Summer 2018 Project Summaries

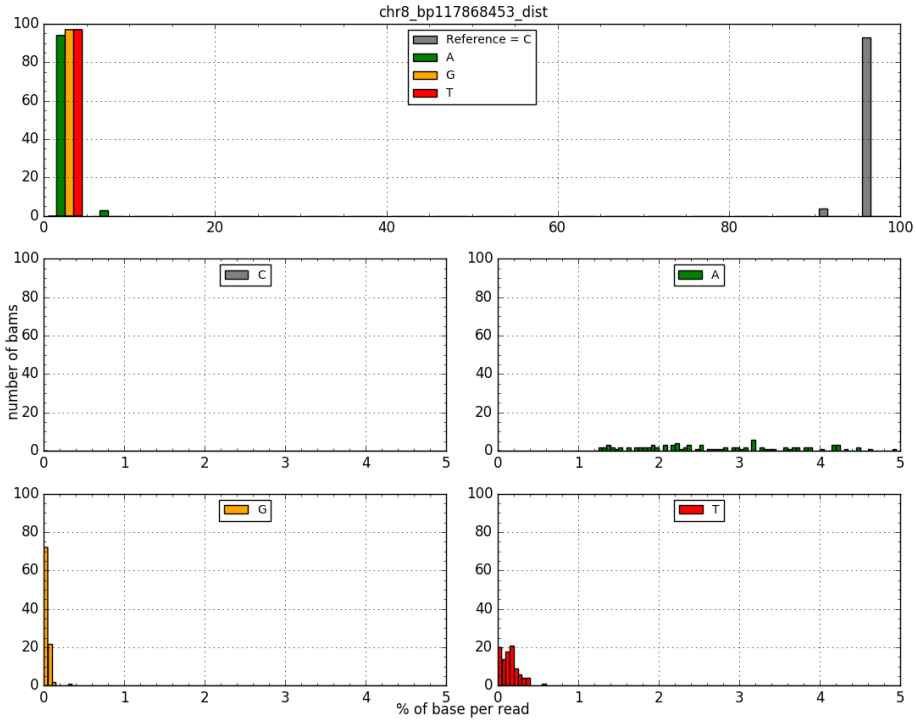
Stanford Healthcare Molecular Pathology Lab

MSI Classification Using Machine Learning to Process NGS Data

Objective: Develop a model that can classify microsatellite instability using NGS data.
Result: The first part of this project involved parsing microsatellite loci from BAM files, which I achieved using Pysam and a partial string matching algorithm to identify regions flanking mononucleotide runs at loci identified in relevant literature. Samples were obtained from The Cancer Genome Atlas. I optimized parameters such as flanking region length and number of mismatches allowed to achieve the maximum number of acceptable reads at each locus. The second part of the project involved creating a model for predicting MSI status using the data obtained from the lengths of these microsatellite regions. I achieved this using TensorFlow, Google's machine learning API. I first developed a model with close to 90% accuracy using a locus-by-locus based call, as is most commonly viewed in the literature. I then altered my model to weigh all features of all loci with ample reads during training using L1 regularization.
Implementation: Once this model reaches sufficient accuracy and validation for clinical implementation, it can be used to confidently predict MSI status from NGS data. In the future, significant MSI loci can be added to the selector for STAMP panels and MSI status can be reported on STAMP results, extending the clinical relevance of the STAMP. This will eliminate the need for PCR-based MSI testing, saving time and providing additional clinically actionable information.

VAF Noise Plots

Objective: Identify noisy positions in the STAMP or Heme-STAMP assay
Result: I implemented the Pysam module to count the proportion of each base at a user-entered position. I then plotted histograms of the distribution of each base using Matplotlib.

