About

Joint Bioconductor Asia - Hong Kong Bioinformatics Symposium

The purpose of this symposium is to share scientific information with other researchers and scientists and to promote communication in the field of bioinformatics in Hong Kong and internationally. Additionally, this conference will also host multiple workshops in R and Bioconductor to enhance the education and training in computational biomedical sciences.

General Information

Date: 15-17 Oct 2023
Website: https://biocasia2023.bioconductor.org/
Venue: Boardroom, 1/F, Faculty Administration Wing (FAW), LKS Faculty of Medicine, 21 Sassoon Road

Co-chairs

Dario Strbenac
Ellis Patrick

Organizer Comittee

Dr. Yuanhua Huang, HKU
Dr. Joshua Ho, HKU
Dr. Xueyi Dong, WEHI
Dr. Yingying Wei, CUHK
Dr. Jiguang Wang, HKUST
Dr. Can Yang, HKUST

Monday, 16 Oct 2023

Evolving an ecosystem for scalable genomic data science

Prof. Vincent Carey

Harvard University

Bioconductor is entering its third decade as a central element of the genomic data science toolbox. The problem space addressed by the Bioconductor community and ecosystem has diversified greatly, while the solution components surfaced in Bioconductor are mostly R packages. There is much more under the surface, and our approaches to tackling technical and substantive problems are changing at various scales. New approaches to genomic data and annotation representation are needed. Configuration complexity must be reduced. Developer support to improve reliability and performance must dovetail with education of the user community to ensure the tooling matches the most pressing needs. I will discuss selected developments in each of these areas.

Improved Estimation of Functional Enrichment in SNP Heritability Using Feasible Generalized Least Squares

Mr. Zewei Xiong

Functional enrichment results typically implicate tissue or cell-type specific biological pathways in disease pathogenesis and as therapeutic targets. We propose generalized Linkage disequilibrium score regression (g-LDSC), a novel method that requires only genome-wide association studies (GWAS) summary level data to estimate functional enrichment. The method adopts the same assumptions and regression model formulation as stratified Linkage disequilibrium score regression (s-LDSC). While s-LDSC only partially utilizes LD information, our method utilizes the whole LD matrix which accounts for possible correlated error structure via a feasible generalized least squares estimation. We demonstrate through simulation studies under various scenarios, that g-LDSC provides more precise estimates of functional enrichment than s-LDSC, regardless of model misspecification. In an application to GWAS summary statistics of 15 traits from the UK Biobank, estimates of functional enrichment using g-LDSC were lower, and more realistic, than those obtained from s-LDSC. In addition, g-LDSC detected more significantly enriched functional annotations among 24 functional annotations for the 15 traits than s-LDSC (118 versus 51).

Systems approach for congruence and selection of cancer models towards precision medicine

Dr. Jian Zou

Chongqing Medical University

Cancer models are instrumental to substitute for human studies and expedite basic, translational and clinical cancer research. For a given cancer subtype, a wide selection of models, such as cell lines, patient-derived xenografts, tumoroids and genetically modified murine models, are often available to researchers. However, how to quantify their congruence to human tumors and to select the most appropriate cancer model is a largely unsolved issue. Here, we develop Congruence Analysis and Selection of CAncer Models (CASCAM), a statistical and machine learning framework for authenticating and selecting the most representative cancer models in pathway-specific and drug-relevant context using transcriptomic data. CASCAM offers harmonization between tumor and cancer model omics data, interpretable machine learning for congruence quantification, mechanistic investigation, and pathway-based topological visualization to determine the final cancer model selection. The workflow is presented using breast cancer invasive lobular carcinoma (ILC) subtype, while the method is generalizable to any cancer subtype for precision medicine development.

High-Dimensional Causal Mediation in Imaging Genetics Study

Prof. George Tseng

University of Pittsburgh

Causal mediation analysis provides a systematic approach to investigate the causal role of one or more mediators in an exposure-outcome association. In omics or imaging data analysis, mediators are often high-dimensional, which brings new statistical challenges. Existing methods either violate causal assumptions or fail in interpretable variable selection. Additionally, mediators are often highly correlated, presenting difficulties in selecting and prioritizing top mediators. To address these issues, we develop a framework using Partial Sum Statistic and Sample Splitting Strategies, namely PS5, for general high- dimensional causal mediation analyses. The method provides a powerful global mediation test satisfying causal assumptions, followed by an algorithm to select and prioritize active mediators with quantification of individual mediation contributions. We demonstrate its accurate type I error control, superior statistical power, reduced bias in mediation effect estimation, and high mediator selection accuracy using extensive simulations of varying levels of effect size, signal sparsity and correlation in mediators. Finally, we apply PS5 to a imaging genetics dataset of chronic obstructive pulmonary disease (COPD) patients (N=8,897) in the COPDGene study to examine the causal mediation of lung images (p=5,810) in the associations between polygenic risk score and lung function and between smoking exposure and lung function, separately. Both causal mediation analyses successfully estimate the global indirect effect and detect mediating image regions. We find a lung region in the lower lobe of right lung with strong and concordant mediation effect for both genetic and environmental exposure, which argues the potential of targeted treatment towards this region to control the severity of COPD due to genetic factor and cigarette smoking

A statistical method for Cross-population fine-mapping by leveraging genetic diversity and accounting for confounding bias

Dr. Mingxuan Cai

City University of Hong Kong

Fine-mapping prioritizes risk variants identified by genome-wide association studies (GWASs), serving as a critical step to uncover biological mechanisms underlying complex traits. However, several major challenges still remain for existing fine-mapping methods. First, the strong linkage disequilibrium among variants can limit the statistical power and resolution of fine-mapping. Second, it is computationally expensive to simultaneously search for multiple causal variants. Third, the confounding bias hidden in GWAS summary statistics can produce spurious signals. To address these challenges, we develop a statistical method for cross-population fine-mapping (XMAP) by leveraging genetic diversity and accounting for confounding bias. By using cross-population GWAS summary statistics from global biobanks and genomic consortia, we show that XMAP can achieve greater statistical power, better control of false positive rate, and substantially higher computational efficiency for identifying multiple causal signals, compared to existing methods. Importantly, we show that the output of XMAP can be integrated with single-cell datasets, which greatly improves the interpretation of putative causal variants in their cellular context at single-cell resolution.

