GVSU Health and Bioinformatics PSM Program Internship Report

David Pearson
Corewell Health

Fall Semester 2023

Table of Contents

Learning Objectives

To begin my internship experience I had three objectives that I was excited to work towards. The first one being to apply all the skills I have learned throughout my higher education in real-world applications. I have spent many hours in a classroom setting where I would get an introduction to many different ideas and tools, but the classroom does not provide the opportunity to really master these skills like an internship can. For example, I wanted to take some of the knowledge that I was briefly introduced to in CIS 661 and build upon those in a professional setting - with real data and the potential to discover something impactful. Secondly, I wanted to learn as many bioinformatics related tools as possible. We were briefly introduced to some tools in CIS 661, but I knew that there was so much more out there that we didn't have time to cover in class. Specifically, I was the most interested in learning new tools that allow for the visualization of large amounts of data. Information visualization is an extremely important part of the bioinformatics pipeline and thus I wanted to learn new tools to improve my skills in this area. My final goal for this internship was to brush up on my biological science knowledge and to increase my knowledge in this space. Much of my time in my PSM program has been spent more on advancing my computer science ability, and I haven't spent as much time as I would have liked learning more biology since I graduated with my B.S. in Human Biology in 2020. My goal was to make strides in my biological knowledge through reading through research as well as making sure I understand the background biology in the new tools I learn.

Introduction

I have been completing my internship requirement at Corewell Health. I was hired in August 2023 and my internship lasts until July 31, 2024. My role at Corewell Health is a Public Health Intern. I was hired through use of the Michigan Sequencing Academic Partnership for Public Health Innovation and Response (MI-SAPPHIRE). This was possible through funding through the Michigan Department of Health and Human Services and a partnership with a CDC Epidemiology and Laboratory capacity grant aimed at sequence generation and analysis. This internship has a primary location in Grand Rapids, MI but most of the time work was done remotely. I was hired into this internship to work under [redacted] Research and Development at Corewell Health. I also worked primarily with my internship coordinator [redacted] My unique focus in this position would be in the realm of bioinformatics. The primary goal of my position entailed studying infectious diseases and how they may impact public health in a variety of ways.

Description of Work

Before beginning a description of the work that has been completed so far through this internship it should be noted that this internship is one year in length, and I will be continuing with work here after the Fall semester concludes. This means that much of my work here is not yet completed and is ongoing. I have primarily worked on two different projects during my time at Corewell Health thus far.

The first being a personal project of my choice. Initially upon being hired by Corewell Health I was asked to pick an area that I am interested in which I would like to conduct research in. The only requirements for this project were that it is related to infectious diseases and their impact on public health in some way. Ultimately, I decided that I wanted to focus my work on the human microbiome and how infectious diseases may cause disruption that could lead to long-term negative consequences (possibly cancer). I am very interested in infectious diseases which cause cancer and the mechanism of action of which that takes place through. I had a few specific questions that I was particularly interested in investigating further:

- How does the composition of an individual's microbiome impact their response to COVID-19 (or a different disease of interest)?
- How do microbes and human cells interact with each other? How is this interaction changed when a disease enters the body? Can this interaction increase the chance of formation of tumors?
- How do microbes interact with immunotherapy or other types of treatment? What changes are happening to the microbiome during this treatment?

To begin my research into these questions I first conducted a literature review and attempted to find a dataset which may be useful to analyze. Through this process I learned much of the background biology that relates to what is known about the microbiome, how it changes upon infection and what those changes can lead to long-term. I also found Illumina NovaSeq paired end sequencing data which I believed could be beneficial to analyze to further investigate in this realm. This data was taken from the gut microbiome of COVID-19 patients in Japan. I was curious as to if metagenomic analysis of this data may provide further insights into how the gut microbiome is changing after infection by COVID-19 and what long-term risks may be associated with that.

The data for this work was found on the National Center for Biotechnology Information section of the National Library of Medicine website. Specifically, the Sequence Read Archive (SRA) was utilized and FASTQ files from the experiment were downloaded. All work through this pipeline was done with the assistance of Jeremy Prokop.

