Optimization of the Pacific Biosciences Iso-Seq analysis workflow for human transcript structure modeling

My research is part of a larger research project headed by Dr. Colin Dewey which seeks to process and analyze transcriptome data from pluripotent stem cells to investigate the structure of the transcriptome, considering all spliceforms, and to determine dependency relationships between the various transcription and processing events. My role is to determine the accuracy of the provided Pacific Biosciences demultiplexing software and explore alternatives to its default settings. When obeying the default parameter settings, we found a significant portion of our reads were being lost because of filtering. Just creating a consensus sequence from the reads filtered out over 60%, and the demultiplexing process performed on that consensus sequence filtered out over 30% of the remaining reads. This is a lot of data to go unused. I am working on optimizing the number of reads used and the accuracy of the demultiplexing product. The more I can optimize these two parameters, the more accurate and precise my research group's transcriptome structure can be.

To achieve this goal, I have constructed pipelines for demultiplexing following the Pacific Biosciences' recommended tools and parameter settings, as well as "baseline" pipelines for demultiplexing while filtering out as few reads as possible. I now need to analyze the resulting data to understand the following: what is causing some reads to still be omitted? What filter parameters are important for accuracy, and to what extent? What filter parameters omit reads but don't significantly affect accuracy? In this stage I am developing my data science capabilities, learning such skills as how to construct and manipulate data tables, as well as produce data visualizations. Once I understand where data is being lost and why, I can begin to develop solutions, most likely in the form of alternative algorithms for one or more of the steps in the Pacific Biosciences pipeline.