**Final Report 2: Read-to-Reference Mapping of Genomic Variations**

GitGud: Mason Tatro, Ben Yunker, Naomi Williamson, Ellie Spahr

***Response to Reviewers***

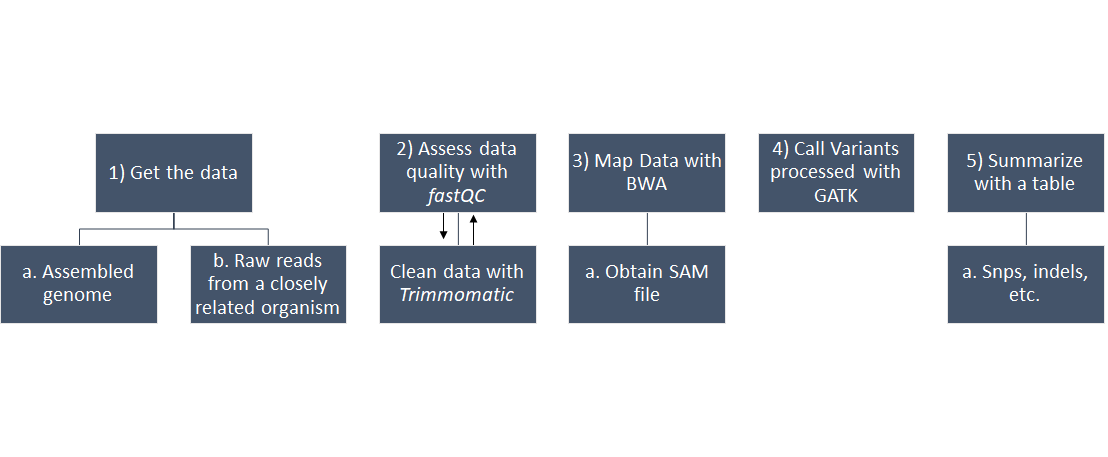
In future proposals, we will consider using the words ‘innovation’ and ‘significance’ within our document to highlight this information for our reviewers.

Spelling and formatting was considered more carefully in the final report, with the format of terminology remaining consistent. In this document, we refer to the causal agent as “*M. tuberculosis*” during discussion of bacterial isolates and their analysis and any reference to the disease state as “tuberculosis”. We believe it is important to retain this distinction when discussing these two elements of the work and have unified terminology for each.

The figure was remade below (Fig 1) to ensure it is easier to read on printed copies; original inclusion of arrows between trimmomatic and fastqc had dropped off during image export and was remedied.

Issues among the group will be resolved internally through discussion and, if necessary, input from Dr. Driscoll will be sought out in situations where a resolution cannot be reached.

Computing was carried out on the WVU Spruce Knob High Performance Supercomputer. This has been stated more clearly in our final report to ensure appropriate institutional support is reflected in our analysis pipeline.



*Figure 1: Amended Pipeline.* Pipeline steps for our proposed work; image contrast has been improved from the proposal document and arrows on step 2 have been fixed.