Functional Adaptive Double-Sparsity Estimator for Functional Linear Regression with Application in Wearable Sensor Data Analysis

Dr. Xinyue Li

City University of Hong Kong

Wearable sensors have been increasingly used in health monitoring and early anomaly detection. Wearable device can collect objective and continuous information on physical activity and vital signs and have great potentials in studying the association with health outcomes. However, how to effectively analyze high-frequency multi-dimensional sensor data is challenging. In this talk, we propose a new Functional Adaptive Double-Sparsity Estimator (FadDoS) based on functional regularization of sparse group lasso with multiple functional predictors, which can achieve global sparsity via functional variable selection and local sparsity via zero-subinterval identification within coefficient functions. We prove that the FadDoS estimator converges at a bounded rate and satisfies the oracle property under mild conditions. Extensive simulation studies confirm the theoretical properties and exhibit excellent performances compared to existing approaches. Application to a Kinect sensor study that utilized an advanced motion sensing device tracking human multiple joint movements and conducted among community-dwelling elderly demonstrates how FadDoS can effectively characterize the detailed association between joint movements and physical health assessments. The proposed method is not only effective in Kinect sensor analysis but also applicable to broader fields, where multi-dimensional sensor signals are collected simultaneously, to expand the use of sensor devices in health studies.

Identifying predictors of glioma evolution from longitudinal sequencing

Dr. Quanhua Mu

The Hong Kong University of Science and Technology

Clonal evolution drives cancer progression and therapeutic resistance. Recent studies have revealed divergent longitudinal trajectories in gliomas, but early molecular features steering post-treatment cancer evolution remain unclear. Here we collected sequencing and clinical data of initial-recurrent tumor pairs from 544 adult diffuse gliomas and performed multivariate analysis to identify early molecular predictors of tumor evolution in three diffuse glioma subtypes. We found that CDKN2A deletion at initial diagnosis preceded tumor necrosis and microvascular proliferation that occur at later stages of IDH-mutant glioma. Ki67 expression at diagnosis was positively correlated with acquiring hypermutation at recurrence in the IDH-wildtype glioma. Strikingly, in all glioma subtypes, MYC gain or MYC-target activation at diagnosis was associated with treatment-induced hypermutation at recurrence. To predict glioma evolution, we constructed CELLO2 (Cancer EvoLution for LOngitudinal data version 2), a machine-learning model integrating features at diagnosis to forecast hypermutation and progression after treatment. CELLO2 successfully stratified patients into subgroups with distinct prognoses and identified a high-risk patient group featured by MYC gain with worse post-progression survival, from the low-grade IDH-mutant-non-codel subtype. We then performed chronic temozolomide-induction experiments in glioma cell lines and isogenic patient-derived gliomaspheres and demonstrated that MYC drives temozolomide resistance by promoting hypermutation. Mechanistically, we demonstrated that, by binding to open chromatin and transcriptionally active genomic regions, c-MYC increases the vulnerability of key mismatch repair genes to treatment-induced mutagenesis, thus triggering hypermutation. This study reveals early predictors of cancer evolution under therapy and provides a resource for precision oncology targeting cancer dynamics in diffuse gliomas.

Epigenetic cancer drug facilitates the proteogenomics discovery of immunotherapy targets

Dr. Stephen Li

The University of Hong Kong

In recent years, immunotherapy has emerged as a promising cancer treatment. By stimulating the host anti-tumor immune response, immunotherapy has improved prognosis and reduced the risk of relapsing. Nonetheless, the availability of neoantigen targets has limited the clinical efficacy of immunotherapy. The conventional method of identifying neoantigens, which focuses on tumor-specific SNVs and INDELs, suffers from patient specificity, dependence on mutation burden, and low immunogenicity. Such limitations underscore the importance of identifying neoantigens from other genetic elements.

Transposable Elements (TE) can serve as a potential source of neoantigens. Recent studies have identified TE-derived products as recurring targets for neoantigens. TE can trigger various immune responses, from producing dsRNA for viral mimicry to producing TE-derived peptide targets for cancer vaccines or CAR-T. However, TE expression is prone to repression by DNA methylation. One strategy to overcome TE epigenetic silencing is to reverse DNA methylation via cancer drugs like decitabine (DAC). However, the use of DAC to discover TE-derived antigens, particularly peptides that involve splicing between TE and exons, is currently rare.

This study aims to identify DAC-driven TE neoantigens across cancer types, using published and in-house transcriptomic data encompassing AML, glioblastoma, and colorectal cancers. The current findings have highlighted DAC-induced upregulation of TE subfamilies and TE-exon splice junctions, with varying subfamily distribution across different types of cancer. Compared with canonical transcripts, TE-derived antigens showed more frequent upregulation than exon-derived splicing isoforms. Translated peptide sequences from DAC-induced TE-derived antigens will enable subsequent proteogenomics search, HLA-affinity prediction, and experimental assays on immunogenicity. The results of this study will provide a novel list of neoantigen targets while highlighting the potential of using DAC in conjunction with immunotherapy across cancer types.

The TEMDA approach in drug design

Prof. Balaji Seetharaman

Unmasking LUAD's Vulnerabilities: Targeting AURKA-TPX2 Interaction for Innovative Therapeutics

Dr. MUKUNTHAN KS

Manipal Institute of Technology

Lung adenocarcinoma (LUAD) is one of the most prevalent and leading causes of cancer deaths globally, with limited diagnostic and clinically significant therapeutic targets. Identifying the genes and processes involved in developing and progressing¬†LUAD is crucial for developing effective targeted therapeutics and improving patient outcomes. Therefore, the study aimed to explore the RNA sequencing data of LUAD from The Cancer Genome Atlas (TCGA) and gene expression profile datasets involving GSE10072, GSE31210, and GSE32863 from the Gene Expression Omnibus (GEO) databases. The differential gene expression and the downstream analysis determined clinically significant biomarkers using a network-based approach. These therapeutic targets predominantly enriched the dysregulation of mitotic cell cycle regulation and revealed the co-overexpression of Aurora-A Kinase (AURKA) and Targeting Protein for Xklp2 (TPX2) with high survival risk in LUAD patients. The hydrophobic residues of the AURKA-TPX2 interaction were considered as the target site to block the autophosphorylation of AURKA during the mitotic cell cycle. The tyrosine kinase inhibitor (TKI) dacomitinib demonstrated the strong binding potential to hinder TPX2, shielding the AURKA destabilization. This in silico study lays the foundation for repurposing targeted therapeutic options to impede the protein-protein interactions (PPIs) in LUAD progression and aid in future translational investigations.

