ITCR 2018 Annual Meeting Poster Abstract Book

Poster 1 - **"Tools to Analyze Morphology and Spatially Mapped Molecular Data: Quantitative Imaging for Pathology (QuIP) Software Platform"** - Joel Saltz, Ashish Sharma, Erich Bremer, Feiqiao Wang, Tammy DiPrima, Joe Balsamo, Ryan Birmingham, Jonas Almeida, Tahsin Kurc
Talk: Day 1, Imaging Session 1

Poster 2 – **"Quantitative Image Informatics for Cancer Research (QIICR)"** - Andrey Fedorov, David Clunie, Reinhard Beichel, John Buatti, Steve Pieper, Jayashree Kalpathy-Cramer, Michael Onken, Jean-Christophe Fillion-Robin, Fiona Fennessy, Ron Kikinis
Talk: Day 1, Imaging Session 1
Demo: Day 1, Demo Session 2

Poster 3 – **"Cancer Imaging Phenomics Toolkit (CaPTk): A Radio(geno)mics Platform for Quantitative Imaging Analytics on Computational Oncology"** - Christos Davatzikos, Despina Kontos, Paul Yushkevich, Russell Taki Shinohara, Yong Fan, Ragini Verma
Talk: Day 1, Imaging Session 1
Demo: Day 1, Demo Session 2

Poster 4 – **"Quantitative Radiomics System Decoding the Tumor Phenotype"** – John Quackenbush, Hugo Aerts
Talk: Day 1, Imaging Session 1

Poster 5 – **"Integrative Imaging Informatics for Cancer Research"** - Daniel S Marcus, Mikhail Milchenko, John Flavin, Rick Herrick, Simon Doran, James Petts, James D'Arcy, Steve Moore, Richard Wahl
Talk: Day 1, Imaging Session 1

Poster 6 – **"Pathology Image Informatics Platform for Visualization, Analysis and Management (PIIP)"** - Metin N. Gurcan, Anant Madabhushi, Dan Hosseinzadeh, Anne Martel
Talk: Day 1, Imaging Session 1
Demo: Day 2, Demo Session 4

Poster 7 – **"Informatics Tools for Tumor Heterogeneity in Multiplexed Fluorescence Images**" – S. Chakra Chennubhotla, D. Lansing Taylor, Brion Sarachan
Talk: Day 1, Imaging/Clinical Session 2

Poster 8 – **"Extensible Open-Source Zero-Footprint Web Viewer for Oncologic Imaging Research"** - Erik Ziegler, Trinity Urban, Rob Lewis, Chris Hafey, Cheryl A. Sadow, Annick D. Van den Abbeele, Gordon J. Harris
Talk: Day 1, Imaging/Clinical Session 2
Demo: Day 1, Demo Session 2

Poster 9 – **"SlicerDMRI: Open-source Diffusion MRI for Cancer Research"** - Isaiah Norton, Lauren O'Donnell
Talk: Day 1, Imaging/Clinical Session 2

Poster 10 – **"Informatics Tools for Optimized Imaging Biomarkers for Cancer Research & Discovery"** - Jayashree Kalpathy-Cramer, Robert Gillies, Dmitry Goldgof, Sandy Napel, Binsheng Zhao, Bruce Rosen
Talk: Day 2, Imaging Session 3

Poster 11 – **"Advanced Development of an Open-Source Platform for Web-Based Integrative Digital Image Analysis in Cancer"** - David Andrew Gutman, Lee Cooper
Talk: Day 2, Imaging Session 3

Poster 12 – **"TCIA Sustainment and Scalability - Platforms for Quantitative Imaging Informatics in Precision Medicine (PRISM)"** - Ashish Sharma, Joel Saltz, Lawrence Tarbox, Kirk Smith, Tracy Nolan, Jonathan Bona, Annie (Ping) Gu, Erich Bremer, Tammy DiPrima, Jonas Almeida, Mathias Brochhausen, Tahsin Kurc, Fred Prior
Talk: Day 2, Imaging Session 3
Demo: Day 1, Demo Session 2

Poster 13 – **"Developing Enabling PET-CT Image Analysis Tools for Predicting Response in Radiation Cancer therapy"** - Xiaodong Wu, Yusung Kim, John Buatti
Talk: Day 1, Break Out Session 2

Poster 14 – **"Development and Validation of Informatics Tools for Immunohistochemistry Analysis in Multi-Institutional Cohort Studies"** - Lee Cooper, Christopher R Flowers, Metin Gurcan, Deepak Chittajallu
Talk: Day 1, Lightning Talks

Poster 15 – **"EMERSE: The Electronic Medical Record Search Engine"** - David Hanauer
Talk: Day 1, Imaging/Clinical Session 2

Poster 16 – **"DeepPhe - A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records"** - Sean Finan, James Masanz, Olga Medvedeva, Eugene Tseytlin, Melissa Castine, Timothy Miller, Olga Medvedeva, David Harris, Harry Hochheiser, Chen Lin, Hadi Amiri, Girish Chavan, Jeremy L. Warner, Rebecca Jacobson, Guergana Savova
Talk: Day 1, Imaging/Clinical Session 2

Poster 17 – **"Advancing Cancer Pharmacoepidemiology Research through EHRs and Informatics"** - Liwei Wang, Lei Luo, Jeremy L. Warner, Yanshan Wang, Jason A. Wampfler, Hua Xu, Ping Yang, Hongfang Liu
Talk: Day 1, Imaging/Clinical Session 2

Poster 18 – **"Scalable Clinical Decision Support for Individualized Cancer Risk Management"** - Guilherme Del Fiol, Kensaku Kawamoto
Talk: Day 1, Imaging/Clinical Session 2

Poster 19 – **"PDX Finder: A Portal for Patient-Derived Tumor Xenograft Model Discovery"** - Terrence F. Meehan, Nathalie Conte, Jeremy Mason, Csaba Halmagyi, Steven Neuhauser, Abayomi Mosaku, Dale A. Begley, Debra M. Krupke, Helen Parkinson, Carol Bult
Talk: Day 1, Imaging/Clinical Session 2

Poster 20 – **"Trinity: Transcriptome assembly for genetic and functional analysis of cancer"** - Brian Haas, Asma Bankapur, Christophe Georgescu, Vrushali Fangal, Carrie Ganote, Cicada Brokaw, Thomas Doak, Aviv Regev
Talk: Day 1, Omics Session 1
Demo: Day 2, Demo Session 4

Poster 21 – **"Visualizing Structural Variation with the JBrowse Genome Browser"** - Robert Buels, Eric Yao, Lincoln Stein, Ian Holmes
Talk: Day 1, Omics Session 1

Poster 22 – **"Informatic Tools for Single-Nucleotide Analysis of Cancer RNA-seq" -** Esther Yun-Hua Hsiao, Yi-Wen Yang, Tracey Chan, Stephen Tran, Jae Hoon Bahn, Xinshu (Grace) Xiao
Talk: Day 1, Omics Session 1

Poster 23 – **"Streamlined sharing and analysis of clinical patient data for cancer research networks" -** Ian Foster
Talk: Day 1, Omics Session 1

Poster 24 – **"Computational Framework for Single-cell Genomics**" - Jude Kendall, Lubomir Chorbadjiev, Vyacheslav Zhygulin, Junyan Song, Joan Alexander[1], Michael Wigler, Alexander Krasnitz
Talk: Day 2, Omics Session 2

Poster 25 – **"The Network Data Exchange in 2018" -** Trey Ideker, Dexter Pratt
Talk: Day 2, Omics Session 2
Demo: Day 1, Demo Session 1

Poster 26 – **"Highly Interactive Next-Generation Clustered Heat Maps (NG-CHMs)" -** Bradley M. Broom, Michael C. Ryan, Chris Wakefield, Bob Brown, Futa Ikeda, Mark Stucky, James Melott, Rehan Akbani, John N. Weinstein
Talk: Day 2, Omics Session 2
Demo: Day 2, Demo Session 3

Poster 27 – **"UCSC Xena - Platform for Functional Genomics Visualization and Analysis" -** Jing Zhu, Brian Craft, Mary Goldman, Eric Collison, Suzanna Lewis, David Haussler
Talk: Day 2, Omics Session 2
Demo: Day 2, Demo Session 3

Poster 28 – **"The cBioPortal for Cancer Genomics: An intuitive open-source platform for exploration, analysis and visualization of cancer genomics data" -** The cBioPortal Consortium, Jianjiong Gao, Tali Mazor, Ersin Ciftci, Pichai Raman, Pieter Lukasse, Istemi Bahceci, Alexandros Sigaras, Adam Abeshouse, Ino de Bruijn, Benjamin Gross, Ritika Kundra, Aaron Lisman, Angelica Ochoa, Robert Sheridan, Jing Su, Onur Sumer, Yichao Sun, Avery Wang, Jiaojiao Wang, Manda Wilson, Hongxin Zhang, Priti Kumari, James Lindsay, Karthik Kalletla, Kelsey Zhu, Oleguer Plantalech, Fedde Schaeffer, Sander Tan, Dionne Zaal, Sjoerd van Hagen, Kees van Bochove, Ugur Dogrusoz, Trevor Pugh, Adam Resnick, Chris Sander, Ethan Cerami, Nikolaus Schultz
Talk: Day 2, Omics Session 2
Demo: Day 2, Demo Session 3

Poster 29 – **"Informatics Links between Histological Features and Genetics in Cancer"** – Kun Huang
Talk: Day 2, Omics Session 2

Poster 30 – **"Multi'omic analysis of subtype heterogeneity in high-grade serous ovarian carcinoma" -** Ludwig Geistlinger, Sehyun Oh, Lucas Schiffer, Marcel Ramos, Michael Birrer, Martin Morgan, Markus Riester, Levi Waldron
Talk: Day 2, Omics Session 2

Poster 31 – **"Reconstruction and interpretation of subclonal tumor evolution from rapid autopsy data reveals novel patterns of aggressive metastatic colonization" -** Xiaomeng Huang, Yi Qiao, Thomas Nicholas, Aaron Quinlan, Gabor Marth
Talk: Day 2, Omics/Clinical Session 3

Poster 32 – **"A Galaxy-Based Multi-Omic Informatics Hub for Cancer Researchers"** - Timothy Griffin, Pratik Jagtap
Talk: Day 2, Omics/Clinical Session 3

Poster 33 – **"Informatics Tools for High-Throughput Analysis of Cancer Mutations"** - Rachel Karchin, Michael Ryan
Talk: Day 2, Omics/Clinical Session 3
Demo: Day 1, Demo Session 1

Poster 34 – **"The Cancer Proteome Atlas: A Comprehensive Boinformatics Resource for Cancer Functional Proteomic Data"** - Jun Li, Yiling Lu, Wei Zhao, Rehan Akbani, Gordon B Mills, Han Liang
Talk: Day 2, Omics/Clinical Session 3
Demo: Day 2, Demo Session 4

Poster 35 – **"The Integrative Genomics Viewer (IGV): visualization supporting cancer research"** - James T Robinson, Helga Thorvaldsdóttir, Jill P. Mesirov
Talk: Day 2, Omics/Clinical Session 3
Demo: Day 2, Demo Session 3

Poster 36 – **"GenePattern Notebooks for Cancer Research"** - Michael Reich, Thorin Tabor, Peter Carr, Edwin Juarez, David Eby, Ted Liefeld, Helga Thorvaldsdóttir, Barbara Hill, Pablo Tamayo, Jill P Mesirov
Talk: Day 2, Omics/Clinical Session 3
Demo: Day 1, Demo Session 1

Poster 37 – **"The TIES Cancer Research Network (TCRN): Computational Pathology Support for Precision Oncology"** - Michael J. Becich, Joel Saltz, Jonathan Silverstein, Roni J. Bolla, Mathias Brochhausen, Chakra Chennubhotla, Michael D. Feldman, Carmelo Gaudioso, Tahsin Kurc, Jack London, Nita J. Maihle, Fred Prior, Ashish Sharma, Lawrence Tarbox
Talk: Day 2, Omics/Clinical Session 3
Demo: Day 2, Demo Session 4

Poster 38 – **"CIViC: Crowdsourced and community-driven standards and interfaces for curation and submission of somatic cancer variants"** - Obi Griffith, Kilannin Krysiak, Arpad Danos, Erica Barnell, Joshua McMichael, Susanna Kiwala, Adam Coffman, Nicholas Spies, Lynzey Kujan, Kaitlin Clark, Yang-Yang Feng, Zachary Skidmore, Cody Ramirez, Alex Wagner, Elaine Mardis, Malachi Griffith
Talk: Day 2, Omics/Clinical Session 3
Demo: Day 1, Demo Session 1

Poster 39 – **"A network-based approach for personalized treatment of Multiple Myeloma**" - Alessandro Laganà
Talk: Day 1, Break Out Session 1

Poster 40 – **"AMARETTO: Regulatory Network Inference for Driver and Drug Discovery in Cancer" -** Rileen Sinha, Thomas Baumert, Olivier Gevaert, Nathalie Pochet
Talk: Day 1, Break Out Session 1

Poster 41 – **"Algorithms for Literature-Guided Multi-Platform Identification of Cancer Subtypes" -** Dongjun Chung, Linda Kelemen
Talk: Day 1, Break Out Session 1

Poster 42 – **"Semantically rich interfaces for cloud-scale cancer genomics with Bioconductor"** - Vincent Carey, Shweta Gopaulakrishnan, Samuela Pollack, Aedin Culhane
Talk: Day 1, Break Out Session 1

Poster 43 – **"OncoMX: an integrated cancer mutation and expression resource for exploring cancer biomarkers"** - Raja Mazumder, Daniel Crichton, K. Vijay-Shanker, Frederic Bastian, Amanda Bell, Hayley Dingerdissen, Samir Gupta, Robel Kahsay, Heather Kincaid, David Liu, ASM Ashique Mahmood, Marc Robinson-Rechavi
Talk: Day 1, Break Out Session 1

Poster 44 – **"Improved Structural Prediction of peptide-HLA Complexes Using DINC-Vina"** - Dinler A. Antunes, Didier Devaurs, Eleni Litsa, Kyle R. Jackson, Mark Moll, Gregory Lizée, Lydia E. Kavraki
Talk: Day 1, Break Out Session 2

Poster 45 – **"CGDnet: Using patient-specific drug-gene networks for recommending targeted cancer therapies**" - Simina M Boca, Jayaram Kancherla, Shruti Rao, Subha Madhavan, Robert Beckman, Héctor Corrada Bravo
Talk: Day 1, Break Out Session 2

Poster 46 – **"Interactive single cell RNA-Seq analysis with the Single Cell Toolkit (SCTK)"** - David Jenkins, Tyler Faits, Emma Briars, Sebasitan Carrasco Pro, Steve Cunningham, Masanao Yajima, W. Evan Johnson
Talk: Day 1, Break Out Session 2

Poster 47 – **"Genome-wide somatic variant calling using localized colored DeBruijn graphs"** - Giuseppe Narzisi, André Corvelo, Kanika Arora, Ewa A. Bergmann, Minita Shah, Rajeeva Musunuri, Anne-Katrin Emde, Nicolas Robine, Vladimir Vacic, Michael C. Zody
Talk: Day 1, Lightning Talks

Poster 48 – **"TRUST: an ultrasensitive software for detecting TCR and BCR hypervariable-region sequences from bulk RNA-seq data"** - Xihao Sherlock Hu, Jian Zhang, Taiwen Li, Shengqing Stan Gu, Jin Wang, Jingxin Fu, Jun Liu, Bo Li, Xiaole Shirley Liu
Talk: Day 1, Lightning Talks

Poster 49 – **"Detection of somatic, subclonal and mosaic CNVs from sequencing"** - Alexej Abyzov
Talk: Day 1, Lightning Talks

Poster 50 – **"Database and tools for functional inference and mechanistic insight into somatic cancer mutations"** - Xue Le, Boshen Wang, Chia-Yi Chou, Alan Perez-Rathke, Jie Liang, Yan-Yuan Tseng
Talk: Day 1, Lightning Talks

Poster 51 – **"MMTF-Spark: Interactive, Scalable, and Reproducible Datamining of 3D Macromolecular Structures**" - Peter W. Rose
Talk: Day 2, Lightning Talks

Poster 52 – **"Integrated Querying of Biological Network Databases"** - Mehmet Koyutürk
Talk: Day 2, Lightning Talks

Poster 53 – **"Multiomics Data Compression"** - Olgica Milenkovic, Mikel Hernaez, Idoia Ochoa, Jian
Talk: Day 2, Lightning Talks

Poster 54 – **"The Exascale Deep Learning and Simulation Enabled Precision Medicine for Cancer"** - George Zaki, Eric Stahlberg, Tom Brettin, Rick Stevens

Poster 55 – **"The NCI Cancer Research Data Commons"** - Tanja Davidsen, Allen Dearry, Juli Klemm, Eve Shalley, Zhining Wang*, Tony Kerlavage

Poster 56 – **"NCI Cancer Research Data Commons Nodes in Development: The Proteomics Data Commons and Imaging Data Commons"** - Izumi Hinkson, Stephen Jett

Poster 1: Tools to Analyze Morphology and Spatially Mapped Molecular Data: Quantitative Imaging for Pathology (QuIP) Software Platform

Joel Saltz[1], Ashish Sharma[2], Erich Bremer[1], Feiqiao Wang[1], Tammy DiPrima[1], Joe Balsamo[1], Ryan Birmingham[2], Jonas Almeida[1], Tahsin Kurc[1]
[1]Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY
[2]Department of Biomedical Informatics, Emory University, Atlanta, GA,
joel.saltz@stonybrookmedicine.edu

This project is developing informatics capabilities for integrative analysis of digital Pathology image data. The underlying software platform, QuIP, has been designed and implemented as a suite of services deployed as containers [1]. The QuIP platform is being actively used in several scientific projects and application development efforts. In this abstract we briefly describe two applications.

QuIP has been used to develop a Web-based application and workflow for characterization of lymphocyte patterns using whole slide tissue images. The workflow uses a deep learning method (specifically a set of Convolutional Neural Networks) for classification of image patches and an iterative process of review and refinement of classification results. It leverages a whole slide image visual editing paradigm for this purpose. The application allows Pathologists to view predictions from the deep learning method and correct the predictions. The corrected predictions are then used to retrain the deep learning method. Patch-level predictions from the deep learning method are stored in the system as heatmaps. A user can view the lymphocyte prediction heatmap as an overlay on an image and use a set of tools to review and refine the algorithm predictions. This application has been used to study patterns of tumor infiltrating lymphocytes across multiple cancer types in The Cancer Genome Atlas. This study was a collaboration between Stony Brook, Emory, MD Anderson, Institute of Systems Biology and the TCGA PanCan Atlas Immune group. Results from about 5000 whole slide tissue images from 13 cancer types have been published in Cell Reports [2].

Another QuIP application implements support for curation of nucleus segmentation results in whole slide tissue images. The motivation for this application is the fact that most image segmentation algorithms are sensitive to input parameters and that there is heterogeneity in tissue morphology across tissue specimens and even within the same tissue. One set of parameters that perform well for a tissue image may not produce good segmentation results for another image. Hence, an image dataset may be analyzed multiple times with different algorithm parameters or different segmentation algorithms. The QuIP curation application assists users with storing and managing analysis results from multiple analysis runs and with reviewing and curating best segmentation results. Curated results can then be loaded to the curation application to share with other users. The curation application is currently being used to analyze and characterize whole slide tissue images from the SEER consortium.