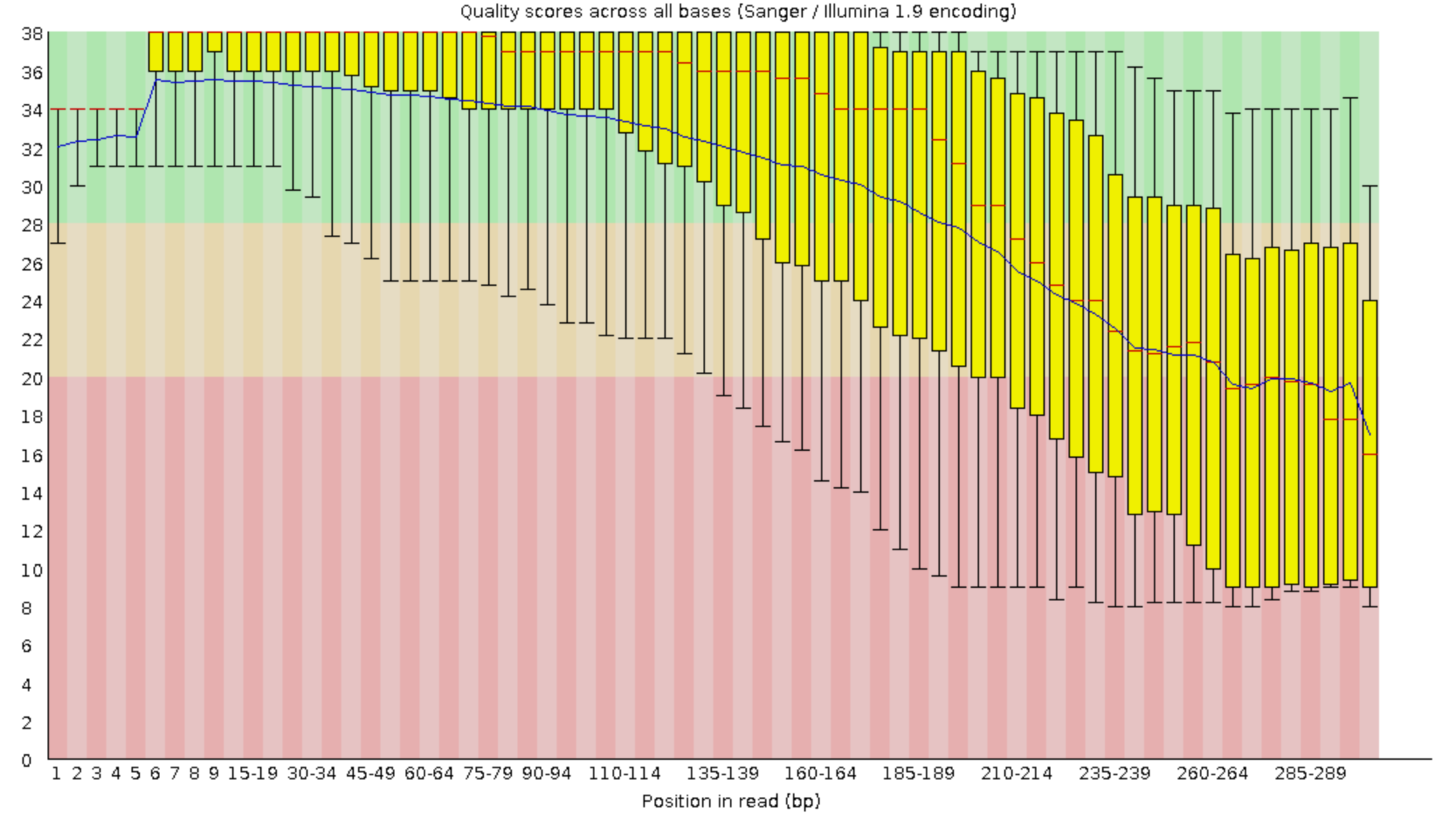
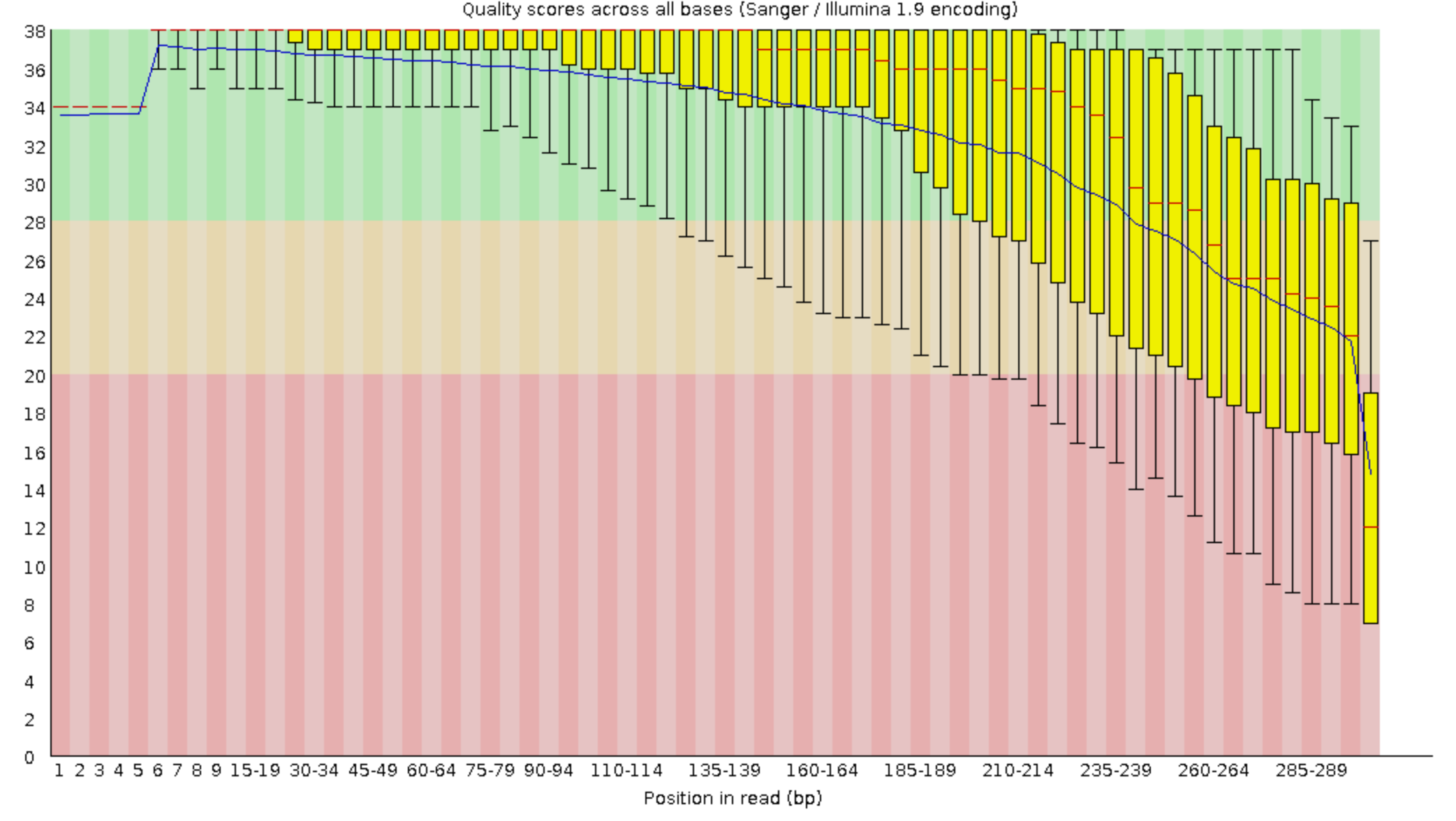
***Results***

