AI MEETS CANCER RESEARCH SYMPOSIUM

SALA D'ACTES
PLAZA EUSEBI GÜELL NR 6
VERTEX BUILDING, CAMPUS NORD AT
UNIVERSITAT POLITECNICA DE CATALUNYA (UPC)
BARCELONATECH,
BARCELONA, SPAIN

NOV
29
30
2023

We graciously welcome you to the inaugural AI Meets Cancer Research Symposium in Barcelona, Spain.

This international event is a joint program with the Barcelona Supercomputing Center (BSC), the Scientific Foundation of the Spanish Association Against Cancer, and two Columbia University Medical Center groups: the Herbert Irving Comprehensive Cancer Center (HICCC) and the Program for Mathematical Genomics, Dept. of Systems Biology (PMG). This symposium will be an exchange of ideas to develop collaboration and innovation in cancer research and education.

SYMPOSIUM CHAIRS:

MOHAMMED ALQURAISHI, PHD, PROGRAM FOR MATHEMATICAL GENOMICS, DEPT. OF SYSTEMS BIOLOGY, COLUMBIA UNIVERSITY MEDICAL CENTER.
RAUL RABADAN, PHD, PROGRAM FOR MATHEMATICAL GENOMICS, DEPT. OF SYSTEMS BIOLOGY, COLUMBIA UNIVERSITY MEDICAL CENTER.
IRENE SÁNCHEZ GARCÍA, PHD, SCIENTIFIC FOUNDATION OF THE SPANISH ASSOCIATION AGAINST CANCER.
EMER SMYTH, PHD, HERBERT IRVING COMPREHENSIVE CANCER CENTER - COLUMBIA UNIVERSITY MEDICAL CENTER.
ALFONSO VALENCIA, PHD, BARCELONA SUPERCOMPUTING CENTER.

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputació
COLUMBIA Columbia University Herbert Irving Comprehensive Cancer Center
THE PROGRAM FOR MATHEMATICAL GENOMICS
fundación científica asociación española contra el cáncer

AI MEETS CANCER RESEARCH SYMPOSIUM

NOV 29, 2023

10:00 AM INTRO AND WELCOME SYMPOSIUM CHAIRS

MOHAMMED ALQURAISHI, PHD
Program for Mathematical Genomics, Department of Systems Biology,
Columbia University Medical Center
RAUL RABADAN, PHD
Program for Mathematical Genomics, Department of Systems Biology,
Columbia University Medical Center
IRENE SÁNCHEZ GARCÍA, PHD
Scientific Foundation of the Spanish Association Against Cancer
EMER SMYTH, PHD
Herbert Irving Comprehensive Cancer Center,
Columbia University Irving Medical Center
ALFONSO VALENCIA, PHD
Computational Biology Life Sciences Group,
Barcelona Supercomputing Center

10:30-12:10 1ST SESSION CANCER RESEARCH

CHAIR: RAUL RABADAN, PHD

MOHAMMED ALQURAISHI, PHD
Program for Mathematical Genomics, Department of Systems Biology,
Columbia University Medical Center

The State of Protein Structure Prediction and Friends

AlphaFold2 revolutionized structural biology by accurately predicting protein structures from sequence. Its implementation however (i) lacks the code and data required to train models for new tasks, such as predicting alternate protein conformations or antibody structures, (ii) is unoptimized for commercially available computing hardware, making large- scale prediction campaigns impractical, and (iii) remains poorly understood with respect to how training data and regimen influence accuracy. Here we report OpenFold, an optimized and trainable version of AlphaFold2. We train OpenFold from scratch and demonstrate that it fully reproduces AlphaFold2's accuracy. By analyzing OpenFold. training, we find new relationships between data size/diversity and prediction accuracy and gain insights into how OpenFold learns to fold proteins during its training process.

10:30-12:10 1ST SESSION CANCER RESEARCH *(CONTINUED)*

CHAIR: RAUL RABADAN, PHD

MACHA NIKOLSKI, PHD
Head of "Computational Biology and Bioinformatics" lab
CNRS -IBGC, Bordeaux University, France

TrialMatchAI: An AI-powered Natural Language Processing System for the Automatic Matching of Cancer Patients to Precision Clinical Trials

Allocating patients to precision clinical trials remains a major hurdle, often leading to long delays in providing life-saving treatments and impeding scientific progress. This challenge is in part due to two factors that underscore the rapidly changing landscape of precision oncology. The first factor is the increasing complexity of patients' genetic and clinical data that accompanies the spread of affordable next-generation genome sequencing and advances in biotechnology. The second factor is inherent to clinical trials in public databases whose eligibility criteria are presented as unstructured free texts, allowing for substantial flexibility in terminology and writing styles. These factors render standard patient-trial matchmaking strategies increasingly impractical. In this presentation, we emphasize the need for an artificial intelligence (AI) approach to find the best suited clinical trials for each cancer patient, automating the matchmaking process with the goal of reducing recruitment times.

