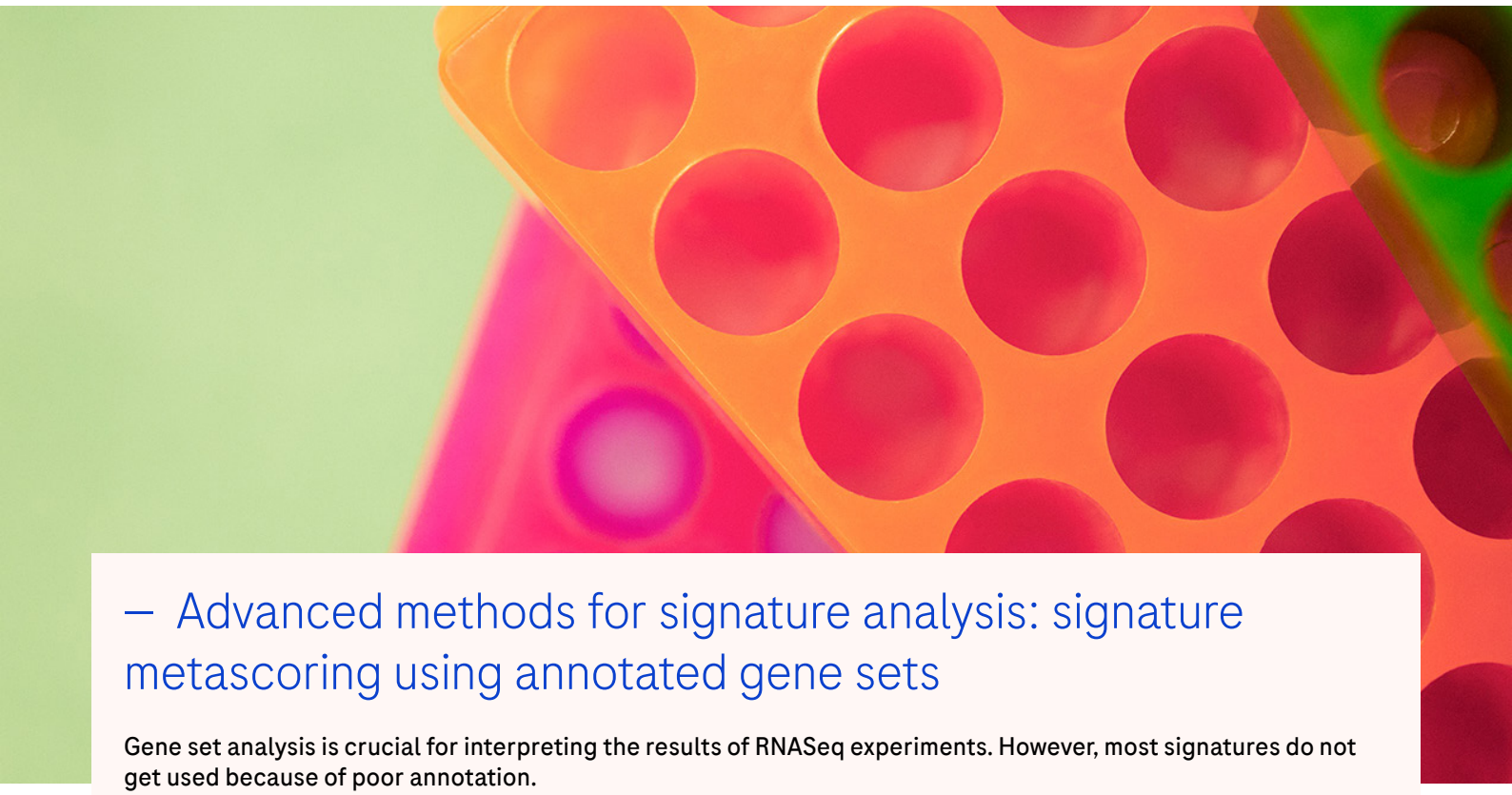
HEALTHCARE XPLORERS
Academic minds for innovation

Roche



– Advanced methods for signature analysis: signature metascoring using annotated gene sets

Gene set analysis is crucial for interpreting the results of RNASeq experiments. However, most signatures do not get used because of poor annotation.

Apply until **19 April 2022** / Xplorers Camp on **29 April 2022**

Question to be solved

- How can we make better use of the rich collection of signatures available both externally and internally?
- How do we resolve multiple signatures which purportedly represent the same biological theme?
- How should signatures be tagged (annotated) to best reflect the biological theme they represent?

General background

A signature or a gene set is a group of genes that represents a biological theme (signaling pathway, disease subtype or cell type). Most signatures are underutilized due to non-existing or poor annotations. Often, the names of the signatures do not provide an intuitive link to the biological theme they represent.

Broad Institute took the approach of generating a single “hallmark” signature for each biological theme, by combining and refining all signatures belonging to that biological theme [1]. One limitation of this approach is that the arduous process of generating all hallmark signatures has to be repeated with the addition of new signatures. Another is the need for manual curation in order to determine the biological theme that a signature belongs to. Moreover, the number of biological themes covered is limited to fifty. Finally, the hallmark signatures could be biased by the selection of the gene expression datasets against which their member genes were picked.

This project proposes a different approach which does away with the aforementioned limitations. The output of a gene set analysis method on an unlimited number of signatures (for example, enrichment scores from GSEA for all signatures in MSigDB) is fed to a subsequent step. This step, called the metascoring step, then gives a final score to each biological theme. To enable this approach, the signatures are to be annotated using natural language processing (NLP) methods and terms from standard ontologies such as cell and pathway ontologies on OBO Foundry and BioPortal..

Data types & technologies

- Data Types
  - Text (signature names, metadata, gene symbols)
- Technologies/platform
  - MongoDB
  - Python
  - FastAPI
  - High Performance Computing (Unix)

Supporting material or links

- Reference
  - The Molecular Signatures Database (MSigDB) hallmark gene set collection. Liberzon et al., Cell Syst. 2015 Dec 23; 1(6): 417–425.
- Work already done on our internal signature database and annotation pipeline (https://docs.google.com/presentation/d/1HkqR8YlyY8WIMK_56bu9-clP-I7z35xM/edit)

Needed skills

- Python programming experience and skills
- Natural language processing (NLP)

Mentor

Dr. Chiahuey Ooi
Principal Scientist, Predictive Modeling and Data Analytics

Form of cooperation

Preferred scale: 6 months full-time in the year 2022
Possible format: Internship

How to present your idea

Preferred presentation medium: Powerpoint/Google slide – please show how you would approach the problem in 5-6 slides (10 minutes). Note: Although a rudimentary pipeline is already in place for annotating and managing the signatures (based on NLP, Python, MongoDB and FastAPI), we are still looking for ideas on how to improve the annotation pipeline.
Specific skills I will check during the pitch session: Python knowledge, familiarity with NLP.