA profiles; epigenetic and conformational studies; and functional assays. Experienced and knowledgeable human review of the computational data processing and analysis of these datasets is essential to confirm and interpret results and even identify subtle signals. The Integrative Genomics Viewer (IGV) was developed to address this need by providing a high-performance, easy-to-use, track-based viewer for the visual exploration of genomic data. Since 2008, IGV has been heavily used by the cancer research community with the goal of better understanding cancer through identification of genomic abnormalities and other significant events that can be linked to functional implications. IGV use has also expanded to evaluate results of clinical sequencing data processing workflows.

IGV is made available and maintained in multiple forms, for end users, bioinformaticians, and software developers:

IGV Desktop is a Java application that is downloaded and run on the user's computer. This was the initial IGV application and is how most investigators use IGV.

igv.js is a JavaScript implementation of IGV targeted for software developers who wish to embed a lightweight IGV viewer in their own data portals, cloud applications, and web pages. The igv.js viewer runs completely in the web browser, with no backend server and no data pre-processing required.

IGV-Web (https://igv.org/app) is an igv.js-based application that allows investigators to view local and remote genomic datasets using the familiar IGV user interface without the need to install any software. Importantly, users can easily share their visualizations of data stored on web servers and popular cloud providers.

igv-reports is a Python utility used to generate self-contained HTML reports that consist of a table of genomic sites of interest and associated interactive igv.js views for each site.

igv-notebook is a Python package that wraps igv.js for embedding in Jupyter, JupyterLab, and Google Colab notebooks. The embedded visualization can be controlled interactively or from code cells in the notebook.

All forms of IGV are freely available via https://igv.org. The software is open source (MIT license); the code is available at https://github.com/igvteam.

WebMev: leveraging cloud and containerization tools for performant, open platform for exploratory omics research

Authors: Derrick DeConti, Brian Lawney, Ilya Sytchev, Saron Nhong, John Quackenbush
Submitter: Derrick DeConti
Submitter Email: deconti@g.harvard.edu

WebMeV is a web-based software system that allows an "exploration first" approach to large, complex transcriptomic and methylomic data analysis. To achieve this objective and make data analysis easy for non-computationally skilled users, we had to overcome numerous obstacles including data format interoperability, data visualization interactivity, and network latency. WebMeV has increasingly made use of cloud technologies to decrease technical complexity and provide an improved user experience while making cost-efficient use of available computing resources. In developing WebMeV, we have also prioritized reproducibility, portability, and open science through the use of containerization technologies and open-source tools. Our focus on allowing exploratory analysis has required that we emphasize development of intuitive user interface designs, many of which we patterned after commercial web application designs for consumer sites so that the ideas and interfaces would be familiar to users. All of this has been developed in a framework in which our code base is modular and expandable, allowing us to add new methods as they are developed by our research group, published in the literature, or implemented through partnerships created as a result of the ITCR set-aside process. Because of its wide variety of methods, the inclusion of public data sets in cancer, and its inclusion of new methods for single-cell and spatial genomic analysis, WebMeV has evolved into a extremely powerful tool that allows scientists to take advantage of genomic data and advanced analytics in an integrated platform to explore the drivers of cancer and other diseases.

Poster Presentation

The GenePattern ecosystem for cancer bioinformatics

Authors: Michael Reich, Thorin Tabor, Ted Liefeld, Edwin Huang, Forrest Kim, Helga Thorvaldsdottir, Jill Mesirov
Submitter: Michael Reich
Submitter Email: mmreich@cloud.ucsd.edu

As the availability of genomic data and analysis tools from large-scale cancer initiatives continues to increase, with single-cell studies adding new dimensions to the potential scientific insights, the need has become more urgent for a software environment that supports the rapid pace of cancer data science. The GenePattern ecosystem, first introduced in 2004 and updated continually to support the changing needs of cancer genomics research, supports the analytical, computational, and reproducibility needs of the world-wide cancer analysis community.

The GenePattern server, available at www.genepattern.org, provides hundreds of analysis methods for scientists at all levels of computational sophistication, with the only requirement for use being a web browser. The GenePattern server offers bulk and single-cell RNA-seq, copy number variation, flow cytometry, network analysis, general machine learning, gene set enrichment analysis, proteomics, and many other modalities. Analysis steps can be linked together into pipelines that can then be edited, shared, and made public. All parameters of an analysis, including the code version, are recorded so that an analytical result can be reproduced at any point in the future.

The GenePattern Notebook system, notebook.genepattern.org, is an integration of the Jupyter Notebook environment with the GenePattern server. It combines the research narrative capabilities of Jupyter with the non-programming approach and breadth of analyses available in GenePattern. Scientists using GenePattern Notebook can create documents that include richly-formatted text and multimedia, executable code, and GenePattern analyses. A single GenePattern notebook can therefore comprise a sophisticated analytical workflow that runs on multiple remote servers.