We will present TrialMatchAI, an open-source tool that leverages natural language processing methods centered on Large Language Models (LLMs) to enable contextual clinical language comprehension and data structuring and harmonization. We fine-tuned pre-trained LLMs, such as BioBERT, to recognize and normalize biomedical entity classes often found in clinical trial texts, such as genes, proteins, diseases, drugs, and mutations. The contextual entity recognition and normalization via LLMs alleviated the problem of inconsistent vocabularies, such as gene aliasing, enabling the structuring of clinical trial text into a queryable format. We leverage the same LLM-based approach to structure patient data, such as reports of actionable mutations and electronic health records, and harmonize it with clinical trial data for downstream patient-trial matchmaking.

Barcelona Supercomputing Center
Centro Nacional de Supercomputació
COLUMBIA
COLUMBIA UNIVERSITY HERBERT IRVING COMPREHENSIVE CANCER CENTER
THE PROGRAM FOR MATHEMATICAL GENOMICS
fundación científica asociación española contra el cáncer

10:30-12:10 1ST SESSION CANCER RESEARCH (*CONTINUED*)

CHAIR: RAUL RABADAN, PHD

ALISON TAYLOR, PHD
Herbert Irving Comprehensive Cancer Center,
Columbia University Irving Medical Center

Functional and computational approaches to uncover selection advantages of cancer aneuploidy

Aneuploidy, including the gain or loss of whole chromosomes or chromosome arms, is a near-universal feature of cancer. We previously applied methods that define chromosome arm aneuploidy to over 10,000 tumors in the Cancer Genome Atlas (TCGA). Cancer subtypes are often characterized by tumor specific patterns of chromosome arm copy number alterations and breakpoints; for example, squamous cell carcinomas (SCCs) from different tissues of origin are characterized by chromosome 3p(chr3p)loss and chromosome 3q (chr3q) gain. From the TCGA aneuploidy data, we developed an algorithm called BISCUT to distinguish peak regions of aneuploidy breakpoints on each chromosome arm. BISCUT identified loci affected by broad copy number alterations that provide fitness advantages or disadvantages both within individual cancer types and across cancers. Our analyses are consistent with selection being the primary driver of aneuploidy events in cancer.

We next wanted to validate some BISCUT peaks without a known driver, to identify potential gene deletions that are beneficial in cancer cells. We focused on chromosome 8p (chr8p), as this arm is frequently deleted across cancer types,but no strong tumor suppressors have been identified. Recent advances in genome engineering allow generation of large chromosomal alterations and validation of findings from patient genomic data. For this study, we used our CRISPR-Cas9 arm-deletion system to delete chr8p in human immortalized epithelial cells. Cells with chr8p deletion showed lower amounts of cell death in culture. Knockdown of WRN, one of the two genesin the smallest chr8p BISCUT peak, was sufficient to reproduce this phenotype, suggesting that WRN haploinsufficiency may be beneficial to tumor development.Our genome engineering approach to model chromosome arm aneuploidies provides a robust model to validate drivers from aneuploidy events, which will be critical to address the gap in our understanding of aneuploidy in cancer.In addition, our methods have identified consequences of individual aneuploidy events, which will lead to new precision oncology targets for patients.

FATIMA AL-SHAHROUR, PHD
Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.

Cancer treatment prediction using single-cell-guided drug prioritization

The heterogeneity of cancer cells poses significant challenges in designing effective treatment strategies. Multi-omics data, and specifically, those derived from single-cell resolution techniques, offer an unprecedented opportunity to address therapeutic heterogeneity in tumors. We developed Beyondcell [1], a method for identifying tumour cell subpopulations with distinct drug responses and propose cancer-specific treatments using single-cell transcriptomics. In this talk, we will present the application of Beyondcell to define therapeutic modules (TMs), that is, groups of drugs for which we predict a similar behavior in specific malignant cells subpopulations. To do so, we have created the Therapeutic Cancer Cell Atlas (TCCA), an annotated and integrated single-cell transcriptomics dataset built from 36 public studies, that spans 1M malignant cells (and 2M cells from the tumor microenvironment), involving 800 patients and encompassing 41 tumor types. Our results in the TCCA show TMs that are shared between malignant subpopulations across human tumours. We will also present how we will apply these findings in the context of the Spanish National Network on Brain Metastasis cohort (RENACER) [2].

