Multidisciplinary Research Program in Medicine Project*: Using digital pathology with spatially resolved gene expression for biomarker discovery*

Hypothesis or Research Question(s):

Questions include:

1) Do regions on tissue images that are predictive of disease also contain gene-expression signals that are predictive disease?

2) Can imaging data be used to predict spatial gene expression data?

PROJECT BACKGROUND & SUMMARY

Deep learning has made breakthroughs in image classification surpassing the performance of trained professionals, for example at diagnosing skin cancer. These advances have led to breakthroughs in digital pathology where deep learning models are applied to images of tissues. Deep learning models trained using H&E slides have been shown to predict tumour genotype, mutations, gene expression, and clinical phenotypes such as heart failure.

The GeoMx Digital Spatial Profiler allows for spatial gene expression profiling after staining tissue slides using up to four morphology markers. However, regions of interest must be selected manually by the user for spatial profiling of gene expression. This present two challenges: 1) selected regions may not contain the most predictive gene signals that discriminates between disease conditions and 2) gene expression cannot be spatially resolved for the entire tissue slide.

The first goal of this project is to train a convolutional neural network (CNN) using tissue imaging data to predict diabetic kidney disease (DKD). Images will be divided into patches/tiles prior to fine-tuning a pre-trained ImageNet model for image classification. Visualizations such as saliency maps will be used to depict regions on the images that are the most predictive of disease. Gene expression will be compared for predictive and non-predictive regions comparing healthy and DKD samples. A secondary goal of this project is to use the imaging data to predict spatial gene expression data using a Spatial Organ Atlas. An autoencoder architecture will be used to predict the 18,000+ protein-coding genes using the imaging data from five organ types (colon, kidney, brain, lymph node and pancreas). Only a limited number of genes are expected to be predicted accurately, however this will be useful for studies where only imaging data is present and gene expression can simply be estimated.

The outcome of this work will be a python with functions to process, analyze and visualize imaging and gene-expression data coming from the GeoMx digital profiler. Users will be able to use simple functions to build complex models using the imaging data to predict disease groups. Other functions to assess model performance, and generate saliency maps will also be provided. The user can then choose to profile these regions for gene expression profiling to compare between disease groups. This software will also provide a function to predict spatial gene expression for the entire tissue section given an input image file. This tool will benefit users of the GeoMx platform and may be extended to other spatial technologies.

BENEFIT TO THE STUDENTS

Student roles

TS1: Curation Scientist TS1 will be responsible for conducting a literature review for journal articles using the GeoMx Digital Profiler and for compiling all publicly available datasets with image and gene

expression data. TS1 will take the lead in writing up a short article (application note) describing the knowledge-to-date and objectives.

TS2: Developer 1 TS2 will be responsible for running models to predict disease status from imaging data alone and imaging + gene expression data. TS2 will train CNNs for disease prediction extending previous work (https://github.com/mahmoodlab/TOAD). Helper functions based on the Captum Python library will be used to create saliency maps.

TS3: Developer 2 TS3 will be responsible for re-running previous analyses from published journal articles which have employed similar models to predict spatial gene expression from imaging data in the context of cancer (e.g. https://github.com/owkin/HE2RNA_code). These scripts will be modified to be compatible with the files from the GeoMx Digital Spatial Profiler. TS3 will develop a figure that depicts the error rate per gene describing which genes are easier/harder to predict.

Benefits to Students

1. Students will learn how to implement each step of the scientific method. Students will be allowed to explore other related-topics based on their research of the topic areas.

2. Students will learn how to employ an agile methodology in the research setting, making uses of project management tools for study planning, milestones and wrap-up. The project will be setup on the PI's GitHub lab page and all students will be given training on how to use GitHub (social platform to store code, collaborate on projects). The use of version control (git) and code collaboration (GitHub) are gold-standard practices that sought-after skills in both industry and academia.

3. Students will also learn reproducible research best practices such that the analyses including running scripts, figure/table generation, and manuscript document can be re-run/generated at any given time.

4. Students developing the code will use industry-standard libraries such as PyTorch for training deep learning algorithms. Developer students will have access to Collab Pro such that they will have access to accelerated computing for training the deep learning models.

5. Students will have the opportunity to present their work orally and via a poster presentation during the Summary Research Day hosted in August of 2022 at the Centre for Heart and Lung Innovation. This is fantastic opportunity to practice oral skills in front of a large audience and receive feedback on their work.

6. If not virtual due to the pandemic, students will have desk space at the Centre for Heart Lung Innovation where they will have access to weekly research in progress seminars, Friday speaker series, Knowledge Translation workshops, mental health awareness meetups, amongst the current 261 scientists and staff.