[1] J Saltz, Sharma A, Iyer G, et al. A Containerized Software System for Generation, Management and Exploration of Features from Whole Slide Tissue Images. Cancer Research. 2017 Nov 1; 77(21)
[2] J. Saltz, R. Gupta, L. Hou, et al., “Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images”, Cell Reports 23, 181–193, April 3, 2018

Poster 2: Quantitative Image Informatics for Cancer Research (QIICR)

Andrey Fedorov[1*], David Clunie[2], Reinhard Beichel[3], John Buatti[3], Steve Pieper[4], Jayashree Kalpathy-Cramer[5], Michael Onken[6], Jean-Christophe Fillion-Robin[7], Fiona Fennessy[1], Ron Kikinis[1]

[1] Brigham and Women's Hospital, Harvard Medical School [2] PixelMed Publishing [3] Department of Electrical and Computer Engineering, University of Iowa [4] Isomics Inc. [5] Massachusetts General Hospital, Harvard Medical School [6] Open Connections GmbH [7] Kitware Inc.
[1*]Brigham and Women's Hospital, Harvard Medical School, andrey.fedorov@gmail.com

Quantitative imaging holds tremendous but largely unrealized potential for objective characterization of disease and response to therapy in the clinic. Automated analysis of clinical imaging data is gradually becoming available both in commercial products and clinical research platforms. As new tools are being introduced, their integration into the clinical or research environment, comparison with similar existing tools and reproducible validation are of critical importance. These tasks were traditionally difficult to achieve due to limited availability of open source research tools and lack of implementation of data standards. With the recent advances in automated imaging-based tissue phenotyping (radiomics) and other relevant artificial intelligence technologies, there is a new realization of the value of the structured machine-readable datasets available at large scale to enable training of AI models. The overarching objective of the QIICR project has been to provide some of the technological components and solutions to help address infrastructure gaps related to quantitative imaging research. In this presentation we will provide a brief summary of the most important accomplishments of the project over the past five years in such areas as development of open source informatics technology, advancement of standardization, and application of the developed tools in the scientific studies. We will discuss collaborations, both within and outside of ITCR. Finally, we will also outline our training and outreach efforts directed towards dissemination of the developed technology and standardization of the image analysis outputs generated by the research and commercial tools.

Online resources for further information

Project home page: http://qiicr.org
Project GitHub organization: https://github.com/qiicr
Catalog of the software tools developed by the project: http://qiicr.org/tools
Major training and outreach efforts:

- DICOM for Quantitative Imaging demonstration and connectathon at RSNA: https://qiicr.gitbooks.io/dicom4qi/
- DICOM tutorial at the MICCAI conference: http://qiicr.org/dicom4miccai/

Poster 3: Cancer Imaging Phenomics Toolkit (CaPTk): A Radio(geno)mics Platform for Quantitative Imaging Analytics on Computational Oncology

Christos Davatzikos, Despina Kontos, Paul Yushkevich, Russell Taki Shinohara, Yong Fan, Ragini Verma

Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, PA, USA, Christos.Davatzikos@uphs.upenn.edu

The growth of multiparametric imaging protocols has paved the way for imaging phenotypes that reflect underlying cancer molecular characteristics and spatiotemporal heterogeneity, can predict treatment response and clinical outcome, and can potentially guide personalized treatment planning. This mounting growth has also underlined the need for efficient quantitative integrative analytics to derive high-dimensional imaging signatures of diagnostic and predictive value in this emerging era of integrated precision diagnostics. Towards this end, we present the Cancer Imaging Phenomics Toolkit (CaPTk), an open-source and dynamically growing software platform for analysis of radiographic images of cancer, currently focusing on brain, breast, and lung cancer. CaPTk offers numerous analysis tools spanning across various steps of advanced computational radiological analysis workflow, including pre-processing, segmentation, feature extraction, and predictive modeling. Specifically, CaPTk leverages the value of quantitative imaging analytics along with machine learning algorithms to derive phenotypic imaging signatures, based on a two-level functionality. First, image analysis algorithms are used to extract comprehensive panels of diverse and complementary features, such as multiparametric intensity histogram distributions, texture characteristics, shape descriptors, kinetics, connectomics, and spatial distribution patterns. At the second level, these quantitative imaging signatures are fed into multivariate machine learning models for prediction, diagnosis, and prognosis. Current incorporated methodologies, as well as pre-trained models, are based on results from clinical studies focusing on: (i) computational neuro-oncology of brain gliomas for precision diagnostics, prediction of outcome, and treatment planning; (ii) prediction of treatment response for breast and lung cancer, and (iii) risk assessment for breast cancer. Specific highlight applications for computational neuro-oncology focus on glioblastoma and include predictions of a) EGFR mutations, b) overall survival, and c) areas of potential recurrence, which describes a computational method recently funded by the Abramson Cancer Center for a clinical trial, in the University of Pennsylvania, of increased dosimetry of radiation therapy in areas identified as areas of high risk of potential recurrence. The main highlight application for breast cancer risk assessment is “LIBRA”, a method for estimating the established breast cancer risk factor of breast density, which has been applied in numerous international/national multi-site studies. Finally, CaPTk is designed under a modular architecture allowing for direct integration of existing algorithms built using any programming language, thereby emphasizing its potential for collaborations, such as the ongoing collaboration with the ITCR-funded project of “Quantitative Image Informatics for Cancer Research (QIICR)” and particularly the DCMQI application to ensure the use of the DICOM standard for output segmentations.

Poster 4: Tools to Advance Radiomics Applications

John Quackenbush and Hugo Aerts

Dana-Farber Cancer Institute, <johnq@jimmy.harvard.edu>

Radiomics aims to quantify tumor phenotypes by extracting quantitative imaging features from CT scans and other radiographic images. These features can then be used either in a statistical modeling setting or through the application of deep learning techniques to create imaging biomarkers predictive of tumor mutational status, disease state, or other relevant clinical outcomes. However, such Radiomics artificial intelligence (AI) has been hampered by a lack of standards and widely-available, benchmarked, and standardized software and tools. We developed *PyRadiomics*, a flexible open-source platform capable that can extract a large panel of features from medical images and to use these in biomarker development and validation. We have seen growing adoption of *PyRadiomics* based on GitHub accesses, manuscript citations, and contributions of new methods, as well as enrollment of subscription to training sessions at large society meetings, including AAPM and RSNA. We have recently extended our radiomics tool kit by creating Radiomics-MeV, a software platform that provides an intuitive graphical user interface for exploring radiomic features and their biological and clinical relevance. Our hope is that this suite of tools becomes the standard for radiomics analysis and methods development and, in doing so, helps to accelerate advancement in the field.

Poster 5: Integrative Imaging Informatics for Cancer Research

Daniel S Marcus[1*], Mikhail Milchenko[1], John Flavin[1], Rick Herrick[1], Simon Doran[2], James Petts[2], James D'Arcy[2], Steve Moore[1], and Richard Wahl[1]

[1*]Department of Radiology, Washington University, dmarcus@wustl.edu
[2]CRUK Cancer Imaging Centre, Institute of Cancer Research

The Integrative Imaging Informatics for Cancer Research (I3CR) project is building new capabilities on XNAT to better support quantitative imaging in cancer research. XNAT is a web-based software platform designed to facilitate common management and productivity tasks for imaging and associated data. It consists of an image repository to store raw and post-processed images, a database to store metadata and non-imaging measures, and user interface tools for accessing, querying, visualizing, and exploring data. XNAT supports all common imaging methods, and its data model can be extended to capture virtually any related metadata. XNAT includes a DICOM workflow to enable exams to be sent directly from scanners, PACS, and other DICOM devices. XNAT's web application provides a number of productivity features, including data entry forms, searching, reports of experimental data, upload/download tools, access to standard laboratory processing pipelines, and an online image viewer. A fine- grained access control system ensures that users are restricted to accessing only authorized data. XNAT also includes a web services API for programmatic access and an open plugin architecture for extending XNAT's core capabilities.

A major component of our work with XNAT is to develop a container service using Docker-style technology to support repeatable and sharable automated analytics. This container service allows containers developed independent of XNAT to be deployed on XNAT through a simple administration interface in XNAT. The service supports retrieving containers from a (e.g. Docker Hub) and loading the container onto a specified Docker server. Analytic routines embedded in the container can then be executed through XNAT's web interface and from automated event-based triggers (e.g. when an exam is stored in the system). XNAT handles the actual interfacing with the container server, including loading data onto the container and storing data generated by the container back to the database.

Another area of active development is enabling interactive workflows associated with image annotation, labelling, and review. For this work, we have adopted the OHIF zero footprint viewer as the primary interface for visualizing and marking up images. Annotations generated in the viewer can be posted back to the XNAT repository and, if desired, trigger a subsequent step in the workflow, including launching a container-based analytic routine. In addition to the standard viewer, we're also building support for Osirix, RT-STRUCT, and Hologic annotations. In order to support these varied formats, a format-agnostic data structure is being implemented to which any of these formats can be converted to/from.

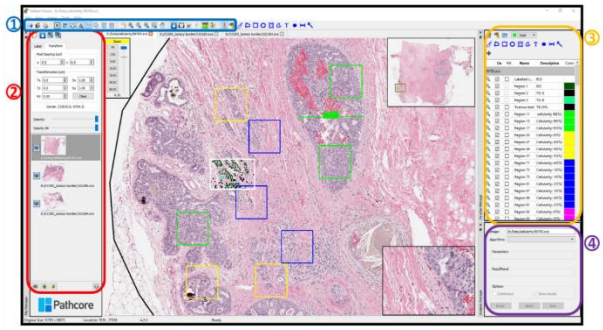
This work is being done in the context of several key cancer imaging uses cases, including surgical navigation based on resting state fMRI-generated functional networks and longitudinal multi-modal analysis of metastatic brain lesions. The primary goal of these projects is to create full-cycle workflows from imaging modalities to navigation and treatment systems.

Poster 6: Pathology Image Informatics Platform for Visualization, Analysis and Management (PIIP)

Metin N. Gurcan[1*], Anant Madabhushi[2], Dan Hosseinzadeh[3], and Anne Martel[4]

[1]Center for Biomedical Informatics, Wake Forest School of Medicine,mgurcan@wakehealth.edu
[2]Case Western Reserve University, [3]Pathcore ,[4]Sunnybrook Research Institute

The Pathology Image Informatics Platform (PIIP) (www.pathiip.org) is a project intended to support the management, sharing, annotation, and quantitative analysis of digital pathology images. One of the project goals is to develop and embed image analysis applications into an existing, freely available, pathology image viewer, Sedeen (http://pathcore.ca/downloads/), in order to create a shared resource for the digital pathology and cancer research communities. Several existing image analysis algorithms including out of focus detection, rigid registration, region of interest transformation, and IHC slide analysis have been embedded into the viewer using the plugin architecture. Biomarker quantification and nuclear segmentation algorithms, written in MATLAB, have also been integrated into the viewer.



1) The Toolbar provides access to controls for image navigation and allows the display layout to be customized. Screen shot, format conversion and cropping tools are also available. 2) The File Manager panel allows the order of images and their visibility to be controlled. Opacity of images and transformation parameters can also be modified which supports manual alignment. 3) The Overlay Manager provides tools for annotation and mark-up of images. Drawing tools include polygons, rulers, arrow and freehand shapes. Annotations can be exported in a human readable XML format and can be imported into other programs and applications. 4) The Analysis Manager allows plugins to be loaded. The SDK provides widgets that allow parameters controlling the analysis application to be set by the user. Applications are distributed as plugins which can be selected from a pull-down menu at run time.

In addition to seven software releases focused on adding support for Visual Studio, x64bit support, more imaging formats and improved support for Matlab-based plug-ins, four plug-ins were released as open-source. One of the plug-ins was for radiology pathology fusion (the rigid registration, and exporting transformed regions of interest). We have demonstrated deploying machine learning models using the Sedeen SDK and Tensorflow C++ API, Boost C++ Library, and Keras (for Python scripting). Hierarchical Normalized Cuts (HNCut) algorithm [1] combines the normalized cuts algorithm with mean shift clustering. HNCut can help pathologists to both quantify and annotate immunohistochemically stained slides by allowing them to identify all pixels that fit within a specific color space. In the TMA Spot Extraction plugin, a color deconvolution algorithm is applied to the down sampled whole slide image to disassociate the hematoxylin and eosin (H&E) staining channels. In the Veta Watershed Segmentation Plugin, an unsupervised automatic method applies color deconvolution and morphological operations to the digital pathology images, followed by the fast-radial symmetry transform to obtain the candidate nuclei locations, which act as markers for a marker-controlled watershed segmentation. The out-of-focus detector addresses a problem that's common in the digital pathology space of finding areas that are out of focus, so that these areas are flagged for human and/or algorithm review.

Poster 7: Informatics Tools for Tumor Heterogeneity in Multiplexed Fluorescence Images

S. Chakra Chennubhotla[1*] and D. Lansing Taylor[1,2] and Brion Sarachan[3*]

[1]Department of Computational and Systems Biology, [2]Drug Discovery Institute, University of Pittsburgh
[3*]Software Science and Analytics Organization, GE Global Research Center, Niskayuna, NY,
chakracs@pitt.edu, sarachan@ge.com

Spatial intratumoral heterogeneity (ITH), quantified as the number and variation of cell phenotypes, as well as the spatial relationships between cells and extracellular molecules within a tumor microenvironment (TME), is of high prognostic and diagnostic value. The acknowledgement of spatial ITH as a key factor in tumor progression has identified a need for new informatics tools to quantify spatial heterogeneity in cancer research applications.

Toward this end, we have created an open source tool, THRIVE (Tumor Heterogeneity Research Interactive Visualization Environment), which 1) permits visualization of large cohorts of whole slide images and tissue microarrays; 2) performs interactive image analysis tasks such as cell segmentation, cell phenotyping, and tumor microdomain discovery via ITH, and 3) contains statistical inference tools to aid in cancer-specific hypothesis testing. We adopt the term tumor microdomain to describe phenotypically distinct regions of the TME, which represent a fundamental unit of spatial heterogeneity. This software platform encapsulates a workflow for quantifying ITH in immunofluorescence (IF) images ranging from a single biomarker to standard multiplexed biomarkers (up to 7) to emerging hyperplexed (>7) images. Each additional biomarker in IF images allows for more insight into cellular and disease mechanisms, but increases cost and data acquisition complexity, so it was important to develop a platform applicable to a range of imaging modalities.

Existing image analysis tools such as CellProfiler, ImageJ/Fiji, and BioimageXD, while useful, are very general tools and thus contain only several of the required features necessary for analyzing spatial ITH, especially from multiplexed and hyperplexed IF images. While some of these contain co-localization pipelines for measuring spatial coincidence of biomarkers within single cells, THRIVE incorporates novel information theoretic measures and current ecological diversity metrics to enhance insights into the spatial organization of tumors by looking at interactions between cells in the TME. We provide the added benefit of designing algorithms with high dimensional image data in mind, collected through multiplexed immunofluorescence, mass spectrometry, or other data collection methods that allow for a large array of molecular probes. THRIVE allows for the creation of custom workflows with plug-in architecture for new functions, can potentially link to genomic and clinical data, and provides multiple spatial and population based heterogeneity metrics, for ease of use by cancer biologists and clinicians alike.

Poster 8: Extensible Open-Source Zero-Footprint Web Viewer for Oncologic Imaging Research

Erik Ziegler[1], Trinity Urban[1,2], Rob Lewis[1], Chris Hafey[1], Cheryl A. Sadow[3], Annick D. Van den Abbeele[4], and Gordon J. Harris[1,2]

[1]Open Health Imaging Foundation, [2]Massachusetts General Hospital, [3]Brigham and Women's Hospital, [4]Dana-Farber Cancer Institute, gjharris@partners.org

The goal of our ITCR U24 grant is to create a vendor-neutral, open-source, extensible, zero-footprint web imaging viewer and libraries for analysis and display of DICOM images, and to build tools on this platform for longitudinal analysis of patients enrolled in oncology clinical trials. For the system architecture, we have implemented Core DICOM libraries such as dicomParser, Cornerstone Core, Cornerstone Tools, and Cornerstone WADO Image Loader (https://github.com/cornerstonejs). The application framework is built on Meteor, a full-stack JavaScript framework which offers reactive UI rendering, and a flexible MongoDB schema-less database for easy development. The application can be split into packages so that third-parties can easily integrate and extend to construct their own zero-footprint imaging applications (https://github.com/OHIF/Viewers). Our project provides two reference applications: LesionTracker for longitudinal analysis, and OHIF Viewer for general purpose DICOM viewing functionality. The viewers are completely web-based and support DIMSE (C-FIND, C-MOVE, etc.) and DICOMWeb (QIDO-RS, WADO-URI, WADO-RS).

Our roadmap includes framework and interoperability support such as DICOM-SR for measurements, and DICOM Segmentation for pixel-wise masks. We are also working towards a simplified deployment against the hospital PACS or as a complete containerized solution. An objective for the LesionTracker viewer is to get the application into routine clinical trials' use. To achieve this, the LesionTracker application has been modified and integrated with the Precision Imaging Metrics (PIM) clinical trials informatics platform. The PIM system is a NCI-shared resource developed by our team at the Dana-Farber/Harvard Cancer Center and used at seven NCI-designated Cancer Centers (with two additional sites under agreement and three others in contract review), and is currently used to manage over 25,000 clinical trial imaging assessments per year. The clinical trials web viewer will be deployed at the first PIM site during summer 2018.

Our U24-supported web imaging project has achieved widespread adoption, and roughly a dozen ITCR-funded projects are using our OHIF Viewers and/or Cornerstone libraries to create web-enabled interfaces (projects include QIICR, 3D Slicer, Martinos QTI, Radiomics, TCIA, LabCAS, THRIVE, CaPTK, C-BIBOP, XNAT, and PET-CoSeg). We have created strong collaborations with a variety of ITCR-funded projects such as Crowds Cure Cancer (Martinos Quantitative Tumor Imaging Lab, TCIA), and Quantitative Image Informatics for Cancer Research (QIICR), both of which showcased our web-viewer at RSNA, November 2017. We also first publicly demonstrated LesionTracker with a new easy installer package at the RSNA 2017 conference, and in the months following the conference, there were over 150 downloads of the package. Recently, the popular Osirix medical image analysis platform announced a new web-enabled version which is built upon the OHIF Viewer. Additionally, numerous industry products have implemented the web viewer packages or libraries into their commercial products such as Ascend Health IT, Asteris, OnePacs, Infervision AI, and Koios. We are continuing to develop the Cornerstone/OHIF platform with added functionality, community collaborations, documentation, and quality management processes.