1.Fustero-Torre C. et al. (2021) Genome Med. 13:187.
2.Valiente M, Ortega-Paino E. (2023) Trends Cancer.S2405-8033(23)00186-3.

12:10 – 12:30 Break

AI MEETS CANCER RESEARCH SYMPOSIUM

NOV 29, 2023

12:30-14:10 2ND SESSION CANCER RESEARCH
CHAIR: RAUL RABADAN, PHD

DAVID TORRENTS, PHD

ICREA Research Professor at the Barcelona Supercomputing Center

AI for genomics. From accessing the data to answering the question.

The development of disease risk predictors for cancer and complex diseases, and the generation of preventive protocols need from the understanding of the underlying genetic architecture of pathological traits, which remains elusive. With time, researchers have identified isolated pieces using approaches based on classical statistical approaches (e.g. regression analysis), in the form of genetic markers associated with the disease. This is providing the initial set of variants and tools to start building disease prediction protocols and to understand diseases. But more integrative approaches are needed to capture the real complexity of these traits. AI offers the possibility of answering more complex and ambitious questions about the genetics behind diseases. But the adaptation of AI methodologies to genomics remains a challenge. The nature of the genome, and its level of implication in disease is different from that of proteins and other types of information, like text. The lack of traceability and transparency are also huge limitations when considering deep learning approaches, like LLM, asat this point, current research in genomics prioritizes marker discovery and understanding the disease, and not opaque classification models. Therefore, we need to develop new AI approaches for genomics, as finding ways of understanding the complexity of the molecular and genetic basis of disease will open the possibility for robust and global prediction and prevention protocols.

BENJAMIN IZAR, MD, PHD

Herbert Irving Comprehensive Cancer Center,
Columbia University Irving Medical Center

Systematic dissection of tumor-immune interactions using single-cell CRISPR-screens

Immune evasion is a hallmark of cancer, yet the underlying mechanisms are often unknown in many patients. How cancer cells balance co-stimulatory and co-inhibitory signals on their surface is poorly understood. Here, we present a framework for studying such complex interactions through building a platform that ranges from patient-centered single-cell genomics,single-cell CRISPR-screening that enables simultaneous immune fitness screening with phenotypic readouts, and precisely informed by the latter, uncover protein-interactions on cancer cells driving immune response and escape. Previously, using single-cell transcriptomics analyses, we had identified the co-stimulator CD58 as part of a cancer-cell-intrinsic immune checkpoint resistance signature in patient melanoma tissue. We subsequently validated CD58 loss as a driver of immune evasion using a patient-derived co-culture model of cancer and cytotoxic tumor infiltrating lymphocytes in a pooled single-cell perturbation experiment, where we additionally observed concurrent upregulation of PD-L1 protein expression in melanoma cells with CD58 loss. Next, we uncovered the mechanisms of immune evasion mediated by CD58 loss, including impaired T cell activation and infiltration within tumors, as well as inhibitory signaling by PD-L1 via a shared regulator, CMTM6. Thus, cancer-cell-intrinsic reduction of CD58 represents a multi-faceted determinant of immune evasion. Furthermore, its reciprocal interaction with PD-L1 via CMTM6 provides critical insights into how co-inhibitory and co-stimulatory immune cues are regulated. We now develop novel experiment and analytical tools to study all potential protein interactions relevant in cancer in a systematic manner.