Integrative Epigenomics and Transcriptomics Analysis to Identify Signature Biomarkers of Lung Adenocarcinoma

Mr. Arnab Mukherjee

Manipal Institute of Technology

Lung carcinoma is one of the most prevalent and life-threatening cancers globally, with tobacco smoking being the most significant cause of lung cancer deaths. Lung adenocarcinoma (LUAD) accounts for approximately 80-85% of reported lung cancer cases and unfolds in a sequential multistage pattern, gradually developing genetic and epigenetic alterations. Alterations in DNA methylation at CpG sites are associated with smoking-induced lung cancer. Smoking-related epigenetic alterations are involved in the modulation of multiple biological pathways. Numerous tumors exhibit atypical methylation patterns, which can involve either increased (hypermethylation) or decreased (hypomethylation) addition of a methyl group to the cytosine. Demethylation of CpG sites is associated with the upregulation of oncogenes and genomic instability observed in multiple solid tumors, including lung cancer. However, hypermethylation is linked to the downregulation of the genes and silencing of tumor suppressors. Enhancers govern gene expression across great distances by looping DNA and offering distant regulatory regions closer to their target gene promoters. Therefore, we employed Illumina HM450k DNA methylation data of patients from The Cancer Genome Atlas (TCGA) to determine enhancers and link enhancer status with the expression of target genes to discover transcriptional targets using The Enhancer Linking by Methylation/Expression Relationship (ELMER) package of Bioconductor. In this study, we investigated a technique for predicting enhancer-target interactions by combining epigenomic and transcriptomic data from a substantial collection of primary tumor samples. This approach allowed us to identify target genes specifically regulated by enhancers with differential methylation patterns in LUAD and revealed the target genes of the differentially methylated sites and the enriched motifs modulating their expression in LUAD progression. The network-based approach aided in determining the hub genes playing a key role as central regulators of ribosome biogenesis, RNA processing, cell cycle regulation, and MMR pathways in LUAD pathogenesis.

Epigenetic clock CpGs are associated with chromatin and are enriched at enhancers

Mr. Yik Chai Charles Lau

The University of Hong Kong

DNA methylation-based epigenetic clocks are biomarkers that can determine chronological age, mortality and propensity for age-associated diseases (1). Since 2011 (2), epigenetic clocks have become known for being among the most robust and accurate ageing predictors, with the popular Horvath 2013 clock capable of accurately predicting age across multiple tissue types (3). The mechanisms underlying epigenetic clocks are still poorly understood as existing clocks have all been generated using microarrays which are biased towards promoter regions, hindering an accurate quantitative and qualitative assessment of the CpGs (cytosine-guanine motifs) that form such clocks and their likely causal relationships with ageing. With the current goals of finding and categorizing features underlying Horvath clock CpGs and CpGs that are significantly correlated with age, and constructing models to systematically identify and classify CpGs for age predicting capacity, a whole genome bisulphite (WGBS) dataset was built and used for epigenetic clock creation and CpG enrichment analysis. This dataset consisting of 617 samples resulted from a comprehensive mining of GEO (4) for all compatible WGBS samples that have age metadata, and as such is diverse not only in age range, but also disease status (cancer and neuropathology, etc.), tissue type and other traits, making it suitable both for the verification of the multitissue Horvath clock and the exploration of relationships between age-associated effects on the methylome and vice-versa. The analyses confirm that both the epigenetic clock methodology and Horvath,Äôs clock are applicable to the data and that it contains age-predicting information. Age acceleration effects, such as deceleration in females and schizophrenic patients are reproduced. Annotation of age-significant CpGs for genomic and chromatin features in the dataset demonstrates significant association with chromatin states, especially enhancers. The results imply that causes of ageing should be sought within chromatin and histone-based mechanisms, with a primary emphasis on enhancer-associated pathways.

1. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat Rev Genet. 2018 Apr 11;19.
2. Bocklandt S, Lin W, Sehl ME, S°nchez FJ, Sinsheimer JS, Horvath S, et al. Epigenetic Predictor of Age. PLOS ONE. 2011 Jun 22;6(6):e14821.
3. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013 Dec 10;14(10):3156.
4. NCBI GEO: archive for functional genomics data sets,Äîupdate | Nucleic Acids Research | Oxford Academic [Internet]. [cited 2023 Sep 15].
Available from: https://academic.oup.com/nar/article/41/D1/D991/1067995?login=false

Copy number state minimum evolutionary path algorithm for inferring whole genome doublings

Mr. Wai Tung Jonathan Tse

The University of Hong Kong

The study of DNA copy numbers is essential to many areas of genomics including research into population diversity and evolutionary processes of diseases. Various copy number inference tools have been made to infer copy number states of tumor samples using sequencing data. However, some tools are limited by the use of non-identifiable models which give ambiguous solutions, potentially giving solutions with large discrepancies in ploidy which suggest the tumor sample to be both diploid or tetraploid. We introduce an algorithm which infers whole genome doubling status to choose an optimal solution between multiple solutions based on minimum steps in the tumor evolutionary path using copy number states.

Transcriptomic analysis of cisplatin resistance in ovarian cancer

Ms. Noel Yue

The University of Hong Kong

High-grade serous ovarian cancer (HGSOC) is the most aggressive type of ovarian cancer; standard treatment of HGSOC consists of debulking followed by platinum-based chemotherapy. Meanwhile, 34% of patients will develop resistance during the treatment. We seek to determine the transcriptomic dysregulation that lead to platinum resistance in high grade serous ovarian cancer with the gene expression profile of the HGSOC samples. Specifically, we would like to encounter cellular heterogeneity such that we will be performing transcriptome deconvolution computationally. However, we confronted common bioinformatics obstacles, including mismatches in gene expression platforms, batch effects, missing normal references for deconvolution, and suboptimal results in publicly accessible algorithms. To address these challenges, we implemented the following strategies: conduct quantile normalization on mismatch-platform gene expression data, referencing the distribution of the TCGA-ovarian cancer RNA-seq dataset, such that all the datasets follow a similar distribution; batch effects are mitigated through the application of ComBat; pseudo normal bulk RNA-seq data is generated using normal single cell RNA-seq data, serving as a substitute for the absent of normal references for deconvolution; conduct evaluation of multiple publicly available algorithms and made modifications to the algorithm to optimize its performance. Successfully addressing these challenges will lead us to a robust dataset that allows us to conduct in-depth analyses, ultimately leading us to determine the transcriptomic dysregulation that lead to platinum resistance in HGSOC patients.