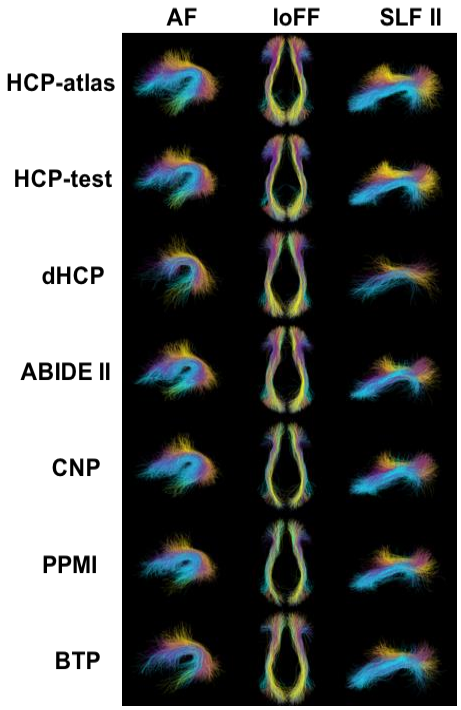
Poster 9: SlicerDMRI: Open-source Diffusion MRI for Cancer Research

Isaiah Norton[1] and Lauren O'Donnell[1]

[1]Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, odonnell@bwh.harvard.edu

Diffusion magnetic resonance imaging (dMRI) is the only non-invasive imaging modality that can map the human brain's connections, which are called white matter fiber tracts. dMRI is also important for the study of the brain tissue microstructure. dMRI works by measuring the diffusion of water molecules. This random molecular motion of water provides unique information because it is affected by the size, shape, and orientation of cells in the brain. dMRI requires the acquisition of multiple images, each sensitized to diffusion in a different spatial orientation, and from this raw diffusion-weighted image data, computational methods[1] are applied to measure tissue properties and to trace brain connections.

SlicerDMRI[2] (http://dmri.slicer.org) software enables diffusion magnetic resonance imaging analysis and visualization in the popular 3D Slicer (https://www.slicer.org) open-source platform for medical image research. Disseminated as an extension through the Slicer Extension Manager "app store," SlicerDMRI has increased in popularity over the past year, with a current average of over 260 downloads per month.



We have made multiple recent improvements to SlicerDMRI. We released a new module for interactive multi-fiber tractography, which allows medical researchers to interactively visualize brain fiber tract connections. We released a new white matter fiber tract atlas, which describes a total of 256 brain connections and was created using a combination of machine learning and expert anatomical annotation. Experiments using our open-source code and new atlas demonstrate robust identification of fiber tracts in an independent test dataset (n=584) acquired on multiple MRI scanners and across the lifespan. At right, three tracts (columns) identified in 7 different datasets (rows).

We have put significant effort into improving online documentation and code performance, as well as addressing SlicerDMRI user questions and requests in the active Slicer Forum (https://discourse.slicer.org). We have improved infrastructure in Slicer to enable asynchronously calling python modules via the command-line module infrastructure, to enable our work in progress integrating dMRI python packages to expand the scope of available processing. Finally, we co-organized an international software "connectathon" for the new Digital Imaging and Communications in Medicine (DICOM) standard for storing fiber tractography (http://archive.rsna.org/2017/17008539.html). Our U01 funded the first open-source implementation of the standard. Connectathon results demonstrated interoperability with multiple academic and commercial software platforms, which can enable future interoperability with FDA-approved systems for neuronavigation (image guidance during brain surgery).

References

1. O'Donnell LJ, Daducci A, Wassermann D, Lenglet C. Advances in computational and statistical diffusion MRI. NMR Biomed [Internet]. Wiley Online Library; 2017 Nov 14; Available from: http://dx.doi.org/10.1002/nbm.3805 PMID: 29134716

2. Norton I, Essayed WI, Zhang F, Pujol S, Yarmarkovich A, Golby AJ, Kindlmann G, Wassermann D, Estepar RSJ, Rathi Y, Pieper S, Kikinis R, Johnson HJ, Westin C-F, O'Donnell LJ. SlicerDMRI: Open Source Diffusion MRI Software for Brain Cancer Research. Cancer Res. 2017 Nov 1;77(21):e101–e103. PMCID: PMC5679308

Poster 10: Informatics Tools for Optimized Imaging Biomarkers for Cancer Research & Discovery
U24 CA180927

Jayashree Kalpathy-Cramer[1], Robert Gillies[2], Dmitry Goldgof[3], Sandy Napel[4], Binsheng Zhao[5], Bruce Rosen[1]

[1]Massachusetts General Hospital, [2]Moffitt Cancer Center, [3]Univ. South Florida, [4]Stanford University, [5]Columbia University Medical Center

The Cloud-based Image Biomarker Optimization Platform (C-BIBOP) is a technical resource for the cancer research community that enables developers of image processing algorithms to share and compare their lesion-segmentation and feature characterization algorithms on publicly available and private data. Biologists can use these algorithms to integrate image phenotype data with molecular and clinical data to better understand the manifestations of cancer, and clinical researchers can use them to derive robust image biomarkers of specific cancer types which will, in turn have utility for precision therapy and monitoring of response. In this project, we are developing both the infrastructure for sharing algorithms as well as the algorithms themselves. Algorithms include tumor segmentation of lung field, lung nodules and brain tumors, radiomics pipelines for the characterization of these tumors and tools for statistical analysis and visualization.

The C-BIBOP platform also supports "challenges", specifically geared towards the imaging analysis community, that enable comparisons of algorithms for image analysis in cancer. We have been working closely with NCI's Quantitative Imaging Network (QIN) to support challenges and collaborative projects. We have also supported challenges at a number of leading conferences including at MICCIA, SPIE, RSNA and others. We have supported a number of collaborative projects and challenges with more than a 1000 individual and team registrations. Our recent developments provide a paradigm shift in challenge methodology where participants upload code, as Docker containers, instead of uploading results of their image analysis pipelines. This enhances reproducibility of the results, allows comparison of methods on similar hardware and allows the algorithm's performance to be estimated on unseen data.

Algorithms within C-BIBOP include those for lesion segmentation and radiomics feature extraction. We will share the results of our deep learning brain tumor segmentation algorithm as well as results of multisite radiomics feature comparison collaborative project. We are utilizing our tools to analyze cancer data including from the IvyGAP collection.

Finally, in order to obtain annotations that can be used to develop machine learning algorithms for image analysis, we conducted a large crowdsourcing effort at the annual Radiological Society of North America (RSNA) conference in November of 2017. The goal of this effort was to crowd source the annotation of over 400 CT studies in TCIA of 4 different cancer types (liver, lung, ovarian, renal) through the volunteer efforts of radiologists attending RSNA. The platform used for this effort was a collaborative project between 4 ITCR groups including ours, QIICR (Kikinis/Fedorov ITCR U24), the zero-footprint viewer (Harris ITCR U24) and TCIA (Prior). Our preliminary analysis suggests that crowd sourcing, especially at events such as RSNA, can be an effective means of acquiring relatively high quality annotations.

Poster 11: The Digital Slide Archive – A Web-Based Platform for Collaborative Pathology Research

Lee AD Cooper[1,2,4], David A Gutman[3,4]

[1]Department of Biomedical Informatics, Emory University School of Medicine
[2]Department of Biomedical Engineering, Emory University / Georgia Institute of Technology
[3]Department of Neurology, Emory University School of Medicine
[4]Winship Cancer Institute, Emory University
Presenter: david.gutman@emory.edu

Increasing adoption of digital pathology scanners is producing large volumes of whole-slide images (WSIs) in both clinical and research domains. Informatics tools to enable the management, sharing, annotation, and quantitative analysis of WSI data are needed support cancer research activities, ranging from support for tissue banking Q&A, to central review for diagnostic confirmation in cohort studies, and basic science investigations. Digital pathology data presents a several unique challenges, a number of which are not adequately addressed by commercially available software. DSA enables investigators to fluidly pan and zoom through images containing multiple billions of pixels. Annotation capabilities allow users to annotate regions or objects in images, and to share deep linked views of specific regions and magnifications of specimens with collaborators. Finally, a suite of image analysis algorithms called *HistomicsTK* allows users to analyse images to generate quantitative descriptions of histology.

This talk will describe recent advances and new features in the DSA, and demonstrate how DSA can be used to facilitate a variety of cancer research applications. In the past year, the annotation capabilities of DSA have been extensively developed and validated through user feedback. In addition, we have focused on additional image and metadata management features that increases the flexibility of DSA to manage histology, radiology and clinical metadata in a single platform with strong API support.

Poster 12: TCIA Sustainment and Scalability - Platforms for Quantitative Imaging Informatics in Precision Medicine (PRISM)

Ashish Sharma[1], Joel Saltz[2], Lawrence Tarbox[3], Kirk Smith[3], Tracy Nolan[3], Jonathan Bona[3], Annie (Ping) Gu[1], Erich Bremer[2], Tammy DiPrima[2], Jonas Almeida[2], Mathias Brochhausen[3] Tahsin Kurc[2], Fred Prior[3]

[1]Department of Biomedical Informatics, Emory University
[2]Department of Biomedical Informatics, Stony Brook University
[3]Department of Biomedical Informatics, University of Arkansas for Medical Sciences

The identification of imaging phenotypes across scale has dramatically improved our understanding of cancer biology and response to therapy. Increasingly image phenotyping is achieved through the use of quantitative image analysis, machine learning, or deep learning. Such methodologies require large collections of well-curated data for algorithm development and validation. The Cancer Imaging Archive (TCIA) is NCI's primary resource for acquiring, curating, managing and distributing images and related data to support Cancer Research.

Through discussions with TCIA users, as well as research programs such as the NCI Quantitative Imaging Network, RSNA, etc., our team has identified a set of near-term challenges that must be met to enhance TCIA's relevance to the research community. These include the ability to manage new data types, including Digital Pathology, Clinical Data, and Quantitative Imaging Features, and use these data types when creating data cohorts. Such capabilities require the ability to integrate, explore and access multiple datasets using the full metadata. Expansion of data publication tools to support reproducible research and expanded support for challenge organizations including a modularized, cloud-deployable technology stack was also emphasized.

This stack is envisioned as a **Platform for Imaging in Precision Medicine (PRISM)** and its adoption will help improve TCIA's ability to manage and analyze integrated datasets. Here integrated datasets refer to tightly coupled Radiology and Pathology images, Radiation Therapy data, clinical context including annotations, markups, and features extracted from Radiology and Pathology studies.

While TCIA collects various non-imaging data types, TCIA lacks a common representation scheme and frequently stores these as spreadsheets. To address this shortcoming, a key component of PRISM is to improve semantic integration of non-image data. To this end, we are in the process of developing a prototype using semantic web technology to improve semantic integration. In our semantic integration prototype, we are converting these to representations that use biomedical ontologies to make the data queryable and accessible for automated reasoning across collections. The result of this work is a semantic graph database with assertions linking patient identifiers to RDF instances representing patients, affected body parts, diagnoses, etc., and relations among those.

Poster 13: Developing Enabling PET-CT Image Analysis Tools for Predicting Response in Radiation Cancer therapy

Xiaodong Wu[1,2], Yusung Kim[2] and John Buatti

[1]Department of Electrical and Computer Enigneering
[2]Department of Radiation Oncology
University of Iowa
{**xiaodong-wu**, yusung-kim, john-buati}@uiowa.edu

Positron emission tomography - computed tomography (PET-CT) has brought about a revolution in modern cancer therapy, from diagnosis, staging, response prediction, treatment planning, to prognosis assessment. In this project, we propose to develop novel algorithms, methods, and general tools for automated and objective analysis of PET-CT images to facilitate the use of the dual modality imaging in the response prediction for radiation therapy.

We have developed an efficient graph-cut based method for simultaneous co-segmentation of tumors from both PET and CT scans while admitting the boundary differences between PET and CT, improving segmentation accuracy. Recently, deep learning has been shown to excel at wide variety of medical image segmentation tasks, due to its ability to learn rich expressive feature representations automatically from data, as opposed to the traditional hand-crafted features. However, in contrast to the graph-cut approach, the deep learning segmentation methods show insufficient to learn global contextual information and multi-scale spatial correlations among image volumes. In computer vision, Conditional Random Fields (CRFs), which explicitly model the contextual and spatial correlations among voxels, has been successfully incorporated into convolutional neural networks (CNNs), yielding state-of-the-art segmentation performance. We thus model our graph-cut co-segmentation method under the framework of CRFs, and to seamlessly integrate directly in CNNs to allow for end-to-end training, resulting in a novel deep-learning based method for PET-CT co-segmentation. Experiments on 60 PET-CT scan pairs of patients with non-small cell lung cancer demonstrated the effectiveness and efficiency of the proposed method.

Following the co-segmentation neural network, additional CNNs are used to extract predictive features from tumors and tumor margins in PET and CT. Traditional binary classification approach for therapeutic response/survival analysis has its severe limitation to be adopted in the CNN framework due to the highly unbalanced event probabilities. Data augmentation techniques are adopted to address this problem, where randomized rotations and transformations of contrast and brightness are used to synthesize additional training data. Two datasets will be used to validate the predictive model: (1) In-house ^{18}F-FDG PET-CT datasets of non-small cell lung cancer (NSCLS), and (2) the NSCLC Radiogenomics from Stanford University at TCIA.

Poster 14: Development and Validation of Informatics Tools for Immunohistochemistry Analysis in Multi-Institutional Cohort Studies

Lee AD Cooper[1,2,4], Christopher R Flowers[1,3,4], Metin N Gurcan[5], Deepak Chittajallu[6]

[1]Department of Biomedical Informatics, Emory University School of Medicine
[2]Department of Biomedical Engineering, Emory University / Georgia Institute of Technology
[3]Department of Hematology and Medical Oncology, Emory University School of Medicine
[4]Winship Cancer Institute, Emory University
[5]Center for Biomedical Informatics, Wake Forest School of Medicine
[6]Kitware, Inc.
Presenter: lee.cooper@emory.edu

Biomarker-driven prognostication is critical for realizing precision treatment strategies and population health management approaches that optimize quality of life and survival for patients. Genomics holds promise for improving the classification and prognostication of malignancies, yet oncology practice continues to rely heavily on immunohistochemically stained tissues (IHC) as a fundamental tool due to its practicality and ability to provide *protein-level* and *subcellular localization* information. Evaluating IHC remains a largely manual and subjective practice outside of breast cancer, where expensive commercial software is used to score ER/PR/HER2. Commercial tools generally cannot account for variations in tissue processing present in multi-institution studies, do not leverage new advances in image analysis and learning algorithms, do not account for intratumoral heterogeneity in IHC staining, and cannot be meaningfully extended or integrated with existing resources to optimize classification and prognostication strategies. New open-source tools are needed for robust IHC analysis and for the integration of IHC measurements with genomics and clinical outcomes.

The goal of this project is to create open-source tools for the quantitative analysis of immunohistochemically (IHC) stained tissues IHC and to enable improved classification and prognostication through the integration of IHC, clinical, and genomic features. These tools will enable pathologists and clinical translational researchers to integrate quantitative IHC measurements with clinical and genomic information to investigate strategies for classifying malignancies and predicting clinical outcomes in large multi-institution cohort studies and clinical trials. The proposed tools will be developed and extensively validated in close collaboration with the NCI-supported *Lymphoma Epidemiology of Outcomes* (LEO) cohort study. Encompassing 8 academic centers, LEO provides a realistic environment where the impact of tissue processing variability can be evaluated, and where methods for standardization and mitigating these variations can be developed and validated.

The software tools produced by this proposal will enable standardization of whole-slide imaging datasets, as well as the characterization of protein expression in subcellular compartments and at the specimen/patient level. Using diffuse large b-cell lymphomas as a driving application, we will develop technology to classify patients based on IHC measurements, including automatic tuning of machine learning algorithms that will enable a broad class of clinically and biologically motivated users to utilize these tools in their investigations. These tools will be released and documented under an open-source model, integrated with HistomicsTK (https://histomicstk.readthedocs.io/en/latest/), and available to the broader cancer research community.

Poster 15: EMERSE: The Electronic Medical Record Search Engine

David Hanauer[1,2,3]

[1]Department of Pediatrics, [2]Rogel Cancer Center, [3]Michigan Institute for Clinical and Health Research, University of Michigan, hanauer@umich.edu

The nationwide adoption of electronic health record (EHR) systems has magnified the possibilities for advancing clinical and translational research. EHR data are routinely used for patient eligibility determination, enriching clinical research studies, and supplementing -omics research with detailed phenotypic data. Yet the promise of utilizing EHR data are limited because much of the richest, most comprehensive data are 'trapped' within the free text clinical notes and siloed within institutions. Further, many software tools cannot meet the complex and rapidly changing research needs for phenotypic data. To address these data needs, our cancer center developed EMERSE (Electronic Medical Record Search Engine) to support the information retrieval (IR) needs for cancer information from EHRs. EMERSE helps users overcome the challenges associated with performing complex information extraction tasks from EHR 'big data'. Through over a decade of work, EMERSE has become both widely used and highly regarded by our cancer center researchers and other teams across our health system. EMERSE has supported over 1,000 studies and more than 160 peer-reviewed publications.

EMERSE supports features that set it apart from other IR tools and make it a valuable asset for the cancer research community. One important EMERSE feature is the ability to share saved searches among users in the form of 'Bundles'. EMERSE also: (1) supports patient lists, which are essential for studies in which cohorts must be searched on an ongoing basis (e.g. adverse events in clinical trials); (2) provides a powerful query expansion feature that includes concepts derived from search logs as well as resources such as the NCI Consumers' Cancer Dictionary; (3) supports customized use of colors for highlighting terms, allowing users to group related concepts with the same color for rapid visual scanning of results; and (4) provides visualizations of search results including one modeled after a heat map and another novel 'Mosaic' view.

With support from the ITCR program we are developing advanced features within EMERSE and are preparing the software for implementation at other sites. The complexities of such 'enterprise-level' software implementations should not be underestimated: with differing IT infrastructures, EHRs, data sources, metadata, and regulatory environments, a high level of effort is required to get tools like EMERSE installed and available for the research community. To enable our dissemination efforts, we are developing a user community as well as a documentation 'Roadmap' to support implementation and ongoing usage. Major features planned include handling of de-identified clinical notes and networking the software instances for cross-institutional patient counts and cohort discovery. The initial partner sites supported through our ITCR collaboration include: University of North Carolina, University of Cincinnati, Case Western Reserve University, University of Kentucky, and Columbia University.

EMERSE can be a valuable tool in the 'toolbox' of other software applications commonly used in clinical research including the i2b2 Workbench and REDCap. EMERSE's focus on the unstructured text provides a valuable complement to these other commonly used research tools. More information on EMERSE can be found at http://project-emerse.org.

Poster 16: DeepPhe - A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records

Sean Finan[1], MS, James Masanz[1], MS, Olga Medvedeva[3], Eugene Tseytlin[4], MS, Melissa Castine[4], Timothy Miller[1,2], PhD, Olga Medvedeva[3], David Harris[1], Harry Hochheiser[3], PhD, Chen Lin[1], MS, Hadi Amiri[1,2], PhD, Girish Chavan[4], MS, Jeremy L. Warner[5], MD, MS, Rebecca Jacobson[4], MD, MS, Guergana Savova[1,2] PhD

[1]Boston Children's Hospital, Boston, MA; [2]Harvard Medical School, Boston, MA; [3]University of Pittsburgh Department of Biomedical Informatics, Pittsburgh, PA; [4]UPMC Enterprises, Pittsburgh, PA; [5]Vanderbilt University, Nashville, TN
Guergana.Savova@childrens.harvard.edu

Background: The labor required for manual extraction of detailed synopses of complex cancer treatment histories presents significant challenges for cancer surveillance and retrospective research. Although natural language processing techniques for automating the extraction of key data from clinical notes show much promise, prior efforts have almost exclusively focused on individual documents (e.g., pathology notes). DeepPhe extends these efforts, using a combination of novel techniques for linking concepts across documents (cross-document co-reference resolution) and domain-specific rules to aggregate individual entity mentions into detailed longitudinal summaries over the entire set of a patient's records.