12:30-14:10 2ND SESSION CANCER RESEARCH *(CONTINUED)*
CHAIR: RAUL RABADAN, PHD

MARTA MELÉ, PHD
Group Leader,Transcriptomics and Functional Genomics Lab
Life Sciences Department, Barcelona Supercomputing Center

Using omics and AI to study the human transcriptome: from genes to individuals

Understanding the consequences of individual transcriptome variation is fundamental to deciphering human biology and disease. My lab uses novel statistical approaches, high throughput functional assays and machine learning to tackle this fundamental question. In this talk, I will give an overview of some projects developed in the lab to tackle how human diversity emerges. I will present a newly developed statistical framework to quantify the contributions of several individual traits as drivers of gene expression and alternative splicing variation across human tissues. Furthermore, we combine transcriptomics and machine learning approaches to analyze histopathological images and revealed a systemic contribution of types 1 and 2 diabetes to tissue transcriptome variation with the strongest signal in nerve, where we identify novel genes related to diabetic neuropathy. We have expanded this framework to study the impact of tobacco smoking on the transcritome, the epigenome and, the tissue architecture across multiple human tissues. We observe a widespread impact of smoking in the human body, with large changes in expression and methylation. We then use convolutional neural networks on lung and thyroid histological images and confirm important architectural differences between never smokers and smokers. Notably, we observed concordant additive effects of smoking and aging in most tissues suggesting that smoking can have similar consequences than those of biological aging. Finally, I will present a novel approach to study the impact of individual traits such as sex and age at single-cell resolution combining information across different cohorts. Overall, our multi-tissue and multi-trait approach provides an extensive characterization of the main drivers of human transcriptome variation in health and disease.

SARA ZACCARA, PHD
Department of Systems Biology,
Columbia University Irving Medical Center

Understanding the complexity of the m6A regulatory program

For decades we believed that information in messenger RNA was confined to its nucleotide sequence. However, new methods to profile mRNA beyond its mere sequence have revealed that mRNA contains additional information in the form of chemical modifications. The most abundant modified nucleotide isN6-methyladenosine (m6A), a methyl modification of adenosine. M6A has the well-established function of triggering the degradation of modified mRNAs, compared to unmodified RNA.

Importantly, in various cancer types, cells can reshape their transcriptome by altering m6A mRNA levels. In the last few years, our work has demonstrated the dependency of acute myeloid leukemia (AML) on m6A. We identified, for the first time, the exact position of each m6A site in AML cells and then showed that depletion of m6A in AML cell lines and primary leukemia blasts led to cell growth inhibition, cell cycle arrest, induction of apoptosis, and differentiation.

Although m6A-mRNA appears highly altered in AML, it is unknown whether (1) these methylation events could be used as a diagnostic tool for detecting the earliest stages of AML and (2) the mechanisms underlying m6A dysregulation during AML progression. Here, we will discuss our current efforts in answering these major questions.

14:10 – 15:40 Lunch Break

AI MEETS CANCER RESEARCH SYMPOSIUM

NOV 29, 2023

15:40 -16:55 3RD SESSION AI
CHAIR: ALFONSO VALENCIA, PHD

NATASA PRZULJ, PHD
ICREA & BSC, natasha@bsc.es
Life Sciences - Integrative Computational
Network Biology, Barcelona Supercomputing Center

Multi-Omics Data Fusion for Enabling Cancer Precision Medicine

Increasing quantities of heterogeneous, interconnected, systems-level, molecular (multi-omic) dataare becoming available. They provide complementary information about cells, tissues and diseases. We need to utilize them to better stratify patients into risk groups, discover new biomarkers, re-purpose known and discover new drugs to personalize medical treatment. This is nontrivial, because of computational intractability of many underlying problems, necessitating the development of algorithms for finding approximate solutions (heuristics).

We develop a versatile data fusion (integration) machine learning (ML) framework to address key challenges in precision medicine from these data: better stratification of patients, prediction of biomarkers, and re-purposing of approved drugs to particular patient groups, applied to cancer and other diseases. Our new methods stem from graph-regularized non-negative matrix tri-factorization (NMTF), a machine learning technique for dimensionality reduction, inference and co-clustering of heterogeneous datasets, coupled with novel network science algorithms. We utilize our new framework to develop methodologies for improving the understanding the molecular organization and disease from the omics network embedding space.

DAVIDE CIRILLO, PHD
Life Sciences - Machine Learning for Biomedical Research,
Barcelona Supercomputing Center

Advancing precision oncology with multilayer networks and synthetic data generation

Multilayer networks and synthetic data generation play pivotal roles in advancing precision oncology by addressing challenges such as privacy preservation, modeling with limited data, and enhanced interpretability. The analysis of community structures within multilayer networks, composed of interconnected bio-entities through relational associations, reveal critical modules for patient stratification. On the other hand, synthetic data generation models, such as variational autoencoders with explainable components, allow transparent and targeted data augmentation, unveiling novel patterns and enabling knowledge discovery. These approaches were applied to characterize adult and pediatric tumors demonstrating their efficacy in navigating the complexities of precision oncology research. The synergy between multilayer networks and synthetic data generation proves instrumental in unlocking insights that contribute to a more comprehensive understanding of cancer, facilitating more precise and personalized healthcare strategies

