CBCB Summer 2022 BRiDGe Potential Projects

Table of Contents

Microbiome Analysis

Assembly graph approaches for identifying structural variation in human and environmental metagenomes

PIs: Mihai Pop and Jackie Michaelis

Background: Large-scale microbiome initiatives increasingly demonstrate the importance of the microorganisms around us. These studies are driven by advances in culture-independent, high-throughput sequencing techniques that facilitate characterization of complex microbial communities. By reconstructing or assembling the genomes of different microbes within a population, we can extract gene content and identify genomic variation, from single nucleotide variants (SNVs) to gene insertion/deletion (indel) events, that may alter microbial functional or pathogenic capacities and contribute to human and environmental health or disease. Detection of variants in metagenomes is challenging. Our lab developed MetaCarvel, a method for de-novo scaffolding and reference-free variant discovery in whole metagenomic shotgun sequencing datasets.

Project Goal: Understand how the choice of metagenome assembly affects structural variant detection.

Students will apply several different metagenome assembly tools to assemble and scaffold a set of samples. The students will then use sequence alignment methods to characterize differences in the assembly output. They will also compare the number and types of structural variants detected using all assembly methods and identify how many variants are shared across assembly methods and how many are uniquely identified by one assembly method. This will help elucidate the impact of assembly methods on downstream analyses and interpretation of results.

Exploring microbial interactions in childhood diarrheal disease

PIs: Mihai Pop and Jackie Michaelis

Background: In low-income countries, moderate to severe diarrhea (MSD) significantly contributes to infant mortality. Although many individual pathogens are linked to disease etiology, the importance of the overall intestinal microbial community structure is less well understood. A 2014 marker gene study from our lab characterized the fecal microbiota of 992 children from four developing countries in Africa and Southeast Asia, comparing MSD cases to healthy controls. Since this publication, we have built upon this initial work and characterized the fecal microbiota of over 3,000 children from the same MSD cohort. In addition to 16S rRNA marker gene data, we have various metadata, qPCR data, and some whole metagenomic shotgun sequencing data. With the increased sample size and taxonomic resolution of the expanded dataset, there are many important questions that interested undergraduates could explore.

Project Goal: Use network analysis techniques to identify bacteria that are co-present or co-absent, suggesting that they are interacting to contribute to diarrheal disease.

Students will organize and format the data, run software specific for network analyses, and tune parameters of the analyses to see how robust the network is to change. Students will identify statistically significant relationships between bacteria and perform literature searches to see if their results correlate with any other research.

Elucidating health-relevant functions of the human gut microbiome

PI: Brantley Hall

Background: The human gut microbiome is a dense assemblage of diverse microbes that play important roles, both positive and negative, in human health. For example, byproducts of gut microbial fermentation such as butyrate supply up to 70 percent of the energy requirements for the human cells lining the colon. Gut microbes also produce metabolites with negative effects on human health, such as TMAO, which is strongly implicated in cardiovascular disease. Our understanding of the human gut microbiome and its effects on human health have improved in the last 20 years, but it is estimated that we still lack understanding of the vast majority of health-relevant host-microbe interactions. This lack of understanding is rooted in the poor annotation of gut microbial genomes. There are millions of genes present in gut microbial genomes that have no known function, and even the genes that are assigned annotations are often mis annotated. The Hall Lab combines computational biology and microbiology to better understand these functions and their distribution across gut microbial species.

Project Goal: Better understand health-relevant functions of the human gut microbiome through better annotation of the functions contained within gut microbial genomes.

In preparation for the REU students, the Hall Lab will prepare a list of microbial functions that are feasible to annotate in the abbreviated REU timeframe and perform an initial literature search. When the REU students arrive, they will be paired with a graduate student in the lab who will guide them through the annotation of the assigned function using a pipeline co-developed by my lab, GutFunFind.