We have recently expanded GenePattern Notebook into a new notebook environment, Genomics to Notebook, g2nb.org, which adds the ability to include analyses on any public Galaxy server as well as using the Integrative Genomics Viewer (IGV) and Cytoscape within notebooks. Each analysis or visualization appears a cell within a notebook, preserving the accessibility and ease of use of the notebook metaphor while retaining the essential aspects of each tool's user interface. When run, the entire analysis appears to execute seamlessly within the notebook. The online workspace also includes a library of featured genomic analysis notebooks, including templates for common analysis tasks as well as cancer-specific research scenarios and compute-intensive methods. Scientists can easily copy these notebooks, use them as is, or adapt them for their research purposes.

Poster Presentation

Determining tissue-independent N6-methyladenosine (m6A) epitranscriptome and its regulatory role in cancer

Authors: Sumin Jo, Ting-He Zhang, Wen Meng, Jianqiu Zhang, Shou-Jiang Gao, Yufei Huang
Submitter: Yufei Huang
Submitter Email: YUH119@pitt.edu

N6-methyladenosine (m6A), the most prevalent mRNA modification in the human transcriptome, regulates vital biological processes by regulating post-transcriptional gene expression. m6A's pervasive role in gene regulation, spanning from normal physiological processes to complex pathophysiological states, has long been recognized. Still, the intricacies of its context-dependent methylation patterns remained elusive. In this study, we meticulously examined 69 methylated RNA immunoprecipitation sequencing (MeRIP-seq) samples across 24 human tissue types. Our investigation unveiled 5,945 tissue-independent (TI) m6A sites present across all 24 tissues. These sites exhibited distinct characteristics highlighted by significant enrichment near the stop codon, higher methylation levels, clustered with greater frequency, and residing in more conserved and shorter exon regions with much less likelihood to be the last exon of genes, compared to their less conserved counterparts. These characteristics imply their intrinsic nature, likely regulated by conserved cell-intrinsic mechanisms. Motif analysis has confirmed the ubiquitous RRACH m6A motif, suggesting an involvement of the canonical m6A methyltransferase complex (MTC). Yet we identified RBM15/RBM15B as the unique RNA binding protein (RPB) that may recruit to deposit these TI sites tissue samples. Also, we uncovered a marked inverse correlation between the methylation levels of these tissue-independent sites and the expression of the genes they methylated. Moreover, a large-scale mapping for RPB binding sites showed that these sites are preferentially bound by m6A reader YTHDF2 and YTHDC1, implicating the involvement of these readers in regulating mRNA decay and splicing. Significantly, genes methylated by these sites are enriched in essential cellular functions and include numerous transcription factors, tumor suppressor genes, oncogenes, and immune genes associated with cancer regulation. Further analysis of these genes using The Cancer Genome Atlas (TCGA) data and Genotype-Tissue Expression (GTEx) data showed their dysregulation across various cancers and tumors. Our work has, for the first time, identified a substantial number of tissue-independent m6A sites within human tissues, providing compelling evidence to show that these sites are modulated by tissue-independent factors and they appear to regulate mRNA degradation and splicing of genes with critical functions in regulation of cancer.

RummagenexRummaGEO: Crossing the Rummagene and RummaGEO Gene Sets for Novel Hypotheses Generation

Authors: Eugenia Ampofo (eugenia.ampofo@mssm.edu), Giacomo Marino, Daniel J.B. Clarke, Stephanie Olaiya, John Erol Evangelista, Sherry L. Jenkins, Avi Ma'ayan
Submitter: Eugenia Ampofo
Submitter Email: eugenia.ampofo@mssm.edu

Recently, the Ma'ayan Lab developed two new resources: Rummagene and RummaGEO. Rummagene is a web server application that provides access to ~750,000 gene sets extracted from supporting materials of 140,000+ articles after scanning 6 million articles available on PubMed Central. Similarly, RummaGEO is a web server application that automates extracting and categorizing over 300,000 human and mouse gene sets from the Gene Expression Omnibus (GEO) for comprehensive gene expression signature search. Since these two new resources produced massive collections of independent annotated gene sets, we sought to cross these two resources to discover gene sets that highly overlap but originate from different seemingly unrelated studies. In total, we compared 748,220 gene sets from Rummagene with 158,062 RummaGEO mouse gene sets and 135,264 RummaGEO human gene sets. This comparison led to the discovery of ~9 million gene set pairs that show high overlap (p-value < 0.001). These 9 million sets are stored in a database and made available for search via the RummagenexRummaGEO website. In addition to providing the overlapping pairs for search, we performed various analyses to characterize the contents within the RummageneXRummaGEO database. Furthermore, the top overlapping sets with abstract dissimilarity were examined for possible new connections between biological and biomedical concepts. For example, drugs and their unknown mechanisms of action, or two diseases with no prior knowledge about their similar molecular mechanisms. The RummageneXRummaGEO search engine is available from: https://rummagenexrummageo.dev.maayanlab.cloud/

NIH/NCI INFORMATICS TECHNOLOGY FOR CANCER RESEARCH (ITCR) ANNUAL MEETING

September 16-19, 2024, Indianapolis, IN