Purpose: We describe the DeepPhe system for extracting rich cancer phenotypes, including the application of DeepPhe to sample data for three cancers (breast, ovarian, and melanoma) from four cancer sites, and the display of relevant results through a preliminary visualization tool.

Methods: DeepPhe ingests multiple clinical documents and optionally discrete data, and outputs a single summary of the patient's clinical phenotype. DeepPhe's processing pipeline extend the cTAKES© system[1], adding components for cross-document co-reference, identification of care episodes, rule-based summarization, and output of resulting data in multiple formats, including FHIR, i2b2/tranSMART, and graph database (Neo4j) representations. The patient visualization tool supports exploration of patient records through displays capable of linking summarized results to individual patient notes through an interactive timeline.

Results: DeepPhe system v2 is currently being trained on breast, skin and ovarian cancer data. DeepPhe system v1 and its evaluation results were described in Savova et al.[3] We will present results with DeepPhe v2 on breast, ovarian and skin cancers across data from four sites. In addition, DeepPhe system v2 implements an episode classifier to group documents into six categories – *pre-diagnostic, diagnostic, decision-making, treatment, follow-up, unknown*. Early evaluation results show excellent performance.

Conclusion: Extraction of specific attributes is promising but requires refinement. Our study emphasizes the importance of research in challenging areas including word sense disambiguation, relation extraction (e.g., coreference, temporal & body location relations) and summarization. Future work will focus on improving individual components, adding support for extraction of genomic observations, and extending the patient visualization tool with cohort visualization capabilities for collections of DeepPhe results. The DeepPhe system v2 will be available for download from http://github.com/DeepPhe/DeepPhe-Release in the summer of 2018.

References

1. Savova G et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES). doi: 10.1136/jamia.2009.001560
2. Hochheiser H et al. An Information Model for Cancer Phenotypes. doi: 10.1186/s12911-016-0358-4
3. Savova G et al. DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. doi: 10.1158/0008-5472.CAN-17-0615

Poster 17: Advancing Cancer Pharmacoepidemiology Research through EHRs and Informatics

Liwei Wang[1] (wang.liwei@mayo.edu), Lei Luo[1,2], Jeremy L. Warner[3], Yanshan Wang[1], Jason A. Wampfler[1], Hua Xu[4], Ping Yang[1], Hongfang Liu[1]

[1] Dept. of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN, U.S.
[2]Dept. of Good Clinical Practice, Guizhou Province People's Hospital, Guiyang, China
[3]Depts. of Medicine and Biomedical Informatics, Vanderbilt University, Nashville, TN, U.S.
[4]School of Biomedical Informatics, UT Houston, TX, U.S.

Lung cancer is the second most common cancer and by far the leading cause of cancer-related death in both males and females, accounting for 1 in 4 cancer deaths in U.S. Accurate identification of lung cancer related information is very important for epidemiological studies, therefore is critical for improving cancer outcomes. Since epidemiologists use electronic health records (EHR) with rich longitudinal data on large populations for epidemiologic research but manual extraction from large volumes of text materials is time consuming and labor intensive, considerable efforts have emerged to automatically extract information from text for lung cancer patients using natural language processing (NLP). As part of the larger U24 grant (Specific Aim 3, Project 2), we developed and evaluated an NLP system in capturing information on stage, histology, grade, chemotherapy, radiotherapy and surgery in lung cancer patients using various narrative data sources from EHR including clinical notes, pathology reports and surgery reports.

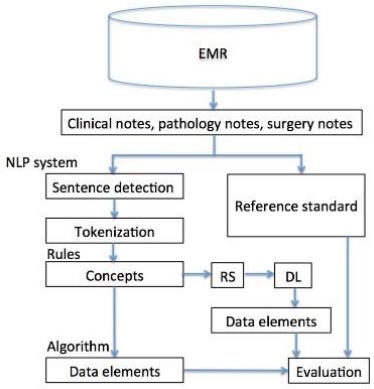


Figure SEQ Figure * ARABIC 1. Study design. EMR: Electronic Medical Record, RS: related sentences, DL: deep learning

We used an existing cohort including 2,311 lung cancer patients with information about stage, histology, grade, and therapies manually ascertained. Based on the cohort, a NLP system was developed using the open source clinical NLP pipeline MedTagger as the platform [1]. Specifically we utilized the sentence detection and tokenization parts in MedTagger. Then the NLP system integrated rules and algorithm to output final normalized concept names for each data element. We finally evaluated the output of NLP system against the human abstracted results from the existing dataset. Deep learning was used to predict values for data elements using sentences labeled by NLP system as input. Then we analyzed NLP results, deep learning prediction results and the reference standard from the existing cohort for error analysis in terms of histology extraction.

Evaluation showed promising results with the recalls for stage, histology, grade, and therapies achieving 89%, 98%, 78%, and 100% respectively and the precisions were 70%, 88%, 90%, and 100% respectively. Error analysis in 100 patients indicated that the NLP system helped to identify more specific histological types, e.g., adenocarcinoma in 8 patients that were not provided in the reference standard, and identify the correct histology type in 1 patient who was mistakenly identified as another type in the reference standard. Findings showed that among 4 cases misidentified by NLP system and deep learning, 2 had no related information recorded and 2 had related information extracted but missed due to the priority of different data sources in our algorithm.

This study demonstrated the feasibility and accuracy of extracting related information from clinical narratives for lung cancer research. Efforts to map the output to our chemotherapy regimen ontology[2] (Specific Aim 2) are ongoing.

References

1. Liu H, Bielinski SJ, Sohn S, et al. An information extraction framework for cohort identification using electronic health records. AMIA Summits on Translational Science Proceedings 2013;2013:149.
2. Malty A, Jain SK, Yang PC, et al. A computerized approach to creating a systematic ontology of hematology/oncology regimens. JCO Clinical Cancer Informatics 2018 (in press)

Poster 18: Scalable Clinical Decision Support for Individualized Cancer Risk Management

Guilherme Del Fiol[1] and Kensaku Kawamoto[1]

[1*]Department of Biomedical Informatics, University of Utah, guilherme.delfiol@utah.edu

Over 10% of US adults are at elevated risk of familial cancers such as breast and colorectal cancer. Increased evidence supports individualizing cancer screening based on risk, with selective application of specific screening and prophylaxis interventions best suited to the individual. Higher risk individuals can be identified through family history assessment in primary care settings. However, many of the affected patients (younger adults under ages of 40 or 50) are seen only occasionally for primary care, and electronic tools are generally not available for enabling efficient cancer risk assessment and management in this population in busy primary care settings.

To help address this challenge, we are developing a clinical decision support (CDS) platform that analyzes data from electronic health record (EHR) systems to automatically identify individuals who meet high risk criteria for familial cancers and refer them to genetic counseling. Our CDS approach (i) builds on a stack of open source software from the *OpenCDS* and *OpenInfobutton* projects; (ii) leverages multiple sources of patient data including family history documentation in structured EHR data, free-text clinical notes, and online patient questionnaires; and (iii) aims to integrate with different EHR systems through widely adopted health information technology (IT) standards.

With EHR adoption reaching over 90% of care settings in the U.S., a unique opportunity exists to disseminate our proposed CDS platform to a variety of health care organizations using different EHR systems. In the first project year, we plan to integrate an initial version of the CDS platform with the Epic® EHR for deployment at University of Utah's Community Clinics and the Huntsman Cancer Institute. In subsequent project years, we also plan to assess the interoperability of the CDS platform with a different EHR (Cerner®) at Intermountain Healthcare.

As a part of the first phase of the project, we concluded a systematic evaluation of family history workflow and documentation patterns in primary care. We observed and interviewed over 150 medical assistants, nurses, and physicians at 12 primary care Community Clinics that are part of University of Utah Health. We identified 7 distinct workflow patterns with family history information being documented in at least 5 different locations in the EHR. These workflows may generalize to other similar primary care settings that are associated with academic medical centers. We are using the findings of this study to elicit requirements for the CDS platform. A manuscript reporting the results of this work is under development.

Poster 19: PDX Finder: A Portal for Patient-Derived Tumor Xenograft Model Discovery

Terrence F. Meehan[1], Nathalie Conte[1], Jeremy Mason[1], Csaba Halmagyi[1], Steven Neuhauser[2], Abayomi Mosaku[1], Dale A. Begley[2], Debra M. Krupke[2], Helen Parkinson[1], Carol Bult[2]

1 European Molecular Biology Laboratory- European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

2 The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

Patient-derived tumor xenograft (PDX) mouse models are a versatile oncology research platform for studying tumor biology and for testing chemotherapeutic approaches tailored to genomic characteristics of individual patient's tumors. PDX models are generated and distributed by a diverse group of academic labs, research organizations, multi-institution consortia, and contract research organizations. The distributed nature of PDX repositories and the use of different standards in the associated metadata presents a significant challenge to finding PDX models relevant to specific cancer research questions. The Jackson Laboratory and EMBL-EBI are addressing these challenges by co-developing PDX Finder, a comprehensive open global catalog of PDX models and their associated datasets. Within PDX Finder, model attributes are harmonized and integrated using a previously developed community minimal information standard to support consistent searching across the originating resources. Links to repositories are provided from the PDX Finder search results to facilitate model acquisition and/or collaboration. The PDX Finder resource currently contains information for more than 1900 PDX models of diverse cancers including those from large resources such as the Patient-Derived Models Repository, PDXNet and EurOPDX. Individuals or organizations that generate and distribute PDXs are invited to increase the "findability" of their models by participating in the PDX Finder initiative at www.pdxfinder.org.

Poster 20: Trinity: Transcriptome assembly for genetic and functional analysis of cancer

Brian Haas[1*], Asma Bankapur[1], Christophe Georgescu[1], Vrushali Fangal[1], Carrie Ganote[2], Cicada Brokaw[2], Thomas Doak[2], Aviv Regev[1]

[1]Broad Institute, Cambridge, Massachusetts
[2]Indiana University, Bloomington, Indiana
[*]presented by: bhaas@broadinstitute.org

RNA-Seq studies indicate that cancer transcriptomes are shaped by genetic changes, variation in gene transcription, mRNA processing, editing and stability, and the cancer microbiome. Deciphering this variation and understanding its implications on tumorigenesis requires sophisticated computational analyses, and the ability to analyse bulk RNA-Seq as well as transcriptomes of individual tumor cells. Most RNA-Seq analyses rely on methods that first map short reads to a reference genome, and then compare them to annotated transcripts or assemble them. However, this strategy can be limited when the cancer genome is substantially different from the reference or for detecting sequences from the cancer microbiome. 'Assembly first' (*de novo*) methods that combine reads into transcripts without any mapping are a compelling alternative. The assembled transcriptome can then be used to identify mutations, fusion transcripts, splicing patterns, expression levels, tumor-associated microbes, and – if collected from single cells – characterize tumor heterogeneity. There is thus an enormous need for computationally efficient, accurate and user friendly tools for transcriptome reconstruction and analysis in cancer. Trinity, first released in 2011 and freely available as open source, is the leading software for *de novo* RNA-Seq assembly, executed millions of times by thousands of researchers, with nearly 6k literature citations, and now including a host of modules for downstream analyses, contributed by the Trinity development team or 3rd party developers.

Through ITCR, we continue to enhance and maintain Trinity and further develop the Trinity Cancer Transcriptome Analysis Toolkit (CTAT) as a leading tool suite for bulk and single-cell cancer transcriptomics. We have tailored analytic modules for critical tasks in cancer biology, working with a network of cancer researchers on Driving Cancer Projects, with a focus on analysing clinical samples. We continue to update the Trinity software to enhance the core algorithm, leveraging new sequencing technologies and integrating genome data with genome-free assembly. We have efforts underway to integrate Trinity CTAT into the NCI cloud computing platform via FireCloud to enable scalable cancer transcriptome data processing and analyses, in addition to developing and maintaining Galaxy-based modules to facilitate cancer transcriptome data exploration.

Future applications and developments of Trinity CTAT target more rigorous characterization of tumor heterogeneity at the single cell level and clinical applications in cancer vaccine development through neoantigen discovery. The algorithm enhancements underway to engineer the next generation Trinity assembler are intended to empower transcriptome assembly and analysis across diverse sample types, sequencing technologies, and data exploration modalities, bringing us closer to more comprehensively analyzing the cancer transcriptome at single cell resolution.

Poster 21: Visualizing Structural Variation with the JBrowse Genome Browser

Robert Buels[1], Eric Yao[1], Lincoln Stein[2], and Ian Holmes[1]

[1]Department of Bioengineering, University of California, Berkeley, ihh@berkeley.edu

[2]Ontario Institute for Cancer Research, Toronto

The JBrowse genome browser is a mature, JavaScript-based genome browser supporting fluid, responsive, browsing of genome annotations and faceted exploration of deep sets of data tracks. JBrowse has been widely adopted in the plant and animal genomics communities, and by a few groups within the cancer informatics community. We are developing JBrowse to further support exploration of cancer informatics data in several ways. First, we are building a framework for genome-oriented client-side visualization components to interoperate, including concertina-collapsed and condensed views, so that multiple JBrowse instances can coexist and talk to one another within the same webpage. This will facilitate synteny views and other side-by-side comparisons of different genomes (or, pertinently for visualizing different regions that have been brought together by an SV). Second, we are developing a circular visualization tool that will enable long-range visualization of breakpoint data from VCF files and similarly-capable formats. Third, we have established a server framework that makes analysis operations available from within the genome browser, linking to workflow engines such as Galaxy on the back end. I will describe our first year of progress toward these goals, including substantial reengineering of the JBrowse codebase and a few design principles for building modern web dashboards for genomics.

Poster 22: Informatic Tools for Single-Nucleotide Analysis of Cancer RNA-seq

Esther Yun-Hua Hsiao[1], Yi-Wen Yang[2], Tracey Chan[3], Stephen Tran[3], Jae Hoon Bahn[2], and Xinshu (Grace) Xiao[1,2,3*]

[1]Department of Bioengineering,
[2]Department of Integrative Biology and Physiology,
[3]Bioinformatics Interdepartmental Program, University of California, Los Angeles,
[*]gxxiao@ucla.edu

We aim to develop informatic tools for single-nucleotide analysis of cancer RNA sequencing (RNA-Seq) data. RNA-Seq is becoming an essential tool in both basic and clinical cancer research. As a result, numerous research groups are generating their own RNA-Seq data sets. In addition, large consortium efforts, such as the TCGA project, are producing an extraordinary amount of RNA-Seq data that are invaluable resources to the research community. The wide adoption of RNA-Seq calls for effective and user-friendly informatic tools that can extract information of important biological relevance. To meet this need, great effort has been dedicated to tool development, resulting in products that are now widely in use, such as short read aligners, transcriptome assembly tools, and methods to detect differential gene expression and alternative splicing. However, a major advantage of RNA-Seq is its capacity to provide information at the single-nucleotide level. Tools that harness this information are relatively scarce. As a result, single-nucleotide analysis is not yet a routine procedure in RNA-Seq informatics. This type of analysis can potentially reveal important biological insights. With sequencing errors excluded, single-nucleotide variants (SNVs) expressed in the RNA reflect existence of genetic variants or RNA editing sites, both of which could be essential players in cancer diagnostics, basic mechanisms and biology. A number of challenges exist in the identification, quantification and functional prediction of these SNVs. We have developed a suite of methodologies to address these challenges. We will further improve these methods and build user-friendly informatic tools and web portals to identify and analyze SNVs in cancer RNA-Seq data. These tools will facilitate a broad spectrum of SNV analysis, ranging from raw read mapping to functional prediction of SNVs in affecting alternative splicing or RNA stability. With no additional experimental cost, information of SNVs is readily extractable in all RNA-Seq data sets. A full exploration of this information could provide novel insights and maximize the scientific value of the still costly RNA-Seq data. Our project will develop tools that will enable incorporation of SNV analysis into routine procedures of RNA-Seq informatics.

Poster 23: Streamlined sharing and analysis of clinical patient data for cancer research networks

Ian Foster

Department of Computer Science, University of Chicago, foster@uchicago.edu

Advances in genomics and data analytics create new opportunities to advance cancer research via large-scale federation of genomic, clinical, imaging and other data from patients across institutions around the world. Successful efforts often require well-annotated patient samples aggregated from multiple sites via national or international networks, with access to substantial computational resources and bioinformatics expertise for large-scale genomic analysis. Yet despite these opportunities and promising early results, cancer research is often stymied by a lack of tools for the reliable, secure, rapid, and easy transfer and sharing of large collections of human data that can be customized to meet the needs of both large and small cancer studies. In the absence of such tools, security and performance concerns often prevent sharing altogether or force researchers to resort to slow and error prone shipping of physical media. If data are received, timely analysis is further impeded by the difficulties inherent in verifying data integrity and managing who can access data and for what purpose.

I will discuss how the Globus research data management platform addresses these obstacles to discovery by extending high-speed, reliable data transfer and sharing technology that automates and optimizes data transfer, sharing, and analysis tasks that would otherwise require unreasonable time, resources, and expertise, and provides intuitive web interfaces for human use and APIs for use in applications. I also describe how Globus technologies are being extended, for example with increased auditing and higher level of identity assurance, to meet the security requirements of human data to enable use in data-intensive cancer research.

Opportunities for partnerships with other ITCR projects will be discussed, such as data portal development, transfer and sharing automation, and federated authentication. I will present illustrative collaborations that apply Globus services, including the development of data distribution and sharing networks for cancer researchers engaged in the study of cancer health disparities among the minority populations of Louisiana and our joint work with Dr. Olufunmilayo Olopade's group at University of Chicago on consensus variant calling of structural variants on a cohort of 420 breast cancer subjects sequenced using targeted gene panel (BRCA). I will also discuss related work within the NIH Data Commons Pilot Project Consortium.

Poster 24: Computational Framework for Single-cell Genomics

Jude Kendall[1], Lubomir Chorbadjiev[2], Vyacheslav Zhygulin[1], Junyan Song[1], Joan Alexander[1], Michael Wigler[1] and Alexander Krasnitz[1*]

[1*]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, krasnitz@cshl.edu
[2]Technological School of Electronic Systems, Technical University of Sofia, Sofia, Bulgaria

While single-cell genomics is by now a firmly established research methodology, its potential as a clinical diagnostic tool still is far from being fully realized. Our ITCR-funded work has been directed towards development of complete computational infrastructure for single-nucleus sequencing (SNS), a method for sparse genomic profiling of individual nuclei, to facilitate applications of SNS both in research and clinical settings. Sparsity of per-nucleus sequencing yield makes it affordable to profile large numbers of nuclei. At the same time, SNS-derived sequencing data are sufficient for detection of cell-specific DNA copy number variants. Given the ubiquity of copy-number variation in cancer, SNS is particularly suitable for cancer genomics.