Single-cell RNA Sequencing

Improved methods for querying single-cell gene expression data

PI: Rob Patro

Background: The development of high-throughput single-cell RNA-sequencing technologies has, in only a few short years, become a transformational tool in helping to understand many complex biological processes. For example, the creation of single-cell “atlases” have helped to discover and characterize new cell types and cell states, and to elucidate the roles they play in different tissues and disease states. Single-cell experiments conducted during embryogenesis have provided a looking glass, at cell-level resolution, into some of the mechanisms that

determine cell fate and tissue differentiation, with potentially broad implications for regenerative biology and tissue engineering.

However, these new technologies have brought with them a host of new computational challenges. Due to the nature of how molecules are “tagged” and “barcoded”, traditional methods developed for processing bulk RNA-seq data are not appropriate for processing data from the highest-throughput single-cell RNA-seq protocols. Fundamental differences in the data, in the type of experimental measurement errors and noise, and in the scale of the data itself, necessitate the development of new algorithms, data structures, and statistical methods for processing and analyzing this data. While tremendous effort has been devoted in the bioinformatics community to developing tools for analyzing processed data (in the form of gene by cell count matrices cataloging how frequently each gene was observed in each cell), comparatively little work has addressed the raw sequencing data should be processed to arrive at accurate gene by cell count matrices. Likewise, few methods have been developed to allow scientists to leverage the large and quickly-growing collection of single-cell atlases via exploratory data analysis. Tools exist to help probe annotated or pre-classified parts of the atlas, but few tools exist to allow unstructured and exploratory queries over millions of cells.

Project Goal: Students would focus on developing and deploying a scalable expression-based search system for large-scale single-cell RNA-seq atlases. Students would help to build a system that would let users query across millions of cells to find those expressing (or repressing) certain subsets of genes. This would involve working on improving the fundamental data structures, optimization of queries, development of statistical measures of significance, or the development of a front end to make the posing and answering of biologically meaningful queries simpler.

Improved methods for quantifying single-cell gene expression data

PI: Rob Patro

Background: We have developed some of the first methods that attempt to process single-cell RNA-seq data in a principled framework that accounts for many of the experimental complexities that arise during measurement. For example, our method “alevin” is the first to propose a principled way of handling sequencing reads that align to (may have been generated from) more than one gene — other approaches simply discard such data, resulting in potential biases in subsequent estimation. Our approach also allows quantifying not just gene expression, but also the inherent uncertainty associated with each estimate, which can inform downstream analyses if available. We have also built the first sequence-level simulator for this type of data to aid in the further development and assessment of new computational processing methods, and we have explored statistical methods for sharing information among the cells sequenced together in a sample to improve the accuracy of expression estimates. Yet, despite these advances, the accuracy and robustness of methods for processing single-cell data still lag far behind their counterparts used in analyzing bulk RNA-seq data. This provides a number of exciting research opportunities for students interested in bioinformatics to contribute to practical methods that are likely to have a significant impact in helping scientists make the most of this rich new source of experimental data.

Project Goal: For this REU, students will participate in one of the following projects:

- Developing methods for the improved visualization of RNA-velocity results. RNA-velocity is an increasingly popular analysis in single-cell RNA-seq where estimates of the spliced and unspliced molecule ratios in cells are used to model the dynamics of cells in expression space. Current approaches to visualize RNA-velocity perform standard dimensionality reduction on expression profiles, and then visualize velocity as a vector field overlaid upon the dimensionally reduced cell embedding. In this project, students will investigate how to modify the loss function of the dimensionality reduction technique itself (e.g. t-SNE / UMAP) to account for the RNA-velocity signal. Specifically, they will explore modifying the loss function to reward velocity coherence of nearby cells, and also to encourage the embedding distance between cells to be inversely proportional to velocity magnitude, to avoid crowding in the final visualization.
- Developing improved methods for statistical information sharing between the cells sequenced together in a sample. Though we have taken useful first steps in this direction, many improvements are possible, and sharing information across subsets of similar (in terms of gene expression) cells holds the potential to substantially improve expression estimation accuracy. Students will explore the effect of different neighbor selection strategies (distance metrics, number of selected neighbors, etc.) and the effect of different prior strengths on information sharing results. Students will also work with Dr. Patro's PhD students to develop a notion of information sharing relevant to the ambiguity in the splicing status (i.e. spliced vs. unspliced) of reads, rather than just the gene of origin of reads.