15:40 -16:55 3RD SESSION AI *(CONTINUED)*
CHAIR: ALFONSO VALENCIA, PHD

MARIA RODRIGUEZ MARTINEZ, PHD
IBM Research - Zurich
AI-driven modelling of immune receptors.
In recent years, deep learning models have led to outstanding breakthroughs in many bioinformatic fields. However, most models behave as black boxes, which might potentially hide data biases, incorrect hypotheses, or even software errors. In this task, I will focus on the problem of predicting T cell receptors (TCRs) specificity, which is vital for developing effective T cell-based immunotherapies. First, I will introduce TITAN (Tcr epITope bimodal Attention Network), a bimodal neural network that explicitly encodes both TCR sequences and epitopes to enable the independent study of the generalization capabilities to unseen TCRs and/or epitopes. Furthermore, TITAN exploits attention mechanisms to identify the most informative amino acids to make a prediction, and in doing so, can identify challenging situations where data is insufficient to allow the model to generalize to unseen epitopes. Next, I will discuss how recent protein language models, both generalist models trained on millions of unlabeled amino acid sequences and domain-specific models trained on immune receptor sequences, can achieve similar performances to traditional deep learning models with significantly smaller architectures, hence reducing data requirements. However, a challenge with current transformer-based protein models is their lack of interpretability. To address this limitation, I will finally discuss DECODE, an easy-to-use, interpretable pipeline that extracts binding rules governing the predictions of any black-box model to predict T cell specificity.

16:55 – 17:15 BREAK

17:15 - 18:30 4TH SESSION: AI
CHAIR: MOHAMMED ALQURAISHI, PHD

JULIO SAEZ RODRIGUEZ, PHD
Molecular Medicine Partnership Unit (MMPU)
EMBL - University Hospital Heidelberg
Knowledge-based machine learning to extract disease mechanisms from omics data
Omics approaches, in particular those with single-cell and spatial resolution, provide unique opportunities to study the deregulation of intra- and inter-cellular processes in cancer using AI-based approaches. The use of prior biological knowledge allows us to reduce the dimensionality and increase the interpretability of the data, thereby supporting the application of these approaches, in particular by extracting from the data features describing the activity of molecular processes such as signaling pathways, gene regulatory networks, and cell-cell communication events. In this talk, I will present resources and methods from our group to effectively capture and deploy prior knowledge from the public domain to extract disease mechanisms from omics data.

AI MEETS CANCER RESEARCH SYMPOSIUM

NOV 29, 2023

17:15 - 18:30 4TH SESSION: AI *(CONTINUED)*
CHAIR: MOHAMMED ALQURAISHI, PHD

TAL KOREM, PHD
Program for Mathematical Genomics, Department of Systems Biology, Columbia University Medical Center

Addressing contamination and bias: towards robust microbiome analysis in cancer

Recent years have seen tremendous potential in studying the interaction between the microbiome and various tumors,from associations of the gut microbiome with tumor presence and treatment response to detecting strong signatures of microbial DNA within the tumor itself. However, this research has also encountered significant challenges, including those of contamination and batch effects. These issues are inherent to microbiome studies, and particularly in ecosystems with low microbial biomass, such astumors.I will present two new approaches for addressing these challenges: first, we introduce a framework for modeling contamination sources, rather than identifying contaminating taxa, which facilitates more accurate in silico identification and removal of contamination; second, we present a new batch correction method that identifies and corrects processing bias in an interpretable manner. We applying these methods to data from previous studies and demonstrate stronger associations with the microbiome.

NURIA MALATS, MD, MPH, PhD
Spanish National Cancer Research Centre (CNIO), Madrid
nmalats@cnio.es

Integrating omics and non-omics data through AI: An epidemiological application

Epidemiological models aim to explore the association of risk factors and biomarkers with the risk of suffering a disease, i.e., cancer. The conceptual model is very basic: cases and controls (phenotype) are distributed according to the exposure of interest and, applying a logistic regression adjusting for potential confounders, we calculate the risk of suffering such cancer of interest for the exposed taking the non-exposed as a reference. However, most cancers are complex or multifactorial diseases in which a risk factors, genetic or non-genetic, intervening and interacting over many years. Therefore, the risk models should mimic such complexity.