Comprehensive genomic characterization of EBV-associated carcinomas

Ms. Yucan Chen

The University of Hong Kong

Epstein-Barr virus (EBV)-associated malignancies display unique clinicopathologic features and molecular alterations, including genetic and epigenetic changes. However, sufficient molecular evidence supporting the interaction between EBV and the host genome is still lacking. We aimed to investigate the host genomic alterations and identify features associated with EBV infection. We downloaded whole-genome sequencing data and RNA-seq data of gastric carcinoma (GC) and Burkitt lymphoma (BL) samples, both annotated with EBV status, and ChIP-seq data of GM12878 and B cells. We investigated the mutational profiles and mutational signatures of EBV-infected and uninfected groups and conducted replication timing, replication strand bias, and transcriptional bias analyses on samples labeled EBV-positive and EBV-negative. We then explore the regional differences in mutation distribution across the genome and the changes of chromatin modifications upon EBV infection. Finally, we performed differential expression and gene set enrichment analyses (GSEA) for hints from transcriptional data. We found the mutations in EBV-infected samples were featured with T>G in the TTT trinucleotide context and TTTTN pentanucleotide context. The mutational signatures showed different SBS28 and SBS17 exposure associated with EBV infection, more prevalent SBS28, and less SBS17 in EBV-infected samples in both GC and BL samples. The trend of replication timing curves of T>G and T>A mutations were higher in early replication timing regions in EBV-infected samples. We found the distinct mutation distribution across genic regions and different chromatin states, and the modification signals in GM12878 versus B cells correlated differentially with mutations between EBV-associated GCs (EBVaGCs) and controls, flatter in EBVaGCs. The GSEA result unraveled the altered expression of genes involved in DNA repair-associated pathways enriched in EBV-positive samples. Our findings characterized EBV-associated genomic features, indicating the distinct etiology behind EBV-associated carcinomas shedding light on the different therapeutic strategies for cancers and their EBV-associated subtypes.

Tuesday, 17 Oct 2023

Exploring Cell Fate and Memory through Single-Cell Multi-Omic Lineage Tracing

Dr. Shou-Wen Wang

Westlake University

Cellular lineage histories and their molecular states encode fundamental principles of tissue development and homeostasis. Current lineage-recording mouse models have insufficient barcode diversity and single-cell lineage coverage for profiling tissues composed of millions of cells. Here, we developed DARLIN, an inducible Cas9 barcoding mouse line that utilizes terminal deoxynucleotidyl transferase (TdT) and 30 CRISPR target sites. DARLIN is inducible, generates massive lineage barcodes across tissues, and enables detection of edited barcodes in 70% of profiled single cells. Using DARLIN, we examined fate bias within developing hematopoietic stem cells (HSCs) and revealed unique features of HSC migration. Additionally, we established a protocol for joint transcriptomic and epigenomic single-cell measurements with DARLIN and found that cellular clonal memory is associated with genome-wide DNA methylation rather than gene expression or chromatin accessibility. DARLIN will enable high- resolution study of lineage relationships and their molecular signatures in diverse tissues and physiological contexts.

Covidscope: An atlas scale COVID,Äì19 resource for single-cell meta analysis at sample and cellular levels

Ms. Danqing (Angela) Yin

The University of Hong Kong

To triangulate the cellular response and processes from the sample level (e.g. individual patient) in the COVID-19 pandemic, single-cell technologies have been widely exploited. The abundance of rising single-cell data atlases poses an unprecedented barrier for researchers to efficiently accessing, organizing, sharing and interpreting massively available information. Here, we provide a carefully integrated atlas-scale resource Covidscope offering a uniform method by utilizing scMerge, scClassify and scFeature to accelerate efficient downstream meta-analysis for COVID-19 single-cell data. We acquired 5 million single PBMC cells from 20 studies consisting of 1000 donors across the globe and made it available online through the Covidscope data portal via https://covidsc.d24h.hk. The ability to search and visualize large single cell expression data matrices in real-time based on users, Äô queries up to 4x10Î4 combinations using patient or cell-level characteristics makes Covidscope distinctive to other existing portals. We believe that Covidscope will have a profound impact on the scientific community, providing a first encyclopedic all-in-one resource for COVID-19 researchers to unravel the hidden and complex insights at the atlas scale.

Temporal preserved representation learning from time-series single-cell transcriptomics data

Ms. Yijun Liu

The University of Hong Kong

Time-series designs of single-cell RNA-seq experiments have been broadly utilised to study various biological processes, including embryonic development, cell differentiation, and disease progression. However, it remains a grand challenge to effectively discover dynamical modules and the according gene regulation from such temporally structured data, largely due to the tight coupling of time stamps and experiment batches and the complexity of cell states. In this study, we introduce temporalVAE, a computational method that leverages a variational auto-encoder to infer the biological time of cells from a compressed latent space, and simultaneously conducts batch correction via an adversarial learning strategy. We illustrate the effectiveness and strengths of temporalVAE on experimental data sets from multiple biological systems, including fine-resolution mouse embryonic development and human endometrium across the menstrual cycle. Notably, our results highlight a strong positive correlation between temporalVAE's prediction capabilities and the time-sensitivity of distinct cell types. As the diversity of single-cell data types continues to expand, we envision that temporalVAE can further integrate different types of data and obtain continuous cellular time, which will facilitate the elucidation of gene regulatory processes at an unprecedented resolution.