Our computational pipeline is now near completion and includes modules for sequence read data mapping; normalization and quality control; derivation of integer-valued DNA copy-number profiles of individual nuclei; reconstruction of clonal structures in cell populations using copy-number features shared by multiple cells; and, finally, data visualization suite. Our tools are easily deployed using Anaconda environment. Modular structure of the pipeline makes it readily adaptable, to keep up with advances in single-cell technology.

Diagnostic utility of both bench and computational components of SNS has recently been demonstrated in a pilot study, where clonal involvement, diversity, complexity and spread were determined and found to complement findings of conventional cytopathology in a series of needle biopsies of in men being diagnosed for prostate cancer. Our results suggest, in particular, that SNS profiling of diagnostic biopsies is an accurate predictor of the tumor grade as determined upon prostatectomy.

Poster 25: The Network Data Exchange in 2018

Trey Ideker[1], Dexter Pratt[1]

[1]UCSD School of Medicine, depratt@ucsd.edu

NDEx is a public database of networks, a resource for cancer researchers to store, share, and publish biological networks. It is a source for networks of many types and sizes and it is a data hub for software using networks. NDEx is tightly integrated with Cytoscape, bringing cloud storage and distribution to the most widely used platform for visualization and analysis of biological networks.

Scientists can store networks in their private NDEx accounts and share their networks with collaborators: in effect, NDEx can be thought of as "DropBox for networks". A 10 GB default storage allowance enables NDEx to be used in everyday work. Networks can be shared with anyone, not just NDEx users, with anonymous access URLs. These links that can be easily used in emails, chat, twitter, and other communication channels. Networks stored in NDEx can range from small pathway networks, to million-edge interaction networks, to data-driven networks loaded directly from analysis pipelines. In the last year, the number of public reference networks available through NDEx has greatly expanded and work is in progress to build pipelines for the automation of periodic updates.

Access to NDEx access is built into the latest release of Cytoscape, enabling users to find networks in NDEx, to download networks for visualization and analysis, and to save networks to their NDEx account. A recently added feature of the NDEx web interface enables one-click download of networks to Cytoscape.

Programs can use NDEx via its REST API to find networks, download networks, and to save networks, typically using the NDEx client libraries available for Python, R and Java. This enables the integration of NDEx with applications, such as the CRAVAT variant analysis website. Current work with The Cancer Proteome Atlas is driving the creation of new methods for rapid integration, enabling application developers to give their users immediate access to network visualization, exploration, and analysis.

Finally, NDEx is a platform for the dissemination and publication of networks, a distribution channel that does not require the creation and maintenance of custom databases or websites. It is a framework to incorporate networks in publications, giving readers access to live data that can be explored, analyzed in Cytoscape, or saved for future use. NDEx is now an official recommended repository of Scientific Data and Springer Nature. The NDEx project can mint DOIs for networks and can be considered a stable repository with a commitment from UCSD to long-term archival support.

Poster 26: Highly Interactive Next-Generation Clustered Heat Maps (NG-CHMs)

Bradley M. Broom[1*], Michael C. Ryan[2], Chris Wakefield[1], Bob Brown[2], Futa Ikeda[2], Mark Stucky[2], James Melott[1], Rehan Akbani[1], and John N. Weinstein[1]

[1]Dept of Bioinformatics and Computational Biology, UT MD Anderson Cancer Center; [2]Insilico Solutions, [*]bmbroom@mdanderson.org

Clustered heat maps are the most frequently used graphics for visualization of molecular profiling data in biology, appearing in tens of thousands of publications—but as static images. For exploration of large data matrices (up to about 20,000 rows and/or columns), we have now developed highly interactive "Next Generation" Clustered Heat Maps (NG-CHMs). NG-CHMs enable the user to zoom and navigate dynamically and link out to dozens of external data resources and computational tools. NG-CHMs (Fig. 1) exploit recent advances in web technology to improve performance, provide a highly responsive user experience, and facilitate deep exploration of the biology behind the image.

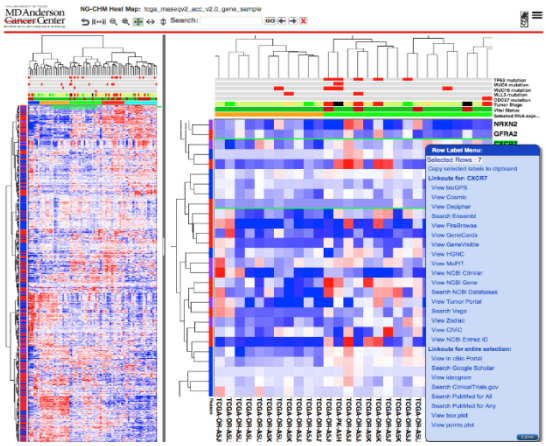


Figure 1 Screenshot of an NG-CHM showing a link-out

NG-CHMs (http://www.ngchm.net) include the following interactive capabilities (among many others):

- Extreme zooming without loss of resolution for drill-down exploration of large data matrices.
- Fluent navigation.
- Flexible real-time recoloring.
- Link-outs to more than a dozen pertinent annotation resources, including GeneCards, PubMed, the Gene Ontology, cBioPortal, MuPIT, TCPA, CIViC, Cancer Digital Slide Archive.
- Annotation with pathway data.
- High-resolution graphics that meet the requirements of all major journals.
- Capture of metadata necessary to reproduce any chosen state of the map months or years later.

NG-CHMs have been widely used throughout The Cancer Genome Atlas (TCGA) projects. Our poster and demo will illustrate the capabilities of NG-CHMs using data from our interactive compendium of 297 NG-CHMs for TCGA data at the protein, RNA, and DNA levels (http://tcga.ngchm.net). However, the system fully supports data from other domains.

Figure 2 Screenshot showing a partial list of NG-CHM plug-ins active on an

For building NG-CHMs, an inter-active web-based system, Galaxy tools, R/R-Studio packages, and Docker images are also available.

Link-outs (Fig. 2) in the NG-CHM system are defined by an extensible system of plug-ins. Computational link-outs can access data in an NG-CHM via an API.

NG-CHMs are currently linked with five other ITCR projects, and links to three other ITCR projects have been proposed (Fig. 3).

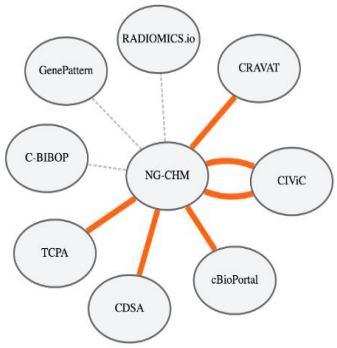


Figure 3 The subset of the ITCR connectivity map

Funding support for NG-CHMs by ITCR: grant no. U24CA199461 from NCI/NIH, TCGA: grant no. U24CA143883 from NCI/NIH, The Michael & Susan Dell Foundation: The Lorraine Dell Program in Bioinformatic, The H.A. Mary K. Chapman Foundation, and an Anonymous donor for Computational Biology in Cancer Medicine.

Poster 27: UCSC Xena - Platform for Functional Genomics Visualization and Analysis

Jing Zhu[1], Brian Craft[1], Mary Goldman[1], Eric Collison[2], Suzanna Lewis[3], and David Haussler[1]

[1*]UCSC Genomics Institute, UC Santa Cruz, jzhu@soe.ucsc.edu
[2]UCSF School of Medicine, UC San Francisco
[3]Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory

The UCSC Xena platform (http://xena.ucsc.edu/) allows biologists and bioinformaticians to securely analyze and visualize cancer functional genomics data. Our unique Visual Spreadsheet shows multiple data types side-by-side, enabling the discovery of correlations across and/or within genes and genomic regions. Dynamic Kaplan-Meier survival analysis, scatter plots, bar graphs, and box plots are all displayed with statistical tests. The Transcript View compares transcript-specific expression for 'tumor' TCGA data to 'normal' GTEX data within and across tissue types. We dynamically link to CRAVAT, TumorMap, and the UCSC Genome Browser, giving users complementary views and data to further deepen their understanding of the underlying biology.

In addition to the commonly available SNPs, INDELs, copy number variation, and gene expression datasets, we support DNA methylation, exon-, transcript-, miRNA-, lncRNA-expression and structural variants. We also support clinical data such as phenotypes, subtype classifications and biomarkers. Our whole genome views allow users to easily visualize non-coding regions for both copy number variation and somatic mutations. Users can dynamically find, filter and subgroup on any of this data. Our expanding public Xena Data Hubs currently host 1500+ datasets from more than 35 cancer types, as well as Pan-Cancer datasets. We showcase seminal cancer genomic datasets to the scientific community, including the latest from the GDC, TCGA, PanCan Atlas, PCAWG, and ICGC. We also host 'normal tissue' datasets from GTEx for comparison with TCGA tumor data. All of our data is available for download via our python API or through AWS S3 buckets.

In addition to exploring these public datasets, the UCSC Xena Browser can easily display an investigator's own genomic and clinical data on a Xena Hub. We empower users to install and load data into their own Xena hub - our architecture ensures that the investigator's data remains private. The lightweight Xena Data Hubs are straightforward to install on Windows, Mac and Linux operating systems. Loading data is easy using either our point-and-click application or command line interface.

We recently integrated TIES NLP search into UCSC Xena, allowing users to search TCGA pathology reports to find their samples of interest. The TCGA pathology reports have clinical data that is not available in the coded clinical datasets, such as whether the patient has entered menopause or the color of their tumor. By giving Xena users access to the data within the pathology reports through TIES NLP technology, we allow users to more finely select samples of interest. After selecting samples, users can filter to just their samples or create and compare two subgroups, allowing iterative analysis and visualization.

Poster 28: The cBioPortal for Cancer Genomics: An intuitive open-source platform for exploration, analysis and visualization of cancer genomics data

The cBioPortal Consortium
Jianjiong Gao, Tali Mazor, Ersin Ciftci, Pichai Raman, Pieter Lukasse, Istemi Bahceci, Alexandros Sigaras, Adam Abeshouse, Ino de Bruijn, Benjamin Gross, Ritika Kundra, Aaron Lisman, Angelica Ochoa, Robert Sheridan, Jing Su, Onur Sumer, Yichao Sun, Avery Wang, Jiaojiao Wang, Manda Wilson, Hongxin Zhang, Priti Kumari, James Lindsay, Karthik Kalletla, Kelsey Zhu, Oleguer Plantalech, Fedde Schaeffer, Sander Tan, Dionne Zaal, Sjoerd van Hagen, Kees van Bochove, Ugur Dogrusoz, Trevor Pugh, Adam Resnick, Chris Sander, Ethan Cerami, and Nikolaus Schultz, cerami@jimmy.harvard.edu, schultz@cbio.mskcc.org

The cBioPortal for Cancer Genomics is an open source software platform that enables interactive, exploratory analysis of large-scale cancer genomics data sets. It integrates genomic and clinical data, and provides a suite of visualization and analysis options, including cohort and patient-level visualization, mutation visualization, survival analysis, alteration enrichment analysis, and network analysis. The user interface is user-friendly, responsive, and makes genomic data easily accessible to scientists and clinicians. The public site (http://www.cbioportal.org) hosts data from more than 215 studies, including data from large consortia (TCGA and ICGC) and individual labs. With newly released functionality, users can now explore and query these studies individually or can combine multiple studies into new "virtual studies". The main features of the portal include OncoPrints, a compact graphical representation of alterations in multiple genes across a cohort, mutational diagrams that show locations and frequencies of mutations in a single gene, Kaplan-Meier survival curves, plots that allow the visualization of correlation between different data types for a single or multiple genes (e.g. the correlation between DNA copy number and mRNA expression), among others. To facilitate interpretation of genomic data, the cBioPortal also now integrates annotations from several leading knowledgebases (OncoKB, CIViC, MyCancerGenome and COSMIC), as well as other resources that can guide variant interpretation (CancerHotspots, MutationAssessor, SIFT and PolyPhen).

The cBioPortal has been widely adopted by the cancer community, with dozens of private instances at academic institutions and pharmaceutical/biotechnology companies. The public portal is currently accessed by approximately 25,000 unique visitors per month. Another notable instance is the cBioPortal for AACR GENIE (http://www.cbioportal.org/genie/), which hosts 39,600 samples from AACR Project GENIE. The cBioPortal is fully open source and all code is available on GitHub (https://github.com/cBioPortal/) under a GNU Affero GPL license. Development is a collaborative effort among groups at Memorial Sloan Kettering Cancer Center, Dana-Farber Cancer Institute, Children's Hospital of Philadelphia, Princess Margaret Cancer Centre, and The Hyve, an open source bioinformatics company based in the Netherlands. Ongoing development efforts are focused on (1) building the open source community; (2) implementing architectural and performance improvements; (3) expanding user support, documentation and training resources; (4) developing novel features to support immunogenomics and immunotherapy; (5) enhancing visualization of patient timelines, multiple tumor profiles, and cohort response; (6) releasing a new public Application Programming Interface (API); and (7) increasing integration with other ITCR Funded platforms.

Poster 29: Placeholder Huang

Poster 30: Multi'omic analysis of subtype heterogeneity in high-grade serous ovarian carcinoma

Ludwig Geistlinger[1], Sehyun Oh[1], Lucas Schiffer[1], Marcel Ramos[1], Michael Birrer[2], Martin Morgan[3], Markus Riester[4], Levi Waldron[1*]

[1*]Department of Epidemiology and Biostatistics, CUNY Graduate School of Public Health and Health Policy, New York NY, levi.waldron@sph.cuny.edu
[2]University of Alabama Comprehensive Cancer Center, Birmingham AB
[3]Roswell Park Cancer Institute, Buffalo, NY
[4]Novartis Institutes for Biomedical Research, Cambridge, MA

High-grade serous ovarian carcinoma (HGSOC) is a molecularly heterogeneous disease in which clinically similar cases can exhibit dramatically different response to treatment. Several studies have identified transcriptome subtypes of HGSOC, yet the interpretation and clinical utility of these subtypes remain controversial. It has been estimated that 90% of HGSOC tumors are polyclonal, and clonal spread of HGSOC has been directly inferred from single-nucleus sequencing. If clonal propagation leads to multiple subtypes, then even unambiguously classifiable tumors might be contaminated by small amounts of another subtype that could lead to relapse after subtype-specific therapy. We therefore test the hypothesis that the previously proposed subtypes tend not to be shared between intratumor clones, implying that they differentiate late in tumorigenesis. This hypothesis is tested in The Cancer Genome Atlas HGSOC cases by (i) considering recurrent subtype-associated DNA copy number alterations, (ii) inferring per-alteration heterogeneity from SNP arrays and whole-exome sequencing, and (iii) testing whether subtype-associated alterations display different intra-tumor heterogeneity than other alterations. Preliminary results suggest that ambiguity in the classification of many tumors may arise from intra-tumoral clones of different subtypes, as opposed to clones of uniformly ambiguous subtype. This talk will describe how ITCR-funded development of the novel Bioconductor data classes RaggedExperiment and MultiAssayExperiment enabled testing this hypothesis.

Poster 31: Reconstruction and interpretation of subclonal tumor evolution from rapid autopsy data reveals novel patterns of aggressive metastatic colonization

Xiaomeng Huang[1], Yi Qiao[1], Thomas Nicholas[1], Aaron Quinlan[1], and Gabor Marth[1]
[1]Department of Human Genetics, University of Utah
gabor.marth@gmail.com

Metastatic breast cancer is an advanced-stage disease in which the cancer cells spread to distant organs. To understand the patterns of metastatic colonization in a patient who presented with aggressive disease, we collected tumor biopsies at initial diagnosis and at mastectomy necessitated by the patient's relapse; as well as twenty-six metastatic tumor sites across seven organs and two normal tissue control skin biopsies via a rapid autopsy procedure within hours after the patient's death. All biopsy samples were subjected to 60X Illumina whole genome sequencing. Our analysis revealed extensive chromosomal changes including amplifications, deletions, LOH and translocations, as well as known driver mutations in RB1, TP53, and PTEN in all biopsy samples. We used the CNV and LOH data to reconstruct the phylogenetic relationships among the tumor samples. Using an extension of our published SubcloneSeeker algorithm, we utilized the somatic SNV allele frequencies in copy number-normal regions of the tumor genomes to refine these phylogenetic relationships, and to construct a detailed map of subclonal expansion that led to metastatic colonization of distal organs.

Subclonal analysis indicates early metastatic escape into the lung, well before the mastectomy procedure could have saved the patient. We identified four distinct waves of metastatic colonization, first into the abdominal organs, then two separate waves into the lymph nodes, and the brain and bones. Detailed subclonal analysis in this aggressive tumor reveals both monoclonal and polyclonal seeding of specific metastatic sites. We also observed, for the first time in a clinical setting, a novel seeding pattern: metastatic "recolonization" of an already established metastatic site in the lung. Our analysis highlights the central role of the lung in the metastatic spread of the tumor, i.e. sites in the lung serving as "subclone incubators" in the colonization of new organs and new sites. Functional annotations of the genomic and transcriptomic alterations indicate high metastatic potential that is already present in the primary tumor, a finding that is consistent with the rampant metastatic spread in this patient. These annotations indicate that, although many of the metastasis-related genomic alterations were present in, and thus potentially targetable at, all tumor sites in the patient, others were only present at a subset of the metastatic sites. To determine which is the case required comprehensive sampling of the metastatic sites. This serves as a cautionary note that targeted treatment strategies based on assaying only the primary site, or only one or two metastatic sites may fail to achieve the desired therapeutic outcome, and more comprehensive surveying of the patient's metastases appears necessary.

The high number of biopsied sites in this study, 30 in all, allowed us to reconstruct the evolution of the aggressive disease in our patient with unprecedented resolution, and to identify characteristic patterns of metastatic colonization. Our ongoing effort to improve the annotation of genomic alterations and transcriptomic changes as part of a collaborative project with the developers of the CiViC data (presented as a poster presentation at the 2018 annual ITCR meeting) was key to interpreting these patterns in the context of the metastatic spread and potential clinical interventions. If confirmed in additional, similarly high-resolution datasets, these patterns will lead to better understanding of the metastatic process and may help guide effective clinical intervention.

Poster 32: A Galaxy-Based Multi-Omic Informatics Hub for Cancer Researchers

Timothy Griffin (tgriffin@umn.edu) and Pratik Jagtap (pjagtap@umn.edu)

Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota

We will update on our ongoing work on developing a unified informatics hub for multi-omics data analysis and interpretation in cancer research. The overarching goal of our work is to implement data analysis and informatics tools for integration of biological mass spectrometry data (proteomics and metabolomics) with genomic / transcriptomic information. We are utilizing the Galaxy platform as the unifying workflow management system to implement and disseminate multi-omic tools and validated workflows. Work includes:

1) Proteogenomics tools, which integrate genomic, transcriptomic and mass spectrometry-based proteomics data. Progress in the last year has focused on validating generic methods for generating customized protein sequence databases from RNA-Seq data, capturing a variety of sequence variants indicated from these transcripts. We are also working on refining methods for matching the mass spectrometry data to these sequences to confirm expression of protein sequence variants. Through a new collaboration with the ITCR-funded Karchin research group, we are also integrating the CRAVAT tool with our proteogenomics outputs, to offer methods to interpret and visualize the potential associations with cancer progression of protein variants.