We employ a wide variety of biomarkers, including omics data, to better characterise the exposome, the genome, and the phenome and integrate them with non-omics data by using machine and deep learning approaches. However, this integration endeavour poses important challenges. In addition to the complexities in the design and analysis, there is the possibility of incurring a selection bias due to the lack of data in certain subgroups of the study population, the heterogeneity of the phenome/cancer under study and the different ways of characterizing it, the way to manage the different nature of the data and their hierarchical dependency, the fairness between omics and non-omics data, the correlation and interaction between both types of data, the limitation of the statistical power of the study taking into account a scenario in which there are more variables (parameters) than individuals, the lack of studies to validate/replicate the results, and the lack of functional information that helps with the functional interpretation of the results. Furthermore, it must be kept in mind that the carcinogenic process is dynamic and is active for many years during which environmental exposures and genetic factors, among others, interact. In a very simplistic way, current studies only take into account the moment in which the subject is included and the information is collected, not what happened years before. All these challenges may explain the lack of the study result replicability and widen the gap in translating the findings from the studies into clinical and public health interventions.

I will exemplify this complexity in the understanding of the complex relationships among the sporadic pancreatic cancer risk factors towards the definition of high-risk individuals that may benefit from entering screening programs.

18:30 - 19:30 NETWORKING - Open discussion to all.

10:30-11:45 5TH SESSION: AI & MOLECULES
CHAIR: RAUL RABADAN, PHD

YUFENG SHEN, PHD
Department of Systems Biology,
Columbia University Irving Medical Center

Predicting genetic effect of missense mutations by machine learning

Accurate prediction of genetic effect of missense variants is fundamentally important for disease gene discovery, clinical genetic diagnosis, personalized treatment, and protein engineering. Commonly used computational methods predict pathogenicity, which does not capture the quantitative impact on fitness in human. We developed a method, MisFit, to estimate selection coefficient of missense variants. MisFit jointly models the effect at a molecular level (D) and a population level (selection coefficient, S), assuming that in the same gene, missense variants with similar D would have similar S. We trained it by maximizing the probability of observed germline variant allele counts in 234,992 individuals of European ancestry. We show that S is informative in predicting allele frequency across ancestries and consistent with the fraction of de novo mutations observed in sites under strong selection. Further, S outperforms previous methods in prioritizing de novo missense variants in individuals with neurodevelopmental disorders. Finally, we show that predicted D and S are consistent with functional readout of deep mutational scan experiments of clinically important genes, including cancer driver genes.

DAVID JONES, PHD
Department Computer Science,
University College London UK. E-mail: d.t.jones@ucl.ac.uk

Using AlphaFold to Model Large Multidomain Proteins

Large multidomain proteins play a pivotal role in human disease, such as cancer, as their complex structural arrangements and interactions often underlie critical cellular processes and pathways, making them important targets for understanding disease mechanisms and potential therapeutic interventions. In this talk I will be discussing the ability of AlphaFold2 to model these kinds of proteins. I will continue by looking at the AlphaFold Protein Structure Database (AFDB), comprising predictions for all of the protein chains in UniProt, and highlight the current limitations in functional annotation stemming from the underlying complex multidomain structures of larger chains. To help address this issue, I will describe Merizo, a deep learning method for accurate automatic domain segmentation, which clusters residues into domains in a bottom-up manner. Merizo's training on CATH domains and fine-tuning on a subset of AFDB models make it applicable to both experimental and AFDB data. As a practical demonstration, we have applied Merizo to models fromthe human proteome, successfully identifying many new putative domains, but also further highlighting current limitations of predicting the structures of multidomain proteins.

CHARLOTTE BUNNE, PHD
Machine Learning for Personalized Medicine, ETH Zurich

TBD

11:45 – 12:05 BREAK

12:05- 13:20 6TH SESSION: AI & MOLECULES
CHAIR: DAVIDE CIRILLO, PHD

JIAN TANG, PHD
Mila-Quebec AI Institute,
CIFAR AI Research Chair

Geometric Deep Learning for Protein Understanding

Proteins are workhorses of living cells. Understanding the functions of proteins is critical to many applications such as biomedicine and synthetic biology. Thanks to recent biotechnology breakthroughs such as gene sequencing and Cro-EM, a large amountof protein data (such as protein sequences and structures) are generated,providing a huge opportunity for AI. As the functions of proteins are determined by their structures, in this talk, I will introduce our recent work on protein understanding based on protein 3D structures with geometric deep learning. I will introduce two lines of work including: (1) the first structure-based pretraining framework for protein representation learning, with applications in antibody design; (2) a generative diffusion model for protein side-chain conformation prediction, complementing existing progress of AlphaFold2 on protein backbone structure prediction.