Identification of potential neural precursor cell populations using atlas-level single-cell data

Prof. Zhichao Miao

Guangzhou National Laboratory

The extent to which neurogenesis occurs in adult primates remains controversial. In this study, using an optimized single-cell RNA sequencing pipeline, we profiled 207,785 cells from the adult macaque hippocampus and identified 34 cell populations comprising all major hippocampal cell types. Analysis of their gene expression, specification trajectories and gene regulatory networks revealed the presence of all key neurogenic precursor cell populations, including a heterogeneous pool of radial glia-like cells (RGLs), intermediate progenitor cells (IPCs) and neuroblasts. We identified HMGB2 as a novel IPC marker. Comparison with mouse single-cell transcriptomic data revealed differences in neurogenic processes between species. We confirmed that neurogenesis is recapitulated in ex vivo neurosphere cultures from adult primates, further supporting the existence of neural precursor cells (NPCs) that are able to proliferate and differentiate. Our large-scale dataset provides a comprehensive adult neurogenesis atlas for primates

Annotating and discovering cell types from single cell RNA-seq data using machine learning

Dr. Yu Li

The Chinese University of Hong Kong

JSNMF enables effective and accurate integrative analysis of single-cell multi-omics data

Prof. Zhixiang Lin

The Chinese University of Hong Kong

The single-cell multiomics technologies provide an unprecedented opportunity to study the cellular heterogeneity from different layers of transcriptional regulation. However, the datasets generated from these technologies tend to have high levels of noise, making data analysis challenging. Here, we propose jointly semi-orthogonal nonnegative matrix factorization (JSNMF), which is a versatile toolkit for the integrative analysis of transcriptomic and epigenomic data profiled from the same cell. JSNMF enables data visualization and clustering of the cells and also facilitates downstream analysis, including the characterization of markers and functional pathway enrichment analysis. The core of JSNMF is an unsupervised method based on JSNMF, where it assumes different latent variables for the two molecular modalities, and integrates the information of transcriptomic and epigenomic data with consensus graph fusion, which better tackles the distinct characteristics and levels of noise across different molecular modalities in single-cell multiomics data. We applied JSNMF to single-cell multiomics datasets from different tissues and different technologies. The results demonstrate the superior performance of JSNMF in clustering and data visualization of the cells. JSNMF also allows joint analysis of multiple single-cell multiomics experiments and single-cell multiomics data with more than two modalities profiled on the same cell. JSNMF also provides rich biological insight on the markers, cell-type-specific region-gene associations and the functions of the identified cell subpopulation.

Differential inference for single-cell RNA-sequencing data

Dr. Fangda Song

The Chinese University of Hong Kong, Shenzhen

Single-cell RNA-sequencing (scRNA-seq) experiments are becoming more and more complicated with multiple biological conditions or treatments. However, guidelines on experimental designs and rigorous statistical methods for a comparative scRNA-seq study that collects data from multiple conditions are still lacking. Here, we derive the conditions for a valid design for a comparative scRNA-seq study so that the batch effects, cell type effects, and condition effects can be decomposed. We develop an interpretable Bayesian hierarchical model, DIFferential Inference for Single-cell RNA-sequencing Data (DIFseq), to rigorously quantify the condition effects on both cellular compositions and cell-type-specific gene expression levels for scRNA-seq data. DIFseq substantially outperforms the state-of-the-art methods in terms of the accuracy of cell type clustering, identification of differential cellular abundance, and detection of intrinsic genes,Äîgenes that are differentially expressed (DE) between cell types at the baseline,Äîand cell-type-specific DE genes,Äîgenes that are DE between conditions for both simulated and real data.

Spatiotemporal transcriptomic and genomic analysis reveals waves of hematopoiesis in human embryonic organoids

Ms. Yiming Chao

1. The University of Hong Kong, HKSAR 2. Centre for Translational Stem Cell Biology, HKSAR 3. Wellcome-MRC Cambridge Stem Cell Institute, UK

Hematopoiesis is a dynamic process that encompasses multiple waves at different sites throughout the lifespan. The first extra-embryonic hematopoiesis wave arises in the yolk sac during the early embryonic stage. It remains unknown how extra-embryonic tissues support the hematopoiesis in the blood island during embryo development. In our recent work, we model human embryonic development and hematopoiesis with stem-cell-derived human embryonic organoid (HEMO) and use advanced sequencing technology including single-cell RNA sequencing and spatial transcriptomics to define the gene expression and cellular interactions. Single-cell resolution spatial transcriptomics defined the yolk sac erythro-megakaryopoietic niche. Vitronectin-integrin signaling remarked the yolk sac niche in HEMO and human fetal samples. Moreover, we exploit the mitochondrial SNVs as an innate cellular barcode to discriminate different hematopoietic waves that correspond to known human embryonic hematopoiesis in the yolk sac. Our study advances the spatiotemporal transcriptomic and genomic analysis of human embryonic development in stem-cell-derived organoids. We emphasize the regulations of extra-embryonic tissues during development and highlight the dynamic hematopoiesis modeled by the organoids.

Unveiling the Complexity: The Current State of Single-Cell Benchmarking

Dr. Yue Cao

University of Sydney

With the rapid development of computational methods for single-cell sequencing data, benchmarking provides a valuable solution for guiding method selection. As the number of single-cell benchmarking studies has surged over the years, it is now timely to assess the current state of the field and to understand best practices behind high-quality benchmarks. To this end, we conducted a systematic literature search and comprehensively assessed a total of 245 papers. Among these, we examined all 95 pure benchmarking papers from the literature search and selected an additional 150 papers across multiple areas of method development that include a benchmarking component in their evaluation. We illustrated the characteristic of benchmarking across ten broad complementary categories covering over 70 variables. As expected, our analysis revealed a differing focus in evaluation between pure benchmarking papers and method development papers. In addition, our analysis highlights a number of challenges such as dataset dependent method performance and the emerging issue related to effective benchmarking. This work lays the foundation for the need of collective effort from the community to establish a repository for benchmarking datasets as well as a ‚Äúliving benchmark‚Äù platform that facilitate methods comparison for single-cell method developers.

Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis

Ms. Yixuan Wang

The Chinese University of Hong Kong

Single-cell RNA-sequencing has become a powerful tool to study biologically significant characteristics at explicitly high resolution. However, its application on emerging data is currently limited by its intrinsic techniques. Here, we introduce Tissue-AdaPtive autoEncoder (TAPE), a deep learning method connecting bulk RNA-seq and single-cell RNA-seq to achieve precise deconvolution in a short time. By constructing an interpretable decoder and training under a unique scheme, TAPE can predict cell-type fractions and cell-type-specific gene expression tissue-adaptively. Compared with popular methods on several datasets, TAPE has a better overall performance and comparable accuracy at cell type level. Additionally, it is more robust among different cell types, faster, and sensitive to provide biologically meaningful predictions. Moreover, through the analysis of clinical data, TAPE shows its ability to predict cell-type-specific gene expression profiles with biological significance. We believe that TAPE will enable and accelerate the precise analysis of high-throughput clinical data in a wide range.

Deciphering sequence-informed regulatory patterns from single-cell multi-omics

Ms. Fangxin Cai

The University of Hong Kong

Single cell multi-omics presents parallel and asynchronous perspectives to biological processes. Spanning transcription and RNA splicing, paired scRNA-seq and scATAC-seq data provide complementary views to gene regulation. To disentangle their relation, previous works categorize transcription patterns by the coupling between ATAC and RNA data, which corresponds to distinct regulatory mechanisms. These patterns can be difficult to interpret especially when based on reduced data dimensions.
We present a sequence-informed autoencoder framework that may extract biologically meaningful cell dimensions. On spliced and unspliced RNA data, latent cell dimensions retain splicing dynamics, from which RNA velocity can be inferred. For more asynchronous ATAC and RNA data, cell embeddings are aligned by sequence elements such that similar TSS and peak sequences are activated by the same dimension. Cross modal prediction reveals coupled and uncoupled chromatin and expression dynamics. Globally, ATAC states align to later RNA states known as transcription priming. The relation between latent RNA velocity and ATAC dimensions remains to be modeled.

Con-AAE: contrastive cycle adversarial autoencoders for single-cell multi-omics alignment and integration

Mr. Xuesong Wang

The Chinese University of Hong Kong

Motivation: We have entered the multi-omics era and can measure cells from different aspects. Hence, we can get a more comprehensive view by integrating or matching data from different spaces corresponding to the same object. However, it is particularly challenging in the single-cell multi-omics scenario because such data are very sparse with extremely high dimensions. Though some techniques can be used to measure scATAC-seq and scRNA-seq simul- taneously, the data are usually highly noisy due to the limitations of the experimental environment.
Results: To promote single-cell multi-omics research, we overcome the above challenges, proposing a novel framework, contrastive cycle adversarial autoencoders, which can align and integrate single-cell RNA-seq data and single-cell ATAC- seq data. Con-AAE can efficiently map the above data with high sparsity and noise from different spaces to a coordinated subspace, where alignment and integration tasks can be easier. We demonstrate its advantages on several datasets.

Biologically-informed self-supervised learning for segmentation of subcellular spatial transcriptomics data

Dr. Xiaohang Fu

The University of Sydney

Recent advances in subcellular imaging transcriptomics platforms have enabled high-resolution spatial mapping of gene expression, while also introducing significant analytical challenges in accurately identifying cells and assigning transcripts. Existing methods grapple with cell segmentation, frequently leading to fragmented cells or oversized cells that capture contaminated expression. To this end, we present BIDCell, a self-supervised deep learning-based framework with biologically-informed loss functions that learn relationships between spatially resolved gene expression and cell morphology. BIDCell incorporates cell-type data, including single-cell transcriptomics data from public repositories, with cell morphology information. Using a comprehensive evaluation framework consisting of metrics in five complementary categories for cell segmentation performance, we demonstrate that BIDCell outperforms other state-of-the-art methods according to many metrics across a variety of tissue types and technology platforms. Our findings underscore the potential of BIDCell to significantly enhance single-cell spatial expression analyses, enabling great potential in biological discovery.

Benchmarking translational potential of spatial transcriptomics imputation from histology images

Ms. Chuhan Wang

The University of Sydney

Spatial transcriptomics (ST) enables the quantification of gene expression within specific spatial coordinates, offering crucial insights into tumour heterogeneity of tissues at high resolution. In light of this, the feasibility of predicting spatial gene expression (GE) from conveniently obtainable and cost-effective haematoxylin-and-eosin-stained (H&E) histology images gains significance. To this end, we conducted a comprehensive benchmarking study encompassing 6 developed methods designed to predict spatial GE from H&E histology images. These methods were reproduced and evaluated using HER2-positive breast tumour (her2st) and human cutaneous squamous cell carcinoma (cSCC) datasets through a 4-fold cross-validation framework, followed by external validation using The Cancer Genome Atlas (TCGA-BRCA) data. Our evaluation demonstrates diverse metrics, including the performance measures of predicted and ground truth GE values, performance variations across different image regions and levels of gene expression, the translational potential indicated by survival analysis, and the overall usability of the methods.

Systematic comparison of sequencing-based spatial transcriptomic methods

Dr. Luyi Tian

Guangzhou Laboratory

Sequencing-based spatial transcriptomic techniques have undergone rapid development in recent years, enabling unbiased, transcriptome-scale measurements of spatial gene expression. However, these methods have yet to be systematically benchmarked, and the considerable variability across technologies and datasets complicates the establishment of evaluation standards. To address this, we have developed a set of reference tissues with well-defined histological structures, utilizing them to generate data and assess six distinct technologies. Despite variations in resolution, capture efficiency, and spatial precision, spatial transcriptomic data exhibit characteristics distinct from single-cell RNAseq data, such as enhanced capabilities for capturing certain genes, along with more pronounced blood contamination. This study aims not only to guide biologists in method selection but also to build a consensus on evaluation criteria, establish a framework for future benchmarking, and provide gold standards for the assessment of computational tools

CellContrast: Reconstructing Spatial Relationships in Single-Cell RNA Sequencing Data via Deep Contrastive Learning

Dr. Shumin Li

The University of Hong Kong

A vast amount of single-cell RNA-seq (SC) data has been accumulated via various studies and consortiums, but the lack of spatial information limits its analysis dissection of complex biological activities. To bridge this gap, we introduce cellContrast, a computational method for reconstructing spatial relationships among SC cells from spatial transcriptomics (ST) reference. By adopting a deep contrastive learning framework and training with ST data, cellContrast projects gene expressions into a hidden space where proximate cells share similar representation values. We performed extensive benchmarking on diverse platforms, including SeqFISH, Stereo-Seq, and 10X Visium, on mouse embryo and human breast cells. The results reveal that cellContrast substantially outperforms other related methods, facilitating accurate spatial reconstruction of SC. We further demonstrate cellContrast's utility by applying it to cell-type co-localization and cell-cell communication analysis with real-world SC samples, proving the recovered cell locations empower novel discoveries and mitigate potential false positives.