2) We have continued developing and validating a multi-omics visualization platform (MVP), which acts as a Galaxy-plugin for visualizing proteogenomics results. MVP allows users to generate metrics for peptide-spectral matches and view selected mass spectra matching to variant peptide sequences. MVP also enables mapping of peptide variants to genomic coordinates and its visualization in the IGVjs framework, embedded directly within the MVP tool. This work has leveraged ongoing development work of IGVjs by the ITCR-funded Mesirov research group.

3) We continue to develop quantitative tools for assessing protein functions expressed by microorganisms present within host samples, which may play a role in cancer progression and have value as biomarkers. We have extended basic workflows that integrate metagenomics and metaproteomic data and implemented tools to quantify microbial protein expression and relate this to dynamic responses of taxonomic composition and function. Tools are also in development to visualize these results.

4) We have implemented a workflow for quantitative mass spectrometry-based metabolomics, which includes customized tools for imputing missing values across samples and also deploying new tools for metabolite identification from high-resolution mass spectrometry data.

5) Via a supplement grant with the IMAT-funded group of Laurie Parker, we have implemented tools in Galaxy to support data-independent acquisition (DIA) mass spectrometry data. The workflow takes DIA data as inputs and produces annotated peptide tandem mass spectrometry data. Tools are in process to generate peptide spectral libraries from this data, which can be used for hypothesis-driven queries of the DIA data, including biosensor data from the Parker lab.

5) We have continued in our dissemination activities, to promote our informatics resources to users in the community who could benefit from these workflows. We have conducted hands-on workshops at annual ABRF, ASMS and GCC conferences (http://galaxyp.org/workshops/). For these workshops we utilized a training instance on the cloud-based scientific computing resource Jetstream. We have also established several gateway instances for public access on Jetstream resources, supporting applications in proteogenomics, metaprotemics and metabolomics (http://galaxyp.org/access-galaxy-p/).

Poster 33: Informatics Tools for High-Throughput Analysis of Cancer Mutations

Rachel Karchin[1*] and Michael Ryan[2]

[1*]Departments of Biomedical Engineering and Oncology, Institute for Computational Medicine, Johns Hopkins University, karchin@jhu.edu , [2*]In Silico Solutions, Falls Church, VA

Variant interpretation is a critical issue for labs involved in germline and somatic sequencing, particularly those looking at high-risk patient populations. The Cancer-Related Analysis of Variants Toolkit (CRAVAT) is a collection of tools for interpreting genomic variation, developed at Johns Hopkins and continually improved to reflect input from 1000s of clinical and experimental users. CRAVAT delivers consensus standards and innovative protocols for variant interpretation in a highly intuitive, visually striking, and interactive environment. The tools leverage an extensive repository of pre-computed genome-scale data and custom machine learning algorithms. CRAVAT is fast, results are explored through an interactive dashboard with publication-ready plots, verbose annotation, and 3D protein structural mapping, all of which can be accomplished via web portal, or through a Docker container run in the cloud or locally. We have designed CRAVAT to be easy for other tools to integrate with through multiple supported methods: powerful customizable URL links, sophisticated web services, Galaxy tools, plug-ins, and interfaces. Table 1 shows our current and in-progress integrations.

Table 1. ITCR and external tools integrated with CRAVAT.

Tool	Category	Description	ITCR	Status
Trinity CTAT	Genomics	Tools for leveraging RNA-seq to interpret cancer transcriptomes	X	Implemented
IGV	Genomics	Visualization tool for large, integrated genomic data analysis	X	Implemented
CIVIC	Clinical	Web resource for clinical interpretation of cancer variants	X	Implemented
Galaxy P	Multi-omic	Platform to integrate genomics, transcriptomics, and proteomics	X	Implemented
UCSC Xena	Multi-omic	Cancer data repository and visualization tools	X	Implemented
NDEx	Systems Biology	Network repository and visualization tools	X	Implemented
NG-CHM	Multi-omic	Dynamic, graphical environment for clustered data analysis	X	Implemented
BMEG	Clinical	Cancer data integration platform based on graphical modeling	X	In-progress
PDX Finder	Animal models	Integrated data archive for PDX mouse models of cancer	X	In-progress
GEMINI	Genomics	Data-mining framework to annotate genomic variation	X	In-progress
VariantValidator	Genomics	Accurate validation, mapping and formatting of variants using HGVS	X	In-progress
Gene Pattern	Multi-omic	Platform for Reproducible Bioinformatics	X	In-progress
GLOBUS	Genomics	Cloud based sequence analysis		In-progress

KAVIAR	Genomics	Comprehensive genomics variant database		Implemented
BRCA Exchange	Clinical	Web portal for BRCA variant interpretation in genetic counseling		Implemented
Peptide Atlas	Multi-omic	Large-scale proteomic database with applications to genome annotation		Implemented
GALAXY	Multi-omic	Web-based platform for data intensive biomedical research		In-progress

CRAVAT has been available through a web portal interface for 7 1⁄2 years. In that time, analysis jobs have been submitted to the portal by 18,554 unique users from 123 countries on six continents. The web portal has processed a total 64,975 submissions, ranging from a few mutations to over 64 million, totaling over 2 billion mutations. The CRAVAT Docker container has been available for almost 2 ½ years and has been downloaded over 1200 times.

This year, we plan to introduce a new modular software architecture, which will enable us to expand tool development to a larger community of external developers. It is designed to support lightweight, highly customized local installs, enabling users to install and run selected modules of interest, rather than the full set of annotators, scorers, and visualizations. The base CRAVAT software will be available in the public pip repository and analysis modules (both those developed by the CRAVAT team and those from external developers) will be installed by users from a “CRAVAT store”.

Poster 34: The Cancer Proteome Atlas: A Comprehensive Bioinformatics Resource for Cancer Functional Proteomic Data

Jun Li[1], Yiling Lu[2], Wei Zhao[2], Rehan Akbani[1], Gordon B Mills[2] and Han Liang[1,2*]

[1*]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center
[2]Department of Systems Biology, The University of Texas MD Anderson Cancer Center
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 two separate web applications. The first one focuses on RPPA data of patient tumors, which contains >8,000 samples of 32 cancer types from The Cancer Genome Atlas and other independent patient cohorts. The second application focuses on the RPPA data of cancer cell lines, and contains >650 independent cell lines across 19 lineages. Many of these cell lines have publicly available, high-quality DNA, RNA and drug screening data. We are working on a third application that contains adaptive RPPA response of cell lines to drug treatments. Collectively, TCPA provides various analytic and visualization modules to help cancer researchers explore high-quality RPPA datasets and generate testable hypotheses in an effective and intuitive manner.

Poster 35: The Integrative Genomics Viewer (IGV): visualization supporting cancer research

James T Robinson[1*], Helga Thorvaldsdóttir[2], and Jill P. Mesirov[1,2,3]

[1]School of Medicine, University of California San Diego, La Jolla, CA, [*]jrobinso@ucsd.edu
[2] Broad Institute of MIT and Harvard, Cambridge, MA
[3] Moores Cancer Center, University of California San Diego, La Jolla, CA

Advances in genomic technologies are revolutionizing our knowledge and understanding of cancer and other diseases. They enable the generation and analysis of diverse types of data to detect and link genomic abnormalities with clinical phenotypes. While much of the analysis can be automated, human interpretation and judgment, supported by rapid and intuitive visualization, is essential for gaining insight and elucidating complex biological relationships. This is true in both basic research and the clinic, where manual review of aligned reads for confirmation and interpretation of variant calls is an important step in variant calling.

First released in 2008, the *Integrative Genomics Viewer (IGV)* is a high-performance desktop tool for interactive visual exploration of diverse genomic data. The IGV supports real-time interaction at all scales of genome resolution, from whole genome to base pairs, even for very large NGS datasets. A key characteristic of IGV is its focus on the integrative nature of genomic studies, with support for both array-based and NGS data, and the integration of clinical and phenotypic data. IGV was one of the first tools to provide NGS data visualization, and it currently provides a rich set of tools for inspection, validation, and interpretation of NGS datasets. While IGV is often used to view genomic data from public sources, its primary emphasis is to support researchers who wish to visualize and explore their own datasets, or those from colleagues. To that end, IGV supports flexible loading of local and remote datasets, and is optimized to provide high-performance data visualization and exploration on standard personal computers.

In 2016, we introduced *igv.js*, a JavaScript implementation of IGV, designed to be easily embeddable in web portals, cloud applications, and investigator's personal web pages. It supports many of the key features of the desktop application including visualization of NGS alignments, germ-line and somatic variants, copy-number data, annotations, and EQTL data. This version has been widely adopted in web portals, including the cBioPortal for Cancer Genomics, Genotype-Tissue Expression (GTEx) portal, FireCloud, and the ISB Cancer Genomics Cloud. Currently we are focused on enhancing support for individual investigators to visualize and share data from common cloud platforms such as Dropbox, Google Drive, and Amazon S3.

More recently, we developed *Juicebox.js*, a cloud-based web application for exploring genomic datasets in the context of contact mapping experiments, such as Hi-C and ChIA-PET, that explore how genomes fold in 3D. Juicebox.js is based on igv.js and investigators can view tracks in both applications from the same data sources.

The IGV desktop application, igv.js, and Juicebox.js are all open source and freely available at www.igv.org and github.com/igvteam.

Poster 36: GenePattern Notebooks for Cancer Research

Michael Reich[1], Thorin Tabor[1], Peter Carr[2], Edwin Juarez[1], David Eby[1], Ted Liefeld[1], Helga Thorvaldsdóttir[2], Barbara Hill[2], Pablo Tamayo[1], Jill P Mesirov[1, 2]

[1]School of Medicine, UCSD, San Diego, CA 92093
[2]Broad Institute, Cambridge, MA 02142, mmreich@cloud.ucsd.edu

As the availability of genetic and genomic data and analysis tools from large-scale cancer initiatives continues to increase, the need has become more urgent for a software environment that supports the entire "idea to dissemination" cycle of an integrative cancer genomics analysis. Such a system would need to provide access to a large number of analysis tools without the need for programming, be sufficiently flexible to accommodate the practices of non-programming biologists as well as experienced bioinformaticians, and would provide a way for researchers to encapsulate their work into a single "executable document" including not only the analytical workflow but also the associated descriptive text, graphics, and supporting research. Crucially, this system would provide easy access to genomic analysis tools without the need for programming. To address these needs, we have developed GenePattern Notebook, based on the GenePattern environment for integrative genomics and the Jupyter Notebook system. GenePattern Notebook unites the phases of *in silico* research – experiment design, analysis, and publication – into a single accessible interface.

GenePattern Notebook presents a familiar lab notebook format that allows researchers to build a record of their work by creating "cells" containing text, graphics, and executable analyses. Researchers add, delete, and modify cells as the research evolves, supporting the initial research phases of prototyping and collaborative analysis. When an analysis is ready for publication, the same document that was used in the design and analysis phases becomes a research narrative that interleaves text, graphics, data, and executable analyses. The online notebook format allows researchers to explain the analytical and scientific considerations of each step in any level of detail, promoting reproducibility and adoption. Notebooks can also be shared between researchers for collaborative development.

GenePattern Notebook features are designed to help nonprogramming users create and adapt notebooks. We have developed enhancements allowing users to choose analyses from any public GenePattern server, seamlessly send data between GenePattern and Python cells, and create richly formatted text, all without the need for programming. For bioinformatics developers, we have created a "UI Builder" that wraps any Python function in an accessible web-style UI with the addition of one line of code. This facilitates the rapid prototyping of new methods and sharing of code to non-programming researchers.

The platform also provides interactive visualizers for many genomic data types, which can be included in notebooks: heat maps, dendrograms, cluster results, and igv.js, the JavaScript implementation of the Integrative Genomics Viewer.

A free online GenePattern Notebook workspace is available at http://www.genepattern-notebook.org, where researchers can develop, share, and publish notebook documents. We have provided a collection of template notebooks that walk users through various genomic and machine learning analyses, and are collaborating with cancer research laboratories to create integrative cancer genomics notebooks.

Poster 37: The TIES Cancer Research Network (TCRN): Computational Pathology Support for Precision Oncology

Michael J. Becich[1], Joel Saltz[2], Jonathan Silverstein[1], Roni J. Bollag[3], Mathias Brochhausen[4], Chakra Chennubhotla[5], Michael D. Feldman[6], Carmelo Gaudioso[7], Tahsin Kurc[2], Jack London[8], Nita J. Maihle[3], Fred Prior[4], Ashish Sharma[9] and Lawrence Tarbox[4]

[1]Department of Biomedical Informatics, University of Pittsburgh School of Medicine
[2]Department of Biomedical Informatics, Stony Brook University
[3]Augusta University, Depts. of Pathology (RB) and Biochemistry and Molecular Biology (NM)
[4]Department of Biomedical Informatics, University of Arkansas for Medical Sciences
[5]Department of Computational and Systems Biology, Univ. of Pittsburgh School of Medicine
[6]Department of Pathology and Laboratory Medicine, Univ. of Pennsylvania School of Medicine
[7]Roswell Park Cancer Institute, Department of Biostatistics and Bioinformatics
[8]Thomas Jefferson University, Kimmel Cancer Center
[9]Emory University School of Medicine, Department of Biomedical Informatics
Presenting Author E-mail = becich@pitt.edu

Advances in cancer research and personalized medicine require significant new bridging infrastructures, including more robust biorepositories that link human tissue to clinical phenotypes and outcomes data. The Text Information Extraction System (TIES) (http://ties.dbmi.pitt.edu) Cancer Research Network (TCRN) (https://cancerdatanetwork.org) is a novel platform developed at the Department of Biomedical Informatics, University of Pittsburgh that enables cancer researchers to mine the content of pathology reports and to share biospecimens across a federated network. Member sites can access pathology data that are de-identified and processed with the TIES natural language processing system (NLP), which creates a repository of rich phenotype data linked to clinical biospecimens. TIES incorporates multiple security and privacy best practices that, combined with legal agreements, network policies, and procedures, enable regulatory compliance. The institutional trust fabric TCRN has developed has been "road tested" by our network and has been used by several funded translational clinical research programs. We have linked cancer registry data with pathology reports and are expanding this feature to all network sites to enable more precise searches and facilitate epidemiological research that will use data from electronic records, including outcomes data. We have piloted the addition of whole slide images (WSIs) linked to pathology reports, containing computationally derived image-based features derived from two innovative Informatics Technology for Cancer Research programs. TCRN will continue to support the "in network" synergistic use of biospecimens and clinical data derived from the six current and two new network expansion sites. All TCRN sites have agreed to share de-identified WSIs and the computational pathology phenotypes from images, NLP processed text, as well as discrete data, publicly with the cancer research community via The Cancer Imaging Archive (TCIA). This work will significantly enhance the computational pathology WSI research resources in TCIA and enable image-based discovery for precision oncology, particularly in the areas of immuno-oncology, tumor heterogeneity and tumor microenvironment. The TIES Cancer Research Network presents a model for data, WSI, and biospecimen sharing locally and nationally.

Poster 38: CIViC: Crowdsourced and community-driven standards and interfaces for curation and submission of somatic cancer variants

Obi Griffith[1], Kilannin Krysiak[1], Arpad Danos[1], Erica Barnell[1], Joshua McMichael[1], Susanna Kiwala[1], Adam Coffman[1], Nicholas Spies[1], Lynzey Kujan[1], Kaitlin Clark[1], Yang-Yang Feng[1], Zachary Skidmore[1], Cody Ramirez[1], Alex Wagner[1], Elaine Mardis[2] and Malachi Griffith[1,*]

[1]McDonnell Genome Institute, Washington University School of Medicine

[2]The Institute for Genomic Medicine, Nationwide Children's Hospital

[*]correspondence: mgriffit@wustl.edu and obigriffith@wustl.edu

The Clinical Interpretations of Variants in Cancer (CIViC; civicdb.org) resource is a crowd-sourced, open-access, and structured knowledgebase that evaluates genomic variants for therapeutic, prognostic, predisposing, or diagnostic utility. This resource helps to alleviate the significant interpretation bottleneck of personalized medicine through contribution from >115 international users on 346 genes, 1,844 variants, and 412 drugs across 226 cancer types. Here we describe the latest features, curation milestones, and changes made to the CIViC interface to incorporate clinical guidelines, increase interoperability, and enhance downstream applications.

We have incorporated new feedback protocols such as entity flagging to instantly notify curators of problematic entities with minimal user effort. We also implemented assertion entries, which aggregate individual evidence items into a single clinical recommendation that is structured, human-readable, and programmatically accessible. Assertion entries are flexible with regards to guideline incorporation and allow for generic labels (e.g. ALK fusions), which are typically unsupported by many resources. Other implemented features include a CIViC Actionability Score for each variant.

CIViC has made a concerted effort to stay abreast of the standards, regulations, and resources available for the use of genomic data in clinical applications. Collaborations with external organizations, such as ClinGen, were valuable in incorporating recognized guidelines, standards, and best practices into variant annotation. These include ClinGen Allele Registry identifiers, gnomAD allele frequencies, ACMG/AMP variant interpretation tiering, FDA companion tests, FDA drug approvals, HPO terms, and WHO/NCCN guidelines. Our collaboration with the ClinGen Somatic Working Group has been formally described in a manuscript recently submitted to Human Mutation. CIViC has also joined with members of the Global Alliance for Genomics and Health (GA4GH) to create a driver project in this space, the Variant Interpretation for Cancer Consortium (VICC, cancervariants.org). This consortium is aggregating the collective knowledge of six established sources of cancer variant interpretations: CIViC, OncoKB, the Jackson Laboratories Clinical Knowledgebase, Precision Medicine Knowledgebase, MolecularMatch, and the Cancer Genome Interpreter. The VICC has identified a set of minimum elements required for describing a cancer variant interpretation to guide harvesting and harmonizing knowledge from the founding resources and is developing an interface to perform a harmonized query (http://g2p-test.ddns.net).

Future work includes: creation of the CIViC DesignStudio to allow individuals to download custom capture probe designs, support for complex genotypes and evidence sources beyond PubMed, automated ClinVar submissions, and an extension of our CIViCpy Python client and analysis toolkit (civicpy.org).

Poster 39: A network-based approach for personalized treatment of Multiple Myeloma

Alessandro Laganà
Email: alessandro.lagana@mssm.edu

Department of Genetics and Genomic Sciences
Institute for Next-Generation Healthcare
Icahn School of Medicine at Mount Sinai, New York, NY.

Multiple Myeloma (MM) is an incurable malignancy of bone marrow plasma cells with a median survival of approximately 6 years. Despite recent therapeutic advancements, the disease remains fatal in the majority of patients. Transformation events at key stages of disease progression are still unelucidated and the causal drivers of MM pathogenesis are still unclear. Clonal heterogeneity, evolution and competition suggest that the employment of a personalized therapeutic approach is likely to improve outcomes for myeloma. In recent years, integrative network-based methods have proven extremely effective in modeling complex biological systems and uncovering novel patterns of genomic perturbation associated with disease initiation and progression.