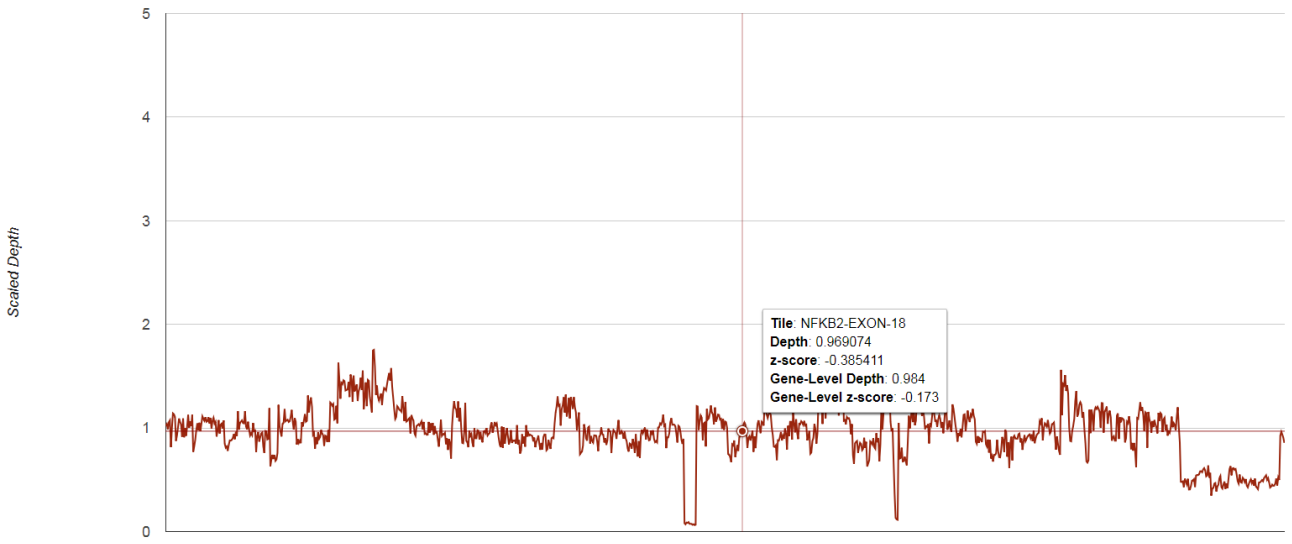


Implementation: This program will help confirm which variants to call as mutations and which are commonly noisy, thus reducing errors in variant calling and increasing the accuracy of reporting. An example of a plot showing a noisy base is shown below.

Copy Number Plots

Objective: Produce a depth plot for the STAMP assay with interactive features to show irregular copy number location

Result: I wrote a python script that reads in information for each tile within a gene and produces an HTML document that uses Google Charts to plot the copy number at that tile. I created a custom HTML tooltip that shows at each point on the plot the copy number and z-score of the tile as well as those statistics for the entire gene.



Implementation: These plots will be implemented on the STAMP and Heme-STAMP run pages on the server to allow the Fellows to more easily identify copy number irregularities. The gene-level information will show when there is an irregularity for an entire gene, simplifying the process of distinguishing 'real' copy number irregularities from sequencing artifacts.

Mutalyzer Lookup

Objective: Add fields to the Heme-STAMP variant report that show the reference transcript, amino acid change, CDS change and chromosomal location using the standard nomenclature.

Result: Using the Requests and JSON modules, I wrote a python script that queries Mutalyzer's web services and parses the result to find these pertinent data fields. The program then produces an appended variant report showing these additional fields.

Implementation: Appended variant reports will save the Fellows the step of manually finding the NCBI transcript in the local file and querying Mutalyzer to find the widely-used HGVS format. This information is crucial in finding these mutations in literature and databases to determine their clinical relevance.

Water Barcode Filter

Objective: Identify unmapped reads in the Heme-STAMP water barcodes.

Result: Using the Subprocess module, my python script reads in a water barcode Fastq sequence file and then subsequently maps it to hg19 and the Phix genome, filtering out any reads that map. It then produces a Fasta file that can be used to BLAST search for potential matches.

Implementation: This step is an important quality control measure to ensure the water blank is not contaminated and that the water barcode does not contain a patient's DNA.

Myeloid Data Entry

Objective: Report on Myeloid master spreadsheet results for CALR long deletion, FLT3 ITD, and CEBPA.
Result: I manually found each patient's Myeloid Panel Report on Epic and recorded the results for these 3 mutations.
Implementation: This additional information makes the spreadsheet more comprehensive and robust to be used in later data analysis projects.