Decoding Fine-grained Cellular Architectures from Histology Images

Mr. Weiqin Zhao

The University of Hong Kong

The cellular architecture of tissues, where distinct cell types are organized in space, underlies cell‚Äìcell communication, organ function and pathology. Emerging spatial transcriptomics technologies provide opportunities to map resident cell types and cell signaling in situ in a scalable manner. The key principle is to integrate spatial RNA-seq with the reference transcriptome signatures of cell types obtained from coupled scRNA-seq profiles. However, spatial transcriptome data has not been utilized in large-scale studies due to the expensive costs. Relatively, histology images are much cheaper and easier to obtain and are routinely generated in clinics. Nevertheless, how to identify fine-grained cell type spatial variations and cell-cell communication directly from histology images remains to be an important problem. To address this issue, we present Hist2Cell, a Graph-Transformer framework to resolve fine-grained cell types directly from histology images and create cellular maps of diverse tissues. Hist2Cell was trained on human breast cancer and human lung spatial transcriptome datasets with spot-level cell abundance estimations. For held-out test patients, Hist2Cell can predict the spatial variation in the localization of fine-grained cell types - at a resolution of around 100Œ°m. Moreover, it can reveal accurate cell-cell communications by feeding predictions to SpatialDM toolbox, figuring out the prioritization of interaction features as well as the identification of interaction spots in the spatial context. More importantly, Hist2Cell can also provide above analysis under higher resolution than spatial transcriptome data in a scalable manner. As independent external tests, Hist2Cell accurately predicts spatial cell localization and cell-cell communications without any modification or tuning. This suggests that it robustly generalizes to new samples. Hist2Cell is more accurate than methods that predicting cell abundance solely from single histology image spot or limited neighboring context, as it leverages both local and global correlations of histology image spots with a Graph-Transformer architecture.

Towards the Visual-Transcriptomics Foundation Model for Spatial Transcriptomics Analysis

Ms. Zhuo Liang

The University of Hong Kong

Histopathology, with its remarkable advancements in tumor analysis and morphology, is becoming indispensable for studying correlations in disease pathology. Spatial transcriptomics, with its high spatial resolution, has emerged as a powerful tool for studying gene expression, enabling us to investigate the relationship between cellular neighborhoods and gene patterns. Existing works have revealed the connection between image morphology and genomics through analyzing images and gene expressions independently, while neglecting the potential advantages of aligning these two modalities. We argue that aligning modalities can open up new possibilities for the inference of genomics from morphology, thereby advancing our understanding of spatial transcriptomics.
In this study, we propose a visual-transcriptomics foundation model named CONTH (CONtrastive learning from Transcriptomics for Histopathology) to align image and gene representations effectively in latent space. This process, referred to as the contrastive task, utilizes cross-entropy loss to evaluate the dissimilarity between two distributions achieved by projecting image and gene expression into low-dimensional space. In addition to assessing the accuracy of the alignment process, the preservation of vital information is crucial. To address this concern, we introduce a generative task. After obtaining image representations, we generate corresponding gene predictions based on these representations. Subsequently, we calculate the image-to-gene reconstruction loss by comparing the predictions with the ground truth. Similarly, we reconstruct images using gene representations and compute the gene-to-image reconstruction loss. These three losses work synergistically to guide the learning process of CONTH.
We demonstrate the effectiveness of our framework via retrieval which serves as an indicator for the generalized consistency between two modalities. On 23 cross-tissue 10xGenomics datasets, we currently achieved a recall of 34%, compared to 20% by randomized selection, proving the superiority of alignment.
We foresee our model will pave the way for a more comprehensive knowledge of the intricate interplay between morphology and genomics.

Multi-task Deep Learning Network for Embryo Quality Assessment

Ms. Lu Yu

The University of Hong Kong

Blastocyst selection is one of the most significant challenges in the in-vitro fertilization (IVF) process because the current assessment is largely based on morphology evaluation by embryologists which brings more subjectivity and less precision. Artificial intelligence can play a role in overcoming the limitations of the manual assessment system by inputting either observation of single static morphology images or time-lapse video sequences.
Here, we design a multi-task deep learning network to assist blastocyst selection using static images. The first branch can automatically grading the fresh blastocyst based on the morphology of expansion (1-6), inner cell mass (A,B,C), and trophectoderm (A,B,C) respectively, which is trained on a gold-standard public dataset. It performs well on our clinical data by using a small size of our clinical data to finetune, which shows good potential of overcoming the bias caused by subjectivity in different embryologists and IVF centers. The second branch which can assess the embryo viability, is trained not only on fresh blastocyst dataset but also by creatively borrowing richer information from post-thawing embryo dataset as frozen embryo transfer has been proved to lead to better live-birth outcomes. YOLO-v3 model is deployed to locate and crop the embryo from the static image to overcome the effects of background noise, which can also extends in time-lapse images. Various preprocessing and online and offline augmentation techniques are applied to balance and flourish the dataset. A series of backbone model pretrained on Imagenet, such as ResNet and Vision Transformer, are combined with grid-search and bayesian optimization to establish optimal model which results in the most accurate predictions.

Experiment-guided clustering to unravel complex drug-diet interactions in nutri-omics data

Prof. Jean Yang

University of Sydney

Easier and more comprehensive access to large omics data has revolutionised not just biological research but also nutrition research. The emerging field of nutriomics is enhancing our understanding of the interactions between nutrients and the host, offering novel insights into personalised nutrition and deepening our understanding of nutrition on health. In this talk, we will discuss how recent approaches in statistical modelling and machine learning methods are addressing data science challenges presented by both experimental and profiling studies in nutriomics. Specifically, we will present eNODAL, an experiment-guided clustering method that takes advantage of both ANOVA-type analyses and unsupervised learning methods in order to extract maximum information from experimental nutriomics studies. Applying eNODAL on a mouse proteomics study, we demonstrate eNODAL ability to identify clusters of proteins that are affected by interactions between nutrition information and drug intake.