We have applied an integrative network biology approach to molecular and clinical data measured from 450 patients with newly diagnosed MM enrolled in the MMRF CoMMpass study, and generated MMNet, a novel network model of myeloma. Our analysis identified new insights into MM disease biology and provided improved molecular features for diagnosing and stratifying patients, as well as additional molecular targets for therapeutic alternatives. More specifically, we (i) identified novel candidate genetic drivers associated to alteration of gene expression; (ii) demonstrated for the first time that high clonality and mutation burden are associated with poor prognosis in newly diagnosed myeloma patients; (iii) identified four novel genes associated to high-risk myeloma at diagnosis; (iv) identified and validated the genes CLEC11A and CDC42BPA as upstream regulators of the aggressive MMSET-associated myeloma subtype; (v) identified three novel patient classes characterized by recurrent transcriptional and genetic events, MYC translocation, and immune activation; (vi) demonstrated the ability of deep sequencing techniques to detect relevant structural rearrangements, providing evidence which encourages wider use of such technologies in clinical practice.

In ongoing work, we are now leveraging MMNet to develop a novel clinical decision support tool for precision medicine of MM. This tool will combine network modeling and computational tools to establish a sequencing-based comprehensive patient profile for disease subtype classification, identification of patient-specific driver events, and to predict the most suitable treatment options. Towards this goal, we have implemented a prototype precision medicine computational pipeline integrating both DNA and RNA sequencing findings and we evaluated the feasibility and effectiveness of our approach in a single-institution clinical trial with 64 subjects. Our results showed clinical response in the majority of patients treated based on our recommendations, demonstrating feasibility of our approach and providing a basis for further expansion by employing network modeling.

Poster 40: AMARETTO: Regulatory Network Inference for Driver and Drug Discovery in Cancer

Rileen Sinha[1,2], Thomas Baumert[3,2], Olivier Gevaert[4,2], Nathalie Pochet[1,2*]

[1]Brigham and Women's Hospital, Harvard Medical School [2]Broad Institute of MIT and Harvard [3]INSERM, University of Strasbourg [4]Stanford University [*]npochet@broadinstitute.org

The availability of increasing volumes of multi-omics profiles from model systems to patient studies across many cancers promises to improve our understanding of the regulatory mechanisms underlying cancer. The main challenges are to integrate these multiple levels of omics data and to translate them across *in vitro* and *in vivo* systems. Here we present our AMARETTO framework that allows learning regulatory networks across biological systems with a broad range of applications, from diagnostic subtyping to driver and drug discovery in cancer. First, AMARETTO infers regulatory networks within each biological system via multi-omics data fusion. Specifically, AMARETTO identifies potential cancer drivers by identifying genes whose genetic and epigenetic cancer aberrations have a direct functional impact on their own transcriptomic or proteomic expression. AMARETTO then connects these drivers with modules of co-expressed target genes that they putatively control, defined as regulatory modules, using a penalized regulatory program. Second, AMARETTO learns communities or subnetworks by connecting the regulatory networks and modules inferred from different systems to identify drivers across diseases or biological systems. Downstream analytic functionalities of AMARETTO include functional annotation of modules, stratifying modules for increasingly specific clinical phenotypes, and ongoing work on automated driver and drug discovery using genetic and chemical perturbations in model systems. AMARETTO offers tools for systematic assessment and benchmarking of the inferred regulatory networks for optimal generalization performance of the models. We recently released the source code of AMARETTO and ongoing efforts will establish links of AMARETTO to tools of other ITCR investigators.

We demonstrate the utility of our AMARETTO framework via two case studies. First, we demonstrate its general applicability across cancers via a pan-cancer study where we identified drivers of smoking-induced cancer and 'antiviral' interferon-modulated innate immune response across twelve cancer (sub)types. Second, we refine our framework for a pan-etiology study of hepatocellular carcinoma where we identified drivers and drugs for hepatitis C and B virus-induced hepatocellular carcinoma. Our analyses confirmed that AMARETTO captures hallmarks of cancer. AMARETTO additionally facilitates identification of known and novel cancer drivers and their targets, and how they can be modulated by known and novel drug compounds. In our pan-cancer study we leveraged genetic perturbation studies in cell lines to validate the inferred regulatory modules and predicted candidate drivers across many cancer sites. In our pan-etiology study of hepatocellular carcinoma we additionally translated chemical perturbation studies from cell lines to patient studies to predict which drug treatments potentially reverse liver disease-associated regulatory modules across different biological systems. Two novel predicted compounds were experimentally validated in a rat model by demonstrating their ability to attenuate liver cirrhosis and hepatocellular carcinoma development. In contrast with existing treatments whose adverse effects limit their clinical use as preventive medicine for hepatitis C and B virus-infected patients at risk of progressing to cancer, these novel compounds potentially offer a safe and low-cost approach for chemoprevention of hepatocellular carcinoma.

Poster 41: Algorithms for Literature-Guided Multi-Platform Identification of Cancer Subtypes

Dongjun Chung[1*] and Linda Kelemen[1]

[1*]Department of Public Health Sciences, Medical University of South Carolina,
chungd@musc.edu

The development of accurate and robust statistical methods for cancer subtype identification using high throughput genomic data is of critical importance to the basic cancer research. These methods will inform molecular-based tumor classification and shared pathogeneses, which offers opportunities for overlapping treatments across various cancer subtypes. In spite of tremendous efforts to develop statistical methods for the analysis of high throughput genomic data profiled in multiple platforms for cancer subtype identification, it still remains a challenging task to implement robust and interpretable identification of cancer subtypes and driver molecular features using these massive, complex, and heterogeneous datasets. The impact derived from improved understanding of tumor classification and driver molecular features can be dramatic, as this knowledge can be used to develop more effective prevention and intervention strategies to reduce the burdens of patients suffered from cancers.

In order to address these challenges and opportunities, we have developed statistical methods and software to improve identification of cancer patient subgroups and driver molecular features by integrating genomic data profiled in multiple platforms with biomedical literature and existing pathway databases. First, we developed InGRiD (Integrative Genomics Robust iDentification of cancer subgroups), a statistical framework and an R package to improve the identification of cancer subtypes and key molecular features by integrating cancer genomic data with pathway knowledge. We applied InGRiD to the genomic data of high grade ovarian cancer patients from The Cancer Genome Atlas (TCGA) and the Australian Ovarian Cancer Study (AOCS) and this analysis showed that the utilization of pathway information can significantly improve robustness and interpretability in identification of cancer subtypes and driver molecular features.

On the other hand, we are also developing statistical models and software to improve the pathway knowledge by integrating PubMed literature mining results with multiple pathway annotations. In this framework, biomedical literature supplements the incompleteness of pathway annotations in existing databases. It will also provide a common knowledgebase to integrate information from diverse pathway databases because biomedical literature provides comprehensive information about the relationship among genes. In this direction, we developed bayesGO, a statistical framework and an R package to infer pathway-modulating genes from the PubMed literature mining results and to facilitate interpretation of their functions using Gene Ontology information. In addition, in order to facilitate users' convenience to investigate gene-gene relationships based on the PubMed literature mining results and utilize this information for downstream analyses such as the bayesGO analysis, we developed GAIL (Gene-Gene Association Inference based on biomedical Literature), a web interface and database that allows the investigation of human gene-gene networks based on the PubMed literature mining results.

These tools are publicly available from the GitHub pages for InGRiD (https://dongjunchung.github.io/INGRID/) and bayesGO (https://dongjunchung.github.io/bayesGO/) and the GAIL web site (http://chunglab.io/GAIL/).

Poster 42: Semantically rich interfaces for cloud-scale cancer genomics with Bioconductor

Vincent Carey[1*], Shweta Gopaulakrishnan[1], Samuela Pollack[2], Aedin Culhane[2]

[1*]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School
[2]Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, [*]stvjc@channing.harvard.edu

Two broad questions arise in contemplating cloud computing for cancer genomics: (1) Do I really need cloud computing to do task T effectively? (2) Do I need skills S and/or resources R to do task T effectively in cloud, and if I do, how do I get them? The answer to question (1) is often negative, as local resources are often sufficient and their use is familiar for many interesting tasks. I will focus on question (2), addressing tasks that become easier, more reproducible and shareable owing to currently ready-to-use cloud computing frameworks. For all cloud-scale data architectures and analysis modalities, accurate and scalable semantic annotation and data validation are essential. We propose that the SummarizedExperiment data plus metadata schema, and its implementation in Bioconductor, constitute foundational elements for organizing and interacting with cloud-scale multiomic assay archives.

An early illustration of our approach is the ivyGlimpse app, a cloud-resident interactive interface to the expression, image annotation, and clinical data from the IvyGAP project. Underlying the app is a Bioconductor package, ivygapSE, that collects the data in a fully annotated and versioned SummarizedExperiment. The app allows users to explore interactions among tumor image annotation, image feature quantification, expression patterns in selected cBioPortal pathways, and survival. More work needs to be done to expand the interactive interface to expression and image data, and to achieve statistical rigor in the face of the complex sequential design giving rise to the full ensemble of expression data.

R users have interactive access to all multiomic data in TCGA and the PanCancer Atlas as curated in the ISB Cancer Genomics Cloud using functions and structures provided in the restfulSE and BiocOncoTK packages. Given a BigQuery authentication token, the pancan_app function provides an interactive interface to acquire an overview of PanCancer Atlas tables and fields. pancan_SE creates fully annotated "virtual" SummarizedExperiment instances that lazily and scalably retrieve quantitative data on any omics assay on any sample, as demanded on the basis of sample and feature selection idioms familiar in general Bioconductor interactive and workflow use.

Finally, in conjunction with the HDF Group, we have created a globally accessible HDF Server instance in AWS EC2 that provides a number of data sets from prominent single-cell experiments in GBM, along with the benchmark 1.3 million neuron 10x data. I will discuss how to interact with this, and with the HDF Object Store, to achieve scalable analysis of cloud-resident multiomics data for cancer research.

Poster 43: OncoMX: an integrated cancer mutation and expression resource for exploring cancer biomarkers

Raja Mazumder[1], Daniel Crichton[2], K. Vijay-Shanker[3], Frederic Bastian[4], Amanda Bell[1], Hayley Dingerdissen[1], Samir Gupta[3], Robel Kahsay[1], Heather Kincaid[2], David Liu[2], ASM Ashique Mahmood[3], Marc Robinson-Rechavi[4,5]

[1]Department of Biochemistry and Molecular Medicine, the George Washington University, mazumder@gwu.edu
[2]Jet Propulsion Laboratory, NASA
[3]University of Delaware
[4]Swiss Institute of Bioinformatics
[5]University of Lausanne

Cancer genomics studies generate a huge volume of highly heterogeneous datasets with numerous variable characteristics, including but not limited to file formats, attribute names, positional coordinates, reference data, processing pipelines, and more. These issues of data variability stemming from the primary study design are further complicated by duplicate, non-uniform presentation of the data across several distributed databases. Many such databases have highly specific research foci, and therefore the presentation of data and the specific subset of data available are frequently customized to unique database needs. One application of cancer data is the evaluation of biomarkers, but existing cancer biomarker databases are limited in their abilities to leverage and search across outside resources for information related to potential biomarkers. Thus, there exists a need for database efforts that can bring together diverse cancer genomics and biomarker data in a unified way. OncoMX, an integrated cancer mutation and expression resource for exploring cancer biomarkers via web portal, is actively being developed to address issues of biomedical data integration to better facilitate cancer biomarker research. OncoMX is a multi-faceted collaboration between the George Washington University, NASA's Jet Propulsion Laboratory, the Swiss Institute of Bioinformatics, and the University of Delaware, soon to include scRNA-seq expression data from Boston University. Briefly, sequencing-based mutation and expression data are integrated and unified by Disease Ontology and Uberon terms into BioMuta and BioXpress, the core knowledgebases for OncoMX. Additional integrated data include normal expression from Bgee, text-mining results for mutation and expression in cancer, functional annotations from both curated and non-curated resources, EDRN biomarker information, and Reactome pathways. Usability of OncoMX is currently being designed with respect to four major user perspectives: (1) exploration of cancer biomarkers; (2) evaluation of mutation and expression in an evolutionary context (both cancer and normal); (3) side-by-side exploration of published literature for gene mutation and expression in cancer; and (4) exploration of a specific gene or biomarker within a pathway context. The proposed access to integrated cancer genomics data and supporting information in OncoMX is expected to benefit basic cancer research and promote efficient consumption of information by end users to ultimately improve biomarker detection capabilities.

Poster 44: Improved Structural Prediction of peptide-HLA Complexes Using DINC-Vina

Dinler A. Antunes[1]*, Didier Devaurs[1], Eleni Litsa[1], Kyle R. Jackson[2], Mark Moll[1], Gregory Lizée[2], Lydia E. Kavraki[1].

[1] Department of Computer Science, Rice University, dinler@rice.edu
[2] Department of Melanoma Medical Oncology - Research, The University of Texas M.D. Anderson Cancer Center

Immunotherapy makes use of the patient's own immune system to identify and eliminate cancer cells, representing a promising new avenue for cancer treatment. The class I human leukocyte antigen (HLA) receptors play a key role in this context, binding and displaying at the cell surface peptides derived from intracellular proteins. These peptide-HLA complexes can be recognized by circulating T-cell lymphocytes, triggering T-cell activation and tumor elimination. Considering the complexity of T-cell activation and regulation, structural analyses of peptide-HLA complexes are becoming essential for the development of new T-cell-based immunotherapies. In this context, we previously proposed an incremental meta-docking approach called DINC (Docking INCrementally), and provided preliminary data supporting its use for general prediction of binding modes of peptide-HLA complexes. DINC originally relied on the popular docking tool Autodock 4 to perform sampling and scoring of potential binding modes. However, another popular docking tool called Autodock Vina has been shown to outperform Autodock 4 in both sampling and scoring, as well as in computational speed. Therefore, we decided to develop a new version of our tool, DINC, where Autodock Vina would replace Autodock 4. Here, we describe this updated version of our approach, which we call DINC-Vina, and show that it improves binding mode prediction of peptide-HLA complexes. For that, we report on a re-docking experiment aiming to reproduce a diverse dataset of peptide-HLA complexes available in the Protein Data Bank. This dataset includes different HLA variants, bound to peptides with different lengths and alternative binding modes. The prediction error in re-docking experiments is usually assessed by calculating the Root Mean Square Deviation between the coordinates of all atoms of the ligand (all-atom RMSD) and those of the reference crystal structure. Our method is able to reproduce all complexes with less than 2 Å all-atom RMSD (with an average of 1.33 ± 0.3 Å across the whole dataset). Our new method also shows improved reproducibility, with high consistency of low RMSD conformations across 20 replicates of the same experiment. On the other hand, consistency in terms of the top scoring conformation has proven to be a much bigger challenge. Our analysis of the conformations generated by DINC-Vina show a good exploration of the conformational space and some degree of success of the scoring function in terms of properly ranking these conformations. However, there is still great variation between individual runs of DINC-Vina, requiring more sampling to increase the chances of correctly predicting the best binding mode. Autodock Vina uses a very general scoring function, mostly suited to ranking small drug-like ligands. The evaluation of alternative scoring methods, focused on peptide-HLA prediction, will be part of our future work. Additionally, we have improved our webserver to make it more robust and reliable, which is very important in light of the increasing user demand and of the diversity of docking jobs that users submit. Future development of DINC-Vina will allow fast and accurate geometry prediction of patient-specific peptide-HLA complexes. In turn, these predictions will have a positive impact on personalized T-cell-based immunotherapies against cancer.

Poster 45: CGDnet: Using patient-specific drug-gene networks for recommending targeted cancer therapies

Simina M Boca[1*], Jayaram Kancherla[2], Shruti Rao[1], Subha Madhavan[1], Robert Beckman[1], Héctor Corrada Bravo[2*]

[1]Innovation Center for Biomedical Informatics, Georgetown University Medical Center
[2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park
*smb310@georgetown.edu

Tumor molecular profiling refers to the use of a panel of genes and proteins which are assessed for potential abnormalities in an individual cancer, including somatic point mutations, gene fusions, copy number alterations, and over or under expression. The findings from molecular profiling can then be matched with targeted therapies that are predicted to work for an individual patient, providing an implementation of precision oncology. In practice, molecular profiling reports provide a list of therapies predicted to lead to benefit or lack of benefit, but do not usually consider the potential cross-talk within and between dysregulated biological pathways. We present current work on CDGnet (Cancer-Drug-Gene network), which has as its goal the prioritization of targeted therapies. Our goal is to have an approach and a tool that is automated, personalized to individual patients, as well as evidence-based. Our current prototype is available at https://siminaboca.shinyapps.io/Search_MP_results_using_FDA_approvals_targets_KEGG/.

The inputs are the specific alterations found in a patient's tumor, such as "KRAS | mutation | G13V" and the tumor type, such as "Colorectal cancer." These data are then integrated with: biological networks relevant to the cancer type and to the specific alterations, FDA-approved targeted cancer therapies and indications, and additional gene-drug connections. Currently the biological networks we consider are the cancer-specific pathways in KEGG – we are working on expanding our tool to use associations from Reactome as well.

We currently consider 4 different therapy categories that can be prioritized for a patient, given their specific tumor alterations: 1. FDA-approved drugs for which the alterations are biomarkers for their tumor type; 2. FDA-approved drugs for which the alterations are biomarkers in other tumor types; 3. drugs for which these alterations or other genes/proteins in the pathway corresponding to the patient's tumor type are targets/biomarkers; and 4. drugs for which these alterations or other genes/proteins in additional cancer pathways are targets/biomarkers. Examples of where the pathway-based approach can be useful include the signalling of KRAS on downstream BRAF in colorectal cancer, which can be coupled to the approval of BRAF inhibitors in melanoma and non-small cell lung cancer, potentially justifying its use in colorectal cancer patients as well. If a colorectal cancer patient with a KRAS mutation has already progressed on BRAF inhibitors, a possible choice could include MTOR inhibitors, as MTOR is downstream of KRAS.

We are currently working on expanding our initial prototype with a visualization component that provides intuition to the different evidence levels inherent in the 4 categories, as well as integrating pathways from Reactome and providing normalization to gene names, disease names, and drug names.

Poster 46: Interactive single cell RNA-Seq analysis with the Single Cell Toolkit (SCTK)

David Jenkins[1,2], Tyler Faits[1,2], Emma Briars[2], Sebasitan Carrasco Pro[2], Steve Cunningham[2], Masanao Yajima[3], and W. Evan Johnson[1,2,*]

[1]Division of Computational Biomedicine, Boston University
[2]Bioinformatics Program, Boston University
[3]Department of Statistics, Boston University
[*]Presenting author, wej@bu.edu

Single cell RNA-sequencing (scRNA-Seq) allows researchers to profile transcriptional activity in individual cells, in contrast to bulk RNA-sequencing which profiles a conglomerate of an entire cell population. Because each sample comes from an individual cell, the small amount of input RNA results in sparse data, introducing new analytical challenges not present in bulk RNA-Seq analysis techniques. Additionally, scRNA-Seq datasets can vary dramatically in the number of cells sequenced, the sequencing depth, and the number of batches in which the samples were sequenced, so many dataset dependent analytical decisions must be made. Here, we present the Single Cell Toolkit (SCTK), an interactive scRNA-Seq analysis package that allows a user to upload raw scRNA-Seq count matrices and perform downstream scRNA-Seq analysis interactively through a web interface. The package is written in R with a graphical user interface (GUI) written in Shiny. Users can perform analysis with modules for filtering raw results, clustering, batch correction, differential expression, pathway enrichment, and scRNA-Seq study design, all in a simple to use point and click interface. The toolkit also supports command line data processing, and results can be loaded into the GUI for additional exploration and downstream analysis. We demonstrate the effectiveness of the SCTK on multiple scRNA-seq examples, including data from mucosal-associated invariant T cells, induced pluripotent stem cells, and breast cancer tumor cells. While other scRNA-Seq analysis tools exist, the SCTK is the first fully interactive analysis toolkit for scRNA-Seq data available within the R language.