*Data retrieval*

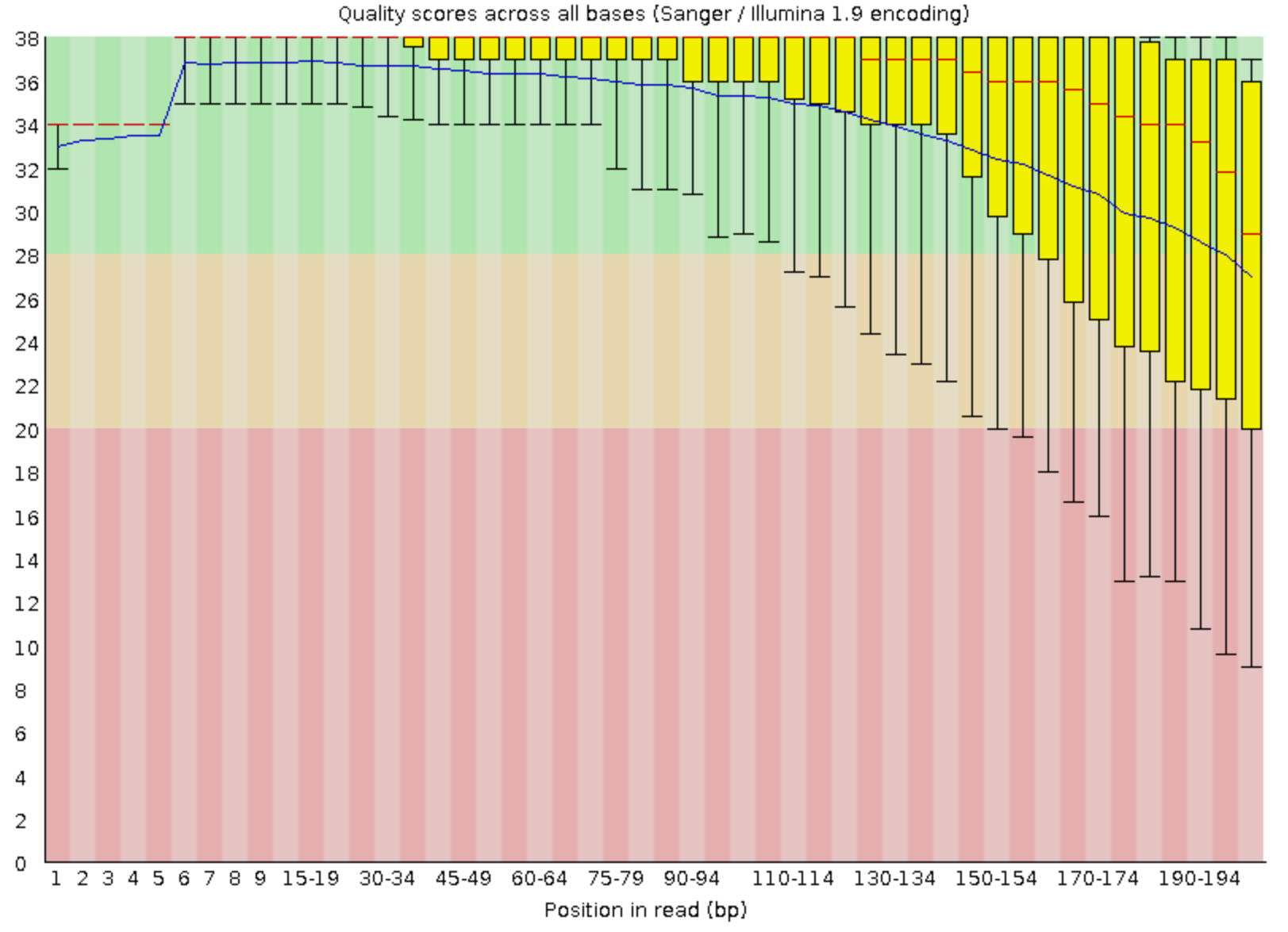
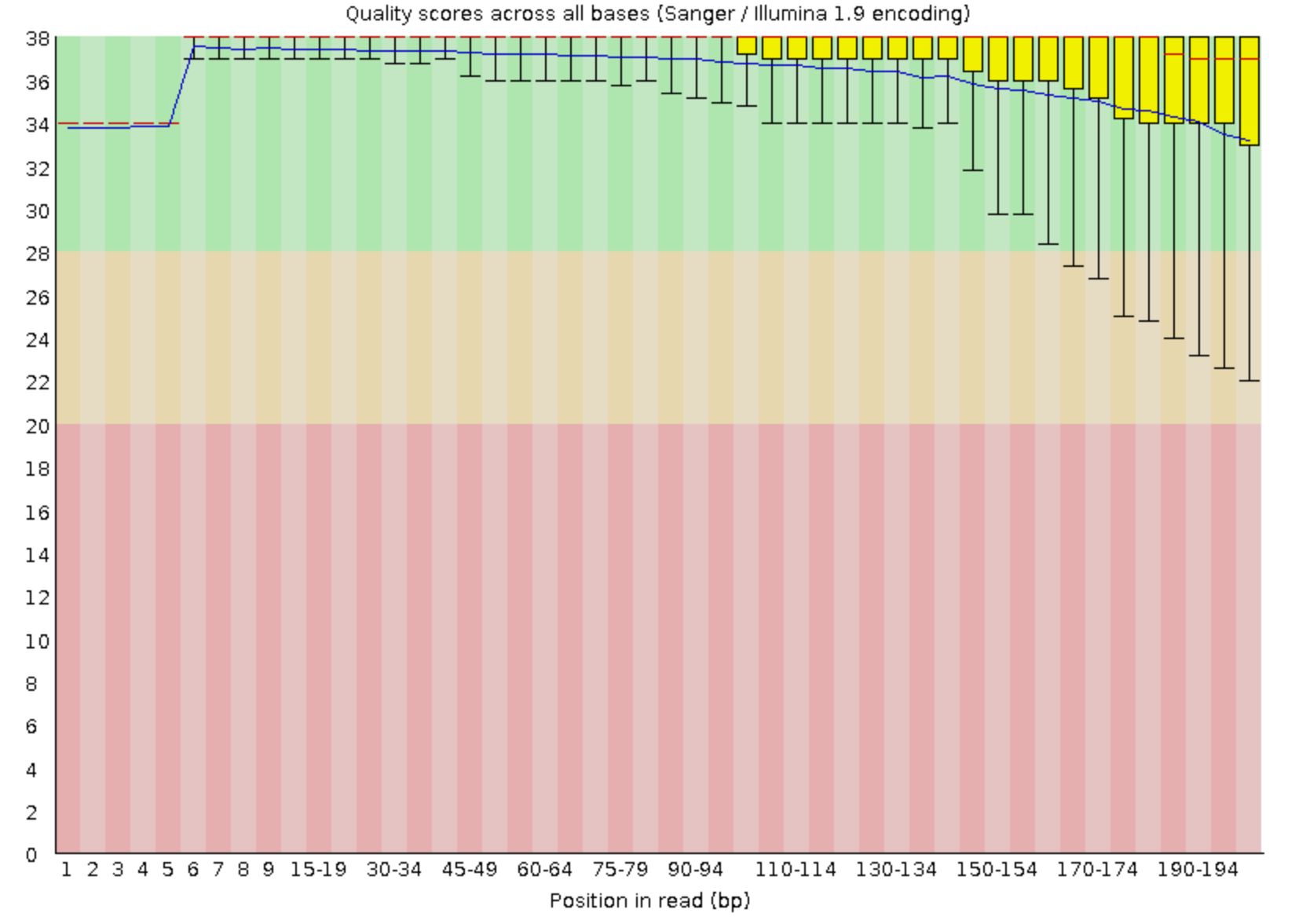
The closed reference genome for *Mycobacterium tuberculosis* H37Rv- an antibiotic-susceptible strain- was retrieved from the NCBI Database (Bioproject: [PRJNA37301](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA37301)) and uploaded to the Spruce Knob Supercomputer environment. Raw reads for the cryptic *M. tuberculosis* isolate (SRA: ERR2099775, Experiment: [ERX2157080](https://www.ebi.ac.uk/ena/data/view/ERX2157080)) were initially retrieved from NCBI’s SRA database but had been interleaved into a single file. Data was redownloaded from the European Nucleotide Archive (Run: ERR2099775) to obtain individual forward and reverse read files so trimming could be conducted in a paired-end format.