Machine Learning

Machine learning and model interpretation for longitudinal data

PI: Michael Cummings

Prerequisites: Experience in Python programming is preferred but not required.

Background: Longitudinal data consists of repeated measures for the same study cohort over a series of time points. An example is the longitudinal monitoring and assessment of Parkinson's Disease, where patient data are collected during each patient visit. The use of data in a time series can help improve the prediction of outcome. Our research focuses on applying machine learning techniques to longitudinal data to improve patient outcome prediction. We aim to build accurate machine learning models and understand the important measures contributing to model performance. Ongoing research includes investigating the possibility of developing an interpretable machine learning framework for longitudinal data. The study provides opportunities for learning data analysis and machine learning methods.

Project Goal: Performing feature engineering, designing and constructing machine learning models, evaluating and interpreting the machine learning models.

Phylogenomics

Parallelizing species tree estimation methods for multi-copy genes

PI: Erin Molloy

Prerequisites: Example project #2 requires programming in C++ or Java and thus is aimed at undergraduate students who have taken courses in these areas.

Background: A critical step in many biological studies is the estimation of evolutionary trees from genomic data. Of particular interest is the species tree, which models how a set of species evolved from a common ancestor. It is now widely recognized that biological processes, such as gene duplication and loss (GDL), can result in individual regions of the genome ("genes") to have evolutionary histories that differ from the species tree. Methods that leverage the phylogenetic signal in genes, including those with multiple copies, are expected to improve our understanding of species evolution. Recently, there has been an explosion of new (and highly promising) methods for this problem. However, these methods are computationally intensive and will not scale to the ultra-large datasets that are currently being assembled by initiatives, such as the 10,000 Plant Genomes Project. Our goal is to scale methods to tens of thousands of species and tens of thousands of genes. In this REU, students will have the opportunity to work on projects that address scalability from different perspectives.

Project Goal: Improving the scalability of methods for estimating species trees from gene families.

Example project #1: Our group has developed two divide-and-conquer approaches for species tree estimation but these have not yet been explored in the context of species tree estimation from multi-copy genes. REU students would have the opportunity to evaluate these divide-and-conquer approaches in this new context, exploring the trade-offs between accuracy and scalability. This project requires running analyses (e.g. with bash scripting) and analyzing the results (e.g. with Python scripting) and thus is appropriate for REU students at a variety of levels. Depending on the abilities and interests of the REU students, they would also be able to propose / explore modifications of these divide-and-conquer approaches that may improve accuracy.

Example project #2: One of the methods for species tree estimation from multi-copy genes was developed in our group. This method, called FastMulRFS, uses a dynamic programming algorithm to solve its NP-hard problem exactly but within a constrained search space. Some related methods (with different optimization criteria) have recently been improved through a combination of vectorization, CPU-threading, and GPU-threading (note: the non-senior authors of this paper were undergraduate students when the work was completed). REU students would have the opportunity to improve the scalability of FastMulRFS (or related methods) by

+ profiling the method to identify compute kernels with computational bottlenecks
+ studying ***one*** of these compute kernels for which there is a bottleneck and identitying parallelism in the algorithm
+ implementing a parallel version of the algorithm (e.g. vectorization, CPU-threading, GPU-threading, message passing)

+ evaluating performance in scaling experiments using both simulated and biological datasets

The broader goal is to have undergraduate students study (and ideally improve) the scalability of species tree estimation that take multi-copy genes as input. Broadly speaky, REU students will gain experience in parallel computing as well as discrete algorithms.