Poster 47: Genome-wide somatic variant calling using localized colored DeBruijn graphs

Giuseppe Narzisi[1], André Corvelo[1], Kanika Arora[1], Ewa A. Bergmann[1,*], Minita Shah[1], Rajeeva Musunuri[1], Anne-Katrin Emde[1], Nicolas Robine[1], Vladimir Vacic[1,†], Michael C. Zody[1]

[1] New York Genome Center, New York, NY, 10013, USA.
[*] Present address: Illumina Cambridge Ltd, Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex, CB10 1XL, UK.
[†] Present address: 23andMe, Inc., Mountain View, CA, 94041, USA.
gnarzisi@nygenome.org

Reliable detection of somatic variants from next-generation sequencing data requires the ability to effectively handle a broad range of diverse conditions such as aneuploidy, clonality, and purity of the input tumor material. The sensitivity and specificity of any somatic mutation calling approach varies along the genome due to differences in sequencing read depths, error rates, mutation types (e.g., single nucleotide variants – SNVs, insertions and deletions – indels, copy number variants – CNVs) and their sizes. Micro-assembly approaches have been successful at calling indels up to a few hundred base pairs in length, allowing inquiry into the twilight zone between longer indels and shorter CNVs. However, existing micro-assembly methods rely on separate assembly of tumor and matched normal data, which has limitations in regions with low supporting coverage, repeats, and large indels. Accounting for these variables requires flexible methods that can adapt to the specific context of each genomic region.

We present Lancet, a new accurate and sensitive somatic variant caller, which detects SNVs and indels by jointly analyzing reads from tumor and matched normal samples using colored DeBruijn graphs. We demonstrated, through extensive experimental comparison on synthetic and real whole-genome sequencing datasets, that Lancet has better accuracy, especially for indel detection, than widely used somatic callers, such as MuTect, MuTect2, LoFreq, Strelka, and Strelka2. Lancet features a reliable variant scoring system, which is essential for variant prioritization, and detects low frequency mutations without sacrificing the sensitivity to call longer insertions and deletions empowered by the local assembly engine.

In addition to genome-wide analysis, selected variants of interest can be exported and rendered in graph space using Lancet. Visual inspection of the colored de Bruijn graph containing the mutations can help to confirm a variant and augment the traditional read alignment visualization available in tools like IGV. Lancet is freely available for academic and non-commercial research purposes as an open-source software project at https://github.com/nygenome/lancet.

Poster 48: TRUST: an ultrasensitive software for detecting TCR and BCR hypervariable-region sequences from bulk RNA-seq data

Xihao Sherlock Hu[1#], Jian Zhang[2], Taiwen Li[3], Shengqing Stan Gu[1], Jin Wang[4], Jingxin Fu[4], Jun Liu[5], Bo Li[6*], and Xiaole Shirley Liu (xsliu@jimmy.harvard.edu)[1,5*]

[1] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute
[2] Beijing Institute of Basic Medical Sciences, Beijing
[3] State Key Laboratory of Oral Diseases, West China Hospital of Stomatology, Sichuan University
[4] School of Life Science and Technology, Tongji University
[5] Statistics Department, Harvard University
[6] Department of Bioinformatics and Immunology, University of Texas Southwestern
[#]Presenting author: Sherlock Hu (xhu@jimmy.harvard.edu)

Tumor-infiltrating immune cells are important components in the tumor microenvironment with anti-tumor impacts. We previously developed the TRUST method to study the receptor region of tumor-infiltrating T cells. TRUST takes single-end or paired-end library reads mapped to the human reference genome in BAM format as the standard input. It automatically detects input library type, selects informative unmapped reads, and *de novo* assembles the complementarity-determining region 3 (CDR3) sequences. We obtained ~3 million TCR CDR3 sequences from >10k RNA-seq samples of The Cancer Genome Atlas (TCGA). We observed a strong association between T cell diversity and tumor mutation load, and we predicted SPAG5 and TSSK6 as putative immunogenic cancer/testis antigens in multiple cancers.

We further improved TRUST to extract the B cell immunoglobulin (Ig) hypervariable regions from the TCGA RNA-seq samples and assembled over 30 million CDR3s of the B cell heavy chain sequences. TRUST then annotated each CDR3 assembly to possible IGHV, IGHJ, and IGH constant genes. We observed widespread B cell clonal expansions and Ig subclass switch in diverse human cancers. The association study of B cell activity and somatic copy number alterations would reveal the complex role of B cells in tumor. The IgH sequences identified from TCGA samples are potentially useful resources for future development of B cell-related immunotherapies.

Taken together, we presented TRUST as a powerful tool for pan-cancer analysis on tumor-infiltrating T cell and B cell repertoires. The TRUST software is open source and publicly available at: https://bitbucket.org/liulab/trust

Poster 49: Detection of somatic, subclonal and mosaic CNVs from sequencing

Alexej Abyzov

Department of Health Sciences Research,
Center for Individualized Medicine, Mayo Clinic
abyzov.alexej@mayo.edu

Progress in technology has made individual genome sequencing a clinical reality, with partial genome sequencing already in use in clinical care. In fact, it is expected that within a few years whole genome sequencing will be a standard procedure that will allow the discovery of personal genomic variants of all types, thus greatly facilitating individualized medicine. However, fast and reliable analysis of such data is challenging, and improvements in analytics are needed before the clinical potential of whole genome sequencing can be realized. Specifically, copy number variations, which account for a large proportion of human genetic diversity, are frequently observed in cancer and have been associated with multiple diseases, cancer susceptibility, cancer progression and invasiveness, individual response to treatment, and patients' quality of life after treatment (i.e., emergence of side effects). Therefore, comprehensive identification and analysis of copy-number variants/alterations (CNVs/CNAs) will help us more fully elucidate the biology of their functional effects on human health (in particular, for cancer emergence and progression) and will facilitate clinical diagnostics and treatment.

However, the ability to detect CNV/CNA from sequencing is not fully utilized due to immature analytical approaches. CNVnator was originally developed and is currently widely used to detect germline CNVs from depth of coverage analysis of next-generation sequencing reads. Its routine application to the detection and analysis of CNAs in cancer samples is challenged by the sample purity and the presence of multiple subclones, each with its own CNAs, and frequent large aneuploidies. Our project is to continue development and enhancement of CNVnator to overcome these challenges to the extent that would allow routine application of the software in clinical setting for the analysis of germline and somatic CNVs/CNAs.

We will demonstrate examples of applying CNVnator to (1) analysis of highly aneuploid cancer cell line K562, (2) detect somatic mosaic CNA in urine samples, and (3) track individual subclones during neoplastic transformation in colorectal cancer.

Poster 50: Database and tools for functional inference and mechanistic insight into somatic cancer mutations

Xue Lei[1], Boshen Wang[1], Chia-Yi Chou, Alan Perez-Rathke[1], Jie Liang[1*] and Yan-Yuan Tseng[2*]

[1]Department of Bioengineering, University of Illinois at Chicago
[2*] Center for Molecular Medicine and Genetics
School of Medicine, Wayne State University
jliang@uic.edu and ytseng@wayne.edu

With the rapid progress of cancer genome studies, many missense variants in populations of cells at different stages of cancer have been identified. However, it remains challenging to understand the roles of cancer-related variants. We have developed a computational method called METS (**M**utational **E**ffect on **T**opological **S**urface**)** for assessing the effects of missense variants with improved accuracy. METS exploits structural information of precisely computed protein surface pockets that are likely involved in biochemical effects. We have mapped > 1.2 million records from the cancer-related variants in Catalogue of Somatic Mutations in Cancer (COSMIC) and from the residue variants in dbSNP onto the surfaces of ~32,000 human protein 3D structures in the Protein Data Bank (PDB). Based on B150 release of dbSNP, 2,756 nsSNP records associated with 659 human disorders in the OMIM database are geometrically mapped to 4,157 structures. Our results show that a large portion of these missense mutations are located on protein surface pockets. To examine their effects on protein structure and function, Wwe describe in detail how mutational effects of each variant can be assessed using examples of several oncoproteins. In addition, we discuss how novel candidate variants likely to be highly relevant to cancer development can be predicted. Furthermore, we discuss our findings on higher-order cooperative units of cancer variants.

Poster 51: MMTF-Spark: Interactive, Scalable, and Reproducible Datamining of 3D Macromolecular Structures

Peter W. Rose

Structural Bioinformatics Laboratory, San Diego Supercomputer Center, UC San Diego, La Jolla, CA

Advances in Structural Bioinformatics are driven by the fast growth in experimental 3D structures and integration with even larger sets of sequence and protein function data. At the same time, the field of Data Science has created new technologies for re-engineering legacy software pipelines to make them scalable, easy to use, reproducible, reusable, and sharable.

Here, we describe the MMTF-Spark/PySpark [1] project that combines three key components to create such an infrastructure: 1. Interactive Jupyter notebooks to run ad hoc analyses, data mining, machine learning, and visualization of 3D structure and sequence datasets, 2. A scalable compute infrastructure to run these analyses interactively across large datasets, e.g., the entire PDB, using a highly efficient, compressed data representations [2, 3] and the Apache Spark framework for distributed parallel computing, 3. A library of methods for data mining and analysis of 3D structure and sequence data, capitalizing on the rich data analytics, visualization, and machine/deep learning tools available in the Python ecosystem.

Scientists face a number of complex and time consuming barriers when applying structural bioinformatics analysis, including complex software setups, non-interoperable data formats and software applications, lack of documentation, simple examples, and tutorials. Given the large datasets, biologists routinely apply computational tools and automation pipelines in their research. However, there is a long tail of ad-hoc, one-off, questions that biologists ask that cannot be answered using available web resources or workflow systems that focus on common tasks. In this project, we provide a self-contained programming environment that caters to scientists with varying computational skills and needs, ranging from biologists with basic programming skills, to structural and computational biologists who want to share their work, to data scientists who seek access to bioinformatics datasets to benchmark new machine learning methods. A key advantage of this environment is interactivity, which enables iterative exploration. By combining documentation, data sets, analysis code, results, and interactive visualizations in Jupyter notebooks, the steps of an interactive session can be captured, reproduced, and shared.

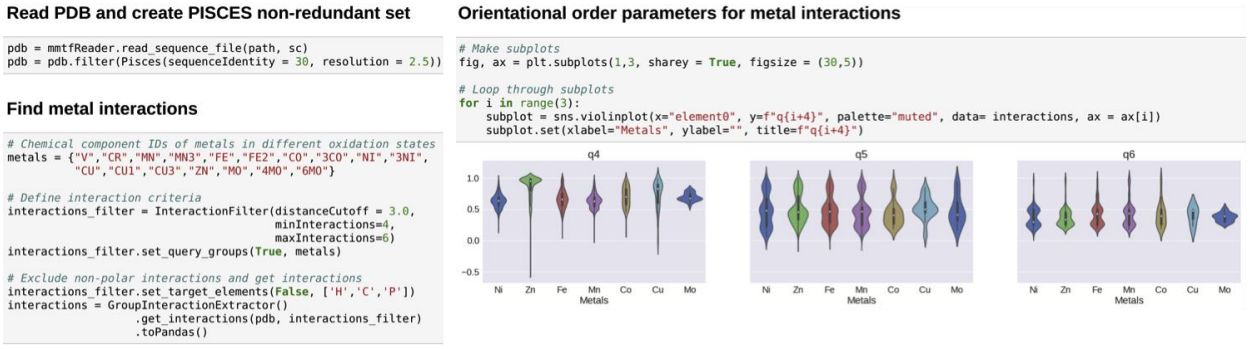


Figure: Example of extracting metal interactions from the PDB and analyzing the metal coordination geometry with orientational order parameters using MMTF-PySpark in a Jupyter Notebook.

Acknowledgements

This project was supported by the NCI of the NIH under award number U01 CA198942.

References

1. https://github.com/sbl-sdsc/mmtf-spark, https://github.com/sbl-sdsc/mmtf-pyspark
2. Bradley AR, et al. (2017) MMTF - an efficient file format for the transmission, visualization, and analysis of macromolecular structures. PLOS Computational Biology 13(6): e1005575.
3. Valasatava Y, et al. (2017) Towards an efficient compression of 3D coordinates of macromolecular structures. PLOS ONE 12(3): e0174846.

Poster 52: Integrated Querying of Biological Network Databases

Mehmet Koyutürk

(1) Department of Electrical Engineering and Computer Science
(2) Center for Proteomics and Bioinformatics
Case Western Reserve University

mehmet.koyuturk@case.edu

In biomedical applications, network models are commonly used to represent interactions and higher-level associations among biological entities. Integrated analyses of these interaction and association data has proven useful in extracting knowledge, and generating novel hypotheses for biomedical research. However, existing computational infrastructure for storing and querying network data target networks at smaller scales. More sophisticated tasks require bulk download of data from different databases, followed by in-house processing, integration, and analysis. In this work, we describe algorithms and indexing techniques that use results from classical linear algebra for fast processing of network proximity queries on very large networks. These algorithms enable compressed storage of integrated networks and efficient processing of sophisticated queries on these networks. We demonstrate applications of these algorithms in the context of drug repositioning, where we model prediction of drug response as a link prediction problem on a heterogeneous network, and use network proximity computed on a series of networks as an indicator of a given sample's sensitivity to a given drug.

Poster 53: Multiomics Data Compression

Olgica Milenkovic[1] Mikel Hernaez[1], Idoia Ochoa[1] and Jian[1]

[1] University of Illinois, Urbana-Champaign

Compression is a data processing task that has the goal to remove redundant and uninformative information from raw datasets. Given the various forms of datasets encountered in multiomics that are used for downstream processing, visualization and learning tasks, it is of importance to identify suitable compression methods for each level of the multiomics data storage platform.

We discuss three compression methodologies - lossless compression, quantization and extrinsic information extraction - adapted to specific data bundles encountered in multiomics studies. We also report on several new lines of work in learning and compression that may be used in multiple downstream data processing pipelines.

Poster 54: The Exascale Deep Learning and Simulation Enabled Precision Medicine for Cancer

George Zaki[1], Eric Stahlberg[1*], Tom Brettin[2] and Rick Stevens[2*]

[1*]Frederick National Laboratory for Cancer Research, eric.stahlberg@nih.gov
[2*]Argonne National Laboratory, stevens@anl.gov

As part of the NCI-DOE collaboration on the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C), the Department of Energy national laboratories (Argonne, Lawrence Livermore, Los Alamos and Oak Ridge) together with Frederick National Laboratory for Cancer Research are jointly working in the development of CANDLE – a distributed learning environment for cancer. Also supported by the DOE Exascale Computing Initiative, the system extends several existing and emerging deep learning frameworks to support scalable deep learning applications in cancer, enabling deep learning to take advantage of both the scale and technologies being developed for Exascale computing.

Working closely with domain scientists at the NCI and Frederick National Laboratory, the CANDLE environment development is driven by addressing emerging deep learning capabilities required in cancer challenge areas at the molecular, cellular and population scales. Deep learning is finding application in feature detection of key protein interactions at the molecular scale, in predicting cancer response to treatments at the cellular level, and in scalable automated feature extraction from large volumes of text-based pathology reports used in precision cancer surveillance. Benchmark examples are provided at the download site for each of these challenge areas, including implementations of example deep learning implementations of auto-encoders, long short-term memory, and recurrent neural networks. As a general framework, the CANDLE environment is extensible to additional application areas able to take advantage of deep learning including image processing, computational steering, and applications beyond cancer.

Available as open source hosted at GitHub, CANDLE version 0.2 was released in April 2018 and is available for download at https://github.com/ECP-CANDLE . CANDLE has been utilized in several modes, including containerized delivery as well as through installation from download.

Poster 55: The NCI Cancer Research Data Commons

Tanja Davidsen, Allen Dearry, Juli Klemm, Eve Shalley, Zhining Wang*, Tony Kerlavage

Center for Biomedical Informatics and Information Technology & *Center for Cancer Genomics, National Cancer Institute, NIH

As the -omics sciences increase the volume of data collection, the need for big data solutions intensifies. Biomedical informatics has reached a turning point where key innovations in data storage and distribution such as compression algorithms, indexing systems, and cloud platforms must be leveraged.

In addition to the data curation and storage needs of modern biomedical research, other challenges include the development of robust analytical tools, as well as infrastructure and funding models to support these efforts. As data generation expands, local storage, and computational solutions become less feasible. Thus, NCI has set out to build the NCI Cancer Research Data Commons (NCI CRDC), a cloud-based infrastructure in support of data sharing, tool development, and compute capacity to democratize big data analysis and to increase collaboration among researchers.

This poster describes three major aspects of the NCI CRDC, first the Genomics Data Commons (GDC), a unified data repository supporting cancer genomic studies, developed by University of Chicago, second the three NCI Cloud Resource projects, each independently developed by the Broad Institute, the Institute for Systems Biology and Seven Bridges Genomics, and third a Data Commons Framework to support common services across the NCI CRDC. The GDC provides the authoritative NCI reference data set that will be available through the cloud platforms that allow researchers to compute on the data. Each NCI Cloud Resource leverages different technical approaches that co-locate the data with computational capacity accessible through Application Programming Interfaces (APIs) to provide secure access to data and tools for biologists, bioinformaticians and developers of analytic tools. The DCF will provide a backbone of core services that components of the NCI CRDC can utilize.

The GDC, NCI Cloud Resources, and the DCF serve to democratize access to cancer -omic data and provide broadly-available, cost-effective computational support that are critical to the demands of modern cancer research and precision medicine.

Poster 56: NCI Cancer Research Data Commons Nodes in Development: The Proteomics Data Commons and Imaging Data Commons

Izumi Hinkson, Stephen Jett

CBIIT, National Cancer Institute

To foster collaboration, increase data access and accelerate discovery, the NCI aims to enhance data sharing within the cancer proteomics and imaging communities, and beyond. Researchers' ability to access diverse datasets, and perform robust and reproducible analyses is currently stifled by the siloed nature of the current informatics infrastructure. To mitigate this, the NCI Cancer Research Data Commons (CRDC) aims to support cancer research initiatives through sustaining multidisciplinary bioinformatics resources such as digital data knowledge bases, or nodes, centered on different research and clinical data including proteomics and imaging. As nodes in the CRDC, the overarching goals of the Proteomic Data Commons (PDC) and Imaging Data Commons (IDC) are to democratize access to cancer-related proteomic and imaging datasets, respectively, as well as to provide sustainable computational support to the cancer research community. The PDC and IDC will be cloud-based infrastructures that interoperate with other nodes within the CRDC such as the Genomic Data Commons (GDC), enabling researchers to seamlessly perform interdisciplinary analyses at scale. The CRDC, via the PDC, IDC and GDC, seeks to empower the cancer-research community, including scientists working in both intramural and extramural laboratories, with the necessary informatics capabilities to carry out large-scale, multi-omic data analysis.