VALENTINA BOEVA, PHD
Department of Computer Science, ETH Zürich, ETH AI Center

Discovery and characterization of shared transcriptional states across cancer patients from single-cell RNA sequencing data

Tumor heterogeneity is regarded as a significant obstacle to successful personalized cancer medicine. Specifically, multiple cancer types have been shown to exhibit heterogeneity in the transcriptional states of malignant cell seven within the same tumor. Some of these specific transcriptional states of malignant cells have been linked to cancer relapse and resistance to treatment. However, today there is no universal and easy-to-use computational method to extract information about shared transcriptional states from single-cell RNA sequencing measurements (scRNA-seq).

To reliably identify shared transcriptional states of cancer cells, we propose a novel computational tool,CanSig. CanSig automatically preprocesses, integrates, and analyzes cancer scRNA-seq data from multiple patients to provide novel signatures of shared transcriptional states; it also associates these states to known and potentially targetable biological pathways. CanSig jointly analyzes cells from multiple cancer patients while correcting for batch effects and differences in gene expressions caused by genetic heterogeneity. For this, CanSig automatically infers copy number variations in malignant cells and trains a deep learning architecture based on conditional variational autoencoders integrating scRNA-seq measurements.

In our benchmarks, CanSig shows state-of-the-art performance in data integration and reliably re-discovers known transcriptional signatures on four previously published cancer scRNA-seq datasets. For instance, CanSig re-discovers four main cellular states of glioblastoma cells previously reported by Neftel et al., Cell, 2019.We further illustrate CanSig's investigative potential by uncovering signatures of novel transcriptional states in several cancer types; some of the novel signatures are linked to such cancer hallmarks as cell migration and proliferation and are enriched in more advanced tumors. We also uncover signatures associated with specific genomic aberrations,such as gene copy number gains.

In conclusion, we demonstrate how one can detect shared transcriptional states in tumors using as input scRNA-seq data for cells with different genetic backgrounds and possibly exhibiting strong batch effects. Our approach can significantly facilitate the analysis of scRNA-seq cancer data and efficiently identify transcriptional signatures linked to known biological pathways.The CanSig method implemented in Python is available at https://github.com/BoevaLab/CanSig

AI MEETS CANCER RESEARCH SYMPOSIUM

NOV 30, 2023

12:05- 13:20 6TH SESSION: AI & MOLECULES *(CONTINUED)*
CHAIR: DAVIDE CIRILLO, PHD

PETIA RADEVA, PHD
Head of *"Artificial Intelligence and Biomedical Applications"* Consolidated Research Group Dept. of Mathematics and Computer Science, Universitat de Barcelona

Deep Learning: A Swiss Army Knife or a Threat to Oncologists?

Recent studies have illuminated a fascinating paradox: Artificial intelligence algorithms are proving to be more proficient at detecting breast cancers than highly experienced physicians, significantly alleviating the workload associated with mammogram readings. But does this signal a future where computers take the reins in determining a patient's cancer status? Deep Learning, a cutting-edge technology currently surpasses human expertise and prompt profound questions about its capabilities and limitations. In this presentation, we will delve into the strides, challenges, and potential implications of Deep Learning technology, fostering a discussion about its transformative potential in the realm of clinical practice.

13:30 - 15:00 LUNCH BREAK

15:00 – 16:40 7TH SESSION: CANCER RESEARCH
CHAIR: MOHAMMED ALQUAISHI, PHD

NURIA LOPEZ-BIGAS, PHD
Institute for Research in Biomedicine, ICREA

In silico saturation mutagenesis of cancer genes

Most mutations identified in tumors in cancer genes are mutations of unknown significance. We have built machine learning models, BoostDM, inspired in evolutionary biology to effectively identify driver mutations in each gene and cancer type. With those models we perform in silico saturation mutagenesis to outline blueprints of potential driver mutations in cancer genes. These blueprints support the interpretation of newly sequenced patients' tumors and the study of the mechanisms of tumorigenesis of cancer genes across tissues.

