The Bacillus Calmette–Guérin (BCG) vaccine was originally developed to prevent tuberculosis; however, it is now well established as a treatment for bladder cancer and is being actively investigated for neurological and autoimmune diseases. In the Faustman Lab, we currently use the Japan BCG strain in clinical trials for type 1 diabetes (T1D). Our studies have shown that BCG can reduce blood glucose levels to near-normal ranges and reprogram glucose metabolism in lymphocytes.

In addition to Japan BCG, our lab also works with another BCG source known as Aeras BCG, which was developed by the Gates Foundation and originally reported to be sub-cultured from the Connaught BCG strain. The goal of this study was to compare the genetic sequences of Japan and Aeras BCG to accurately define their lineage and assess whether Aeras BCG could be reliably used as a substitution in future clinical trials.

To accomplish this, I optimized and performed genomic DNA isolation from both BCG strains and submitted the samples to the Harvard Bauer Core for sequencing using both short-read (Illumina) and long-read (PacBio) technologies. The resulting sequencing data were analyzed by a bioinformatics specialist. Long-read alignment was first performed by mapping Japan BCG to the Tokyo reference genome and Aeras BCG to the Connaught reference genome. A GATK-based variant-calling pipeline was then applied to identify single-nucleotide polymorphisms (SNPs), insertions, and deletions. De novo genome assemblies were also generated to further characterize structural variation. Following these analyses, I received processed datasets detailing insertion sequences, deletion events, and identified SNPs.

Using the insertion and deletion sequences identified in the Aeras BCG sample, I retrieved the corresponding reference sequences from NCBI and wrote Python functions to assess whether our sequences matched the reference sequences exactly. This analysis revealed several regions containing nucleotide-level differences. To determine whether these changes were biologically meaningful, I wrote additional functions to evaluate whether the mutations disrupted start codons or introduced premature stop codons. I then translated the DNA sequences into protein sequences and compared them with reference proteins to assess whether the encoded proteins remained conserved despite underlying nucleotide variation.

Overall, our analysis identified both insertions and deletions in the Aeras BCG genome relative to reference strains. Notably, several PE and PE-PGRS family proteins present in the Pasteur and Danish strains were inserted in our sample, while corresponding PE and PE-PGRS family proteins derived from the Connaught strain were deleted. This reciprocal pattern suggests that insertions and deletions complement each other. PE and PE-PGRS family proteins are mycobacteria-specific proteins implicated in host–pathogen interactions, immune modulation, and antigenic variation, highlighting the biological significance of these strain-level differences.