Hypothesis or Research Question(s): We hypothesize a consensus generated from up to 200 citizen science results demarking clusters in 2D flow cytometry plots will provide a more accurate result than any single random sample when aided by information such as 2D density distribution and an estimation of the number of clusters to be identified, and that millions consensus can be used to train a ML approach for cluster identification.

PROJECT BACKGROUND & SUMMARY

Flow cytometry (FCM) is the predominant technique used to characterize and quantify the widest variety of cell types, and is widely used in immunology, including for cancer diagnosis/treatment, and vaccine development. The current standard practice for data analysis involves manual inspection of a hierarchy of bivariate plots of cell types. It is extremely time consuming, the principal source of variation in the application of the technology, and incomplete due to the complexity of datasets. Many algorithms have been developed to accelerate and standardize the analysis and have shown success in many applications, but none have the performance or capabilities that have led to widespread adoption. Machine Learning (ML) approaches hold the most promise, but the lack of sufficient training data remains a major bottleneck. The development of a tool that facilitates the investigation of the massive amount of information generated by the latest instruments would represent a major breakthrough in the field. We aim to completely automate the gating process providing high accuracy results, in a short amount of time. However, to do so requires expertise in biological sciences (e.g., flow cytometry), complemented by ML and image recognition domain knowledge.

In response to this urgent need, we have led the development of a citizen science infrastructure for the analysis of FCM data, with a focus on the annotation of FCM data from peer-reviewed COVID-19 studies. In collaboration with the video game developer CCP we developed a mini game within Eve Online, a massively online multiplayer game gathering 300k players monthly. This initiative has engaged hundreds of thousands of participants and generated 202M+ of novel FCM annotations. Our unique approach to parallelize analysis has enabled a much deeper analysis than can be done by any single lab by themselves on their own data.

TSs will aid in the development of a method to post-process results, generating a consensus of players' results. This will require development of a robust mechanism to, for each sample, estimate the number of desired cell populations to combine the results into. We propose to develop a ML approach to do so, based on expertly generated manual results on a heterogeneous, curated dataset of 2D FCM samples with the number of cell populations provided. This will be used to train with a generic image boundary detection algorithm developed based on large curated object database. We further hypothesize that the predicted cluster number or any player's plot could be refined by 2D density information (e.g., the number of peaks present). In parallel TSs will collaborate in developing a further approach to use this data to train a hierarchical ML approach for cell population identification. We propose an approach based on matching the consensus bivariate FCM expression data based on the similarity expression densities to target data and applying a training model based on a decision tree.

BENEFIT TO THE STUDENTS

TSs will have opportunities to gain a deep understanding and new skills in the areas of machine leaning (ML), big (dimension, total size) data analysis, clustering, parallel computing, and flow cytometry - all at the cutting edge of science. This high profile (TV, news, internet) project will also provide TSs experience

in research taking place in a highly collaborative, environment with a small group of tightly knot researchers in Montreal, Geneva, Wellington (New Zealand) and Reykjavík, alongside massive network of thousands of citizen scientists worldwide. TSs will work most closely with local team members, including daily interactions with the graduate student who is leading the data science for the entire project and senior staff members leading infrastructure efforts. TSs will provide project updates during the weekly all hands meetings of the lab, and individual review presentations at project end, arrange meetings with co-supervisors for any challenges or questions as well as attending shorter twice a week scrum meetings. This will give TSs an appreciation of different reporting and project management methods. Through this project we will bring in additional experience in ML through the co-supervisor with whom we previously collaborated with, who will help guide TSs efforts. As such TSs will join a well-functioning group at a time where all the pieces are in place for them to step in and contribute at the highest level. TSs have access to significant computing resources, deep domain knowledge across all aspects of the project alongside comprehensive documentation including extensive background information. TSs will also gain experience in the R programming language, statistics, visualization and other aspects of big data analysis. TSs will also gain experience in documentation processes, crucial for such a massive effort. TSs will gain valuable experience in literature review and obtaining information on flow cytometry (FCM) data and data processing. TSs will also able to develop skills on independent learning of programming and the specific applications of FCM libraries available in R. TSs will also learn about taking ownership and responsibility of their work, promoting self-awareness and conducting self-reflection. TSs will gain an appreciation of how one individual's research can have a significant impact on an entire clinical and research community. The local research team is highly diverse with many different backgrounds and cultures, providing the students with an opportunity to develop communication skills in an open and welcoming environment. TSs will produce manuscript-ready graphics based on their work for publication in peer-reviewed research. They will document their methodology in a manner suitable for academic publication in the expected peer-reviewed manuscript.

TS1 will experiment with pre-checked plots and obtain the most optimal 2D density thresholds to use in functions for quality checking (QC) algorithms. Threshold parameterization will be guided by an already developed set of rules that have been shown to be effective in supervised analysis. TS1 will learn to define parameters, specifications, and outputs for the QC algorithms. TS1 will work with the Wellington collaborator on integrating the QC component to the ML framework already in place and tested. TS1 will develop an approach that uses a combination of the aggregated input of up to 200 players per sample, information provided by the 2D density distribution and the ML approach trained on expert derived cluster number data to derive a final cluster number estimate upon which to base the consensus of player results. TS1 will work with the senior staff lead on the project to test, and as necessary improve, the current aggregation approach.

TS2 till work with the graduate student co-supervisor in the final development and testing of the near duplication detection approach. They will work with the local research team in generating test datasets including by co-mixing datafiles to generate a spectrum of related files.

TS3 will work with graduate student supervisor in the development and testing of the final ML algorithm. A focus will be on the evaluation of individual and aggregated manual result vs gold standard expertly analyzed data and evaluation of the ML clustering algorithm against the same gold standard data. Approaches will be developed and tested for automation of the cluster matching and for scoring metrics.