Several tools were used after the data download. First FastQC was used as a method of quality control. FastQC is a way to perform quality control checks on raw sequencing data. Essentially the FASTQ files were imported, and a report is generated that shows where there may be problems in the data. Several graphs and tables are also generated through this process. Some of the modules that are generated in a FastQC report include "Per Base Sequence Quality", "Per Sequence Quality Scores", "Per Base Sequence Content", "Per Sequence GC Content", "Per Base N Content", "Sequence Length Distribution", "Duplicate Sequences", "Overrepresented Sequences", "Adapter Content", "Kmer Content" and "Per Tile Sequence Quality". After this step, a MultiQC report was generated which simply summarizes the found information using the statistics above.

After the quality control step, further analysis was conducted using a variety of tools - Salmon, Kraken2 Immunarch and MiXCR. Salmon is a tool for quantifying the expression of transcripts using RNA-Seq data. Kraken2 is a taxonomic classification method which matches by using kmers. Immunarch is an R package that is used to analyze both T cell receptor and B cell

receptor repertoires. MiXCR is also a software used for T cell receptor and B cell receptor analysis. This step of the analysis pipeline is still in progress, as a new and more urgent project was introduced about halfway through the semester.

The second half of the semester I worked on a separate project that expanded on using the Immunarch package in R. This project is also currently ongoing. The goal of this work is to compare immune repertoires from several different data sources that Corewell Health has acquired. These data sources include patients with unique infections such as MODS, RSV and COVID-19. Importantly, one of these studies focuses on NICU patients which have a variety of genetic diseases. Immunarch can be used to compare the immune repertoires of the NICU patients with the patients from other studies……

Research in this area is currently ongoing, so for the purposes of this report it would be more appropriate to discuss how the Immunarch package can be utilized to analyze immune repertoires. There are three main areas in which Immunarch is useful: Basic analysis and clonality, repertoire analysis and comparison, and clonotype analysis and comparison.

To begin with some of the basic analysis that can be performed using Immunarch, there are two main functions that are useful - “repExplore” and “repClonality”. The “repExplore” function can be used to calculate basic statistics at the repertoire-level such as the number of unique clonotypes, the distribution of clonotype abundances, the distribution of CDR3 sequence lengths or the number of clones per repertoire. The “repClonality” function can be used to estimate the diversity of samples. For example, you can compute the percentage of clonotypes required to occupy a specified percent of the total immune repertoire.

Regarding repertoire-level analysis and comparison, the Immunarch tool has been beneficial by its functions to both create large tables with all clonotypes from the list of repertoires and functions for diversity estimation. It is important for the research being conducted to understand what clonotypes are the most abundant in different studies. Immunarch has allowed me to create several different tables which allow me to compare the most abundant clonotypes between certain populations. For example, I have created tables to compare clonotypes of MODS patients vs RSV patients vs NICU patients vs COVID-19 patients, and I have also created tables which more broadly compare COVID-19 patients with all other patients. Repertoire diversity can also be calculated for samples using a variety of different methods. For example, the “repDiversity” function can be used to estimate species richness in several samples. Analysis of clonotype abundance and repertoire diversity is still ongoing.

Clonotype-level analysis is also possible with Immunarch. Specifically, the “trackClonotypes” function can be utilized to either track the most abundant clonotypes in one sample throughout all other samples or track a list of clonotypes chosen by the user across several samples. It can be seen how this function can be used together with the tables that were generated at the repertoire-level to track clonotypes that were abundant in one study and see their abundance in the other studies. For example, clonotypes with a known association with COVID-19 could be tracked across all repertories and it would be possible to visualize which samples in the repertoire are likely infected with COVID-19 (or vaccinated) and which are not. This analysis of clonotype tracking is also still ongoing.

Immunarch can clearly be an extremely useful tool for analyzing immune repertoire data. Over the next few months, I plan on delving deeper into this analysis and exploring the true potential that this package provides.