15:00 – 16:40 7TH SESSION: CANCER RESEARCH *(CONTINUED)*
CHAIR: MOHAMMED ALQURAISHI, PHD

JEAN GAUTIER, PHD
Herbert Irving Comprehensive Cancer Center,
Columbia University Irving Medical Center

3D organization of chromatin in the maintenance of genome stability

Chromatin is generally considered a crowded and entangled environment with stable, gel-like chromosome territories.Recent biophysical studies have challenged this view revealing that chromatin behaves more like a liquid. DNA transactions are organized within small nuclear domains assembled into biomolecular condensates through the combination of forces, protein-protein and protein-RNA interactions yielding fluid and dynamic chromatin compartments.

We study how multiscale organization of chromatin delineates interactions between close and distant loci to regulate transcription, repair and replication. Genome-wide analyses of rearrangements, chromosome translocations and insertions, reveal the dynamic nature of chromatin with recombination occurring between all chromosomes.

Live-cell microscopy and chromosome conformation capture approaches demonstrate the contributions of nuclear actin-based forces and condensate properties in assembling functional repair domains, which enhance faithful repair reactions at the expense of rare rearrangements with pathogenic potential. Furthermore, our work reveals an unanticipated connection between transcription and repair.

AI MEETS CANCER RESEARCH SYMPOSIUM

NOV 30, 2023

15:00 – 16:40 7TH SESSION: CANCER RESEARCH *(CONTINUED)*

CHAIR: MOHAMMED ALQURAISHI, PHD

EDUARD PORTA, PHD

Computational Biology Life Sciences Group,
Barcelona Supercomputing Center

Unraveling the spatial architecture of Cancer Hallmarks

Tumors are complex ecosystems with dozens of interacting cell types. The concept of Cancer Hallmarks distills this complexity into a set of underlying principles that govern tumor growth. In this talk I will show how we exploit this abstraction to explore the physical distribution of Cancer Hallmarks across 63 primary untreated tumors from 10 cancer types using spatial transcriptomics. We found that Hallmark activity is spatially organized–with 7 out of 13 Hallmarks consistently more active in cancer cells than within the non-cancerous tumor microenvironment (TME). The opposite is true for the remaining six Hallmarks. Additionally, we discovered that genomic distance between tumor subclones correlates with differences in Cancer Hallmark activity, even leading to clone-Hallmark specialization in some cases. Finally, we demonstrate interdependent relationships between Cancer Hallmarks at the junctions of TME and cancer compartments. In conclusion, including the spatial dimension, particularly through the lens of Cancer Hallmarks, can improve our understanding of tumor ecology.

TERESA PALOMERO, PHD

Department of Pathology & Cell Biology,
Columbia University Irving Medical Center

Unraveling Mechanismsof Transformations in Peripheral T-cell Lymphoma: Insights and Perspectives"

Peripheral T-cell lymphomas (PTCL) are a highly heterogeneousgroup of mature T-cell malignancies generally associated with poor prognosis. Malignant transformation in PTCL is the result of a complex interplay between tumor cells, tumor microenvironment and viral infections.Even with recent advances in the characterization of the molecular landscape of PTCL, diagnosis and classification remain challenging.Genomic characterization of angioimmunoblastic T-cell lymphoma (AITL) and PTCL not otherwise specified (PTCL, NOS), the most frequent subtypes of PTCL, has identified recurrent alterations in epigenetic regulators (including TET2, DNMT3A and IDH2), elements of the T-cell receptor pathway, and the TP53 tumor suppressor.Recently, our group described genomic alterations affecting the VAV1 oncogene and showed that expression of Vav1-Myo1f, a recurrent PTCL-associated VAV1 fusion, induces oncogenic transformation of CD4+T cells. Notably, mouse Vav1-Myo1f lymphomas show T helper type 2 immunophenotypic and transcriptional features analogous to high-risk GATA3+human PTCL. Single-cell transcriptome analysis reveals that Vav1-Myo1f alters T cell differentiation and leads to accumulation of tumor-associated macrophages (TAMs) in the tumor microenvironment, a feature linked with aggressiveness in human PTCL. Importantly, therapeutic targeting of TAMs induces strong anti-lymphoma effects, highlighting the lymphoma cells' dependency on the microenvironment. Better understanding of the molecular mechanisms leading to PTCL transformation will provide a strong foundation to improve the diagnosis, prognosisand targeted therapies for the treatment ofperipheral T-cell lymphomas (PTCL), aiming to enhance patient outcomes and quality of life.

16:40 SYMPOSIUM CLOSE - SYMPOSIUM CHAIRS