Cell-type aware CRISPR editing outcomes prediction

Mr. Weizhong Zheng

The University of Hong Kong

The CRISPR-Cas system has revolutionized gene editing by enabling precise and efficient modifications, encompassing a wide range of applications, including gene activity modulation and the generation of model organisms. However, the inherent variability in CRISPR-induced DNA repair processes, influenced by sequence context and cell lines, poses a significant challenge in accurately predicting editing outcomes before conducting experiments. In this study, we introduce inDecay, a flexible system for predicting CRISPR editing outcomes. InDecay predicts the probability of a large number of editing outcomes derived from the target sequence, and incorporates cell-type-specific repair preferences through a multi-stage design. By utilizing informative and parameter-efficient features for each indel event, inDecay outperforms existing prediction methods across multiple evaluation measures, including identifying the most common DNA alterations and the overall occurrence of frameshift mutations. Furthermore, inDecay exhibits superior few-shot learning capabilities when transferred to novel cell types. The unique features of inDecay establish it as a valuable tool for planning CRISPR editing experiments that require precise control over editing outcomes or for experiments conducted in previously unexplored cellular environments.

Integrating long-read RNA sequencing improves locus-specific quantification of transposable element expression

Ms. Sojung Lee

The University of Hong Kong

Endogenous transposable elements (TEs) are implicated in human diseases due to their propensity to compromise genome integrity. Although short-read sequencing is now frequently used to examine TE expression, the highly repetitive nature of TEs limits their accurate quantification at the locus-specific level. We have developed LocusMasterTE, an improved method that integrates information from long-read RNA sequencing to enhance TE quantification. The fractional transcript per million (TPM) from long reads serves as a prior distribution during the Expectation-Maximization (EM) model in short-read TE quantification, thereby enabling the reassignment of multi-mapped reads to correct expression values. Using simulated short-reads, our results indicate that LocusMasterTE outperforms existing quantitative approaches and is especially favorable for quantifying evolutionarily younger TEs. Using matched cell line RNA-seq data, we further demonstrate improved locus-specific TE quantification by LocusMasterTE with stronger enrichment in active, and depletion at repressive, histone marks. Finally, by integrating colorectal cancer cell line long-read sequencing data with short read RNA-seq data from The Cancer Genome Atlas colorectal cancer cohort, we demonstrate LocusMasterTE,Äôs ability to identify survival-related TEs and uncover new expression associations between locus-specific TEs and neighboring genes. By providing more accurate quantification, LocusMasterTE offers the potential to reveal novel functions of TE transcripts.

Improving long-read scRNA-seq data analysis with FLAMES

Mr. Changqing Wang

Walter And Eliza Hall Institute Of Medical Research

Long-read sequencing enables accurate determination of novel isoforms, yet pipelines dedicated for such analyses are limited. The previous FLAMES pipeline covers from data preprocessing all the way to differential analyses. We recently updated the FLAMES pipeline to integrate with standard Bioconductor containers , adding support multi-sample datasets, support for using external packages to replace certain steps and additional data visualisation functions. The updated pipeline is now modularised, with each step using standard file formats as inputs and outputs, allowing more external packages to be included in FLAMES. We hope to integrate the software ecology of this field in FLAMES to provide a convenient one-stop-shop for single cell long-read RNA-seq data processing and analysis, and hence promoting further development in both the software side and the application side.

Biodiversity Genomics - From evolution to conservation policy

Prof. Jerome Hui

The Chinese University of Hong Kong

Lipidomic signatures of resilience to coronary artery disease learnt from chimpanzees

Mr. Andy Tran

University of Sydney

Coronary artery disease continues to be one of the leading causes of death globally. Despite sharing over 96% of our DNA sequence, our closest animal relative, chimpanzees, have remarkable differences in susceptibility to coronary artery disease, where there has only been a single report of death from myocardial infarction, despite thousands of necropsies being performed. In this study, blood plasma has been collected from 20 chimpanzees where over 800 lipid species were isolated and analysed. The same protocol was performed on the BioHEART-CT Discovery cohort, with approximately 1000 patients. We investigate the differences between humans and chimpanzees to reveal evolutionary differences in their lipidomic profiles, and discover lipidomic signatures of resilience to coronary artery disease.

TDbasedUFE and TDbasedUFEadv: bioconductor packages to perform tensor decomposition based unsupervised feature extraction

Prof. Y-h. Taguchi

Chuo University

Motivation: Tensor decomposition (TD) based unsupervised feature extraction (FE) [1] has proven effective for a wide range of bioinformatics applications ranging from biomarker identification to the identification of disease-causing genes and drug repositioning. It can process any kind of omics data set formatted as tensor. For example, gene expression of various tissues of many human subjects can be analyzed to select tissue specific differentially expressed genes. Alternatively, by selecting drugs that target selected genes, we can also perform drug repositioning. We can also make use of this method to seek biomarkers of diseases. However, TD-based unsupervised FE failed to gain widespread acceptance due to the lack of user-friendly tools for non-experts.
Results: We developed two bioconductor packages -TDbasedUFE and TDbasedUFEadv- enable researchers unfamiliar with TD to utilize TD-based unsupervised FE [2]. TDbasedUFE facilitates the identification of differentially expressed genes and multiomics analysis. TDbasedUFE was found to outperform two state-of-the-art methods, DESeq2 and DIABLO. TDbasedUFEadv can deal with various specialized cases, including vertical (features) and horizontal (samples) integration, reduction of memory with summation of vertical or horizontal direction which can be later recovered and the usage of singular value decomposition as preprocessing. All of these feature allow users to analyse wide range of bioinformatics analysis. Availability and implementation TDbasedUFE and TDbasedUFEadv are freely available as R/Bioconductor packages, which can be accessed at https://bioconductor.org/packages/TDbasedUFE and https://bioconductor.org/packages/TDbasedUFEadv, respectively.
References: [1] https://doi.org/10.1007/978-3-030-22456-1
[2] https://www.frontiersin.org/articles/10.3389/frai.2023.1237542/abstract