Internship Discussion

Referring to the three objectives I had set for myself over the course of this internship, I believe that I have achieved all three of them - and will continue to build upon them over the remaining eight months. My first objective was to apply the skills that I’ve learned in the classroom in a more real-life setting. The data that I’m currently working with was obtained from actual patients, and the results of what we may potentially determine through analysis is currently unknown. Any findings that we unveil could have a real impact on public health. I have found this aspect of the internship to be the most exciting - the potential that we could make a real difference. My second goal was to increase my skill set with as many different bioinformatics tools as possible. I have spent most of my time with the Immunarch package as I mentioned, and I feel as though I am starting to get a real grasp on the potential of the tool and what opportunities it holds when analyzing different repertoires. I have also gotten experience with FastQC, MultiQC, Salmon, Kraken2 and MiXCR. I have spent a significant amount of time reading up on these tools as well as getting some hands-on experience with them. I feel more comfortable using these tools and better understand what they can be used to accomplish. Finally, my last objective was to brush up on my biology knowledge and get a better understanding of the biological research that is currently being conducted. I accomplished this goal in two ways. First, my literature review that I conducted at the beginning of the semester advanced both my biological knowledge as well as gave me a better idea of what kind of research is being conducted regarding the human microbiome and infectious diseases. Second, through my research into the various tools I described above I have tried my best to also understand the biology behind everything. I do not find it beneficial to just be able to understand how to use each tool, as it is also important to understand the background on what

your data means in a biological sense. When I produce any sort of visualization, I want to be able to describe to someone what that visualization is showing biologically, and in order to do that I had to understand the biology behind the tool. I hope to further my biological knowledge even further during the remaining time of my internship.

In addition to the scientific skills that I have already discussed throughout this report, I also have improved on some of my professional skills throughout my time at Corewell Health. My time management skills have certainly improved since starting my internship here. With the job being remote and sometimes going days without being in contact with any co-workers, it was critical to keep up to date with all my work. The environment is very different from a classroom environment where you get an assignment that's due the next week and thus requires more planning to ensure that all work is on track to being completed on time. I also believe I have improved at working within a group and bouncing ideas around in said group to progress everyone's research. Throughout my academic career, group work has usually been more individual where each member of the group simply completes their single section of the project and very little communication occurs. Group environments at Corewell Health were much more collaborative where everyone communicated to each other in a way that is beneficial to everyone, and as a result I improved my communication skills in this area.

My experience with the PSM coursework did prepare me for the scientific content of this internship, specifically the more computational side of things. Before beginning the PSM program I had very little experience in computer science and so had many different areas that I needed to improve in. Specifically, CIS 635 and STA 610 had a largely positive impact as they introduced me to R and enhanced my comfortability using that program. Much of my current work is done using R Studio and so this background knowledge has been helpful. CIS 661 was also beneficial in preparing me for the scientific content of this internship as it introduced me to some of the research being conducted which makes use of bioinformatics and some of the tools that bioinformaticians are making use of. Having this background knowledge has been largely helpful to me during this internship.

The professional content that was provided by the PSM program was hugely beneficial to me in both securing an internship and while working with Corewell Health. I believe that the courses I took which helped with both job applications and the interview process were very helpful and pointed me in the right direction in terms of how to create a resume, write a cover letter and conduct myself during interviews. These courses have also been helpful in showing me how to conduct myself in a work environment and giving me knowledge of how many of these companies are structured.

The biggest challenge that I faced during this internship was getting up to date on where scientific research is at. I had very little idea as to what research is currently being conducted or where the next ideas for research to be conducted were going. This made formulating a research question at the beginning of the semester challenging as I wasn't sure what a "good" question to ask would be. To overcome this challenge, I put in work completing a literature review looking into what recent research is out there. Every paper that I read I would make sure to emphasize their "Future Work" section to try and get a good idea for what questions scientists have currently that are yet to be answered, and what areas I can have an impact in.

Overall, the internship experience has been extremely positive for me. This has been a great introduction to the world of scientific research and the role of bioinformaticians in it. Corewell Health has done an excellent job at catering to what I've wanted to focus on during this internship. All three of my learning objectives that I set for myself at the beginning of the internship have progressed substantially. My biological knowledge and comparability using bioinformatics-related tools have both increased tremendously and I am excited for the remainder of my time at Corewell Health to further improve in these areas.