*Quality Control and Trimming*

Initial quality control of our raw *M. tuberculosis* data was conducted through FastQC. This suggested poor quality data in the 3’ ends of reads in both sets (Fig 2). Using Trimmomatic, data was first trimmed to 200 bases and any reads below that length were then discarded. Post-trimming quality control analysis through FastQC showed marked improvement in average read quality (Fig 3). After trimming, read count for paired sets had been reduced to 816,481 reads at a size of 200bp for each distinct file; total size for all reads totaled 326.57 Mbp. Compared to the 4.4 Mbp H37Rv reference, this reduced our read data to ~74x coverage from the initial, pre-trim coverage of 189x. Additional read sets (ENA Accessions: ERR2099776. ERR2099777, ERR2099778, ERR2099779, ERR2099780, ERR 2099781) were evaluated through FastQC but none provided a larger initial nucleotide count nor improved quality as estimated through sequence length distribution and phred score, implying similar coverage would result after trimming.



*Figure 2: Initial Quality Control of M. tuberculosis WGS Illumina data*. Quality scores across base positions for raw forward (left) and reverse (right) cryptically-resistant M. tuberculosis read sets. Red bar denotes average read quality at each position.



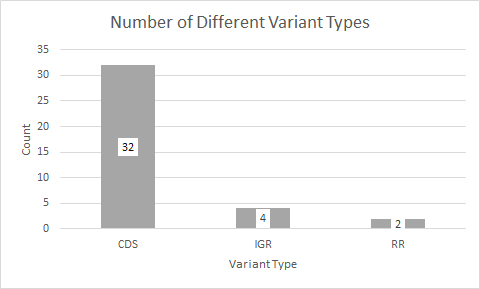
*Figure 3: Post-Trimming Quality Control of M. tuberculosis WGS Illumina data*. Quality scores across base position for trimmed forward (left) and reverse (right) cryptically-resistant M. tuberculosis read sets when cropped to 200bp and reads <200bp were discarded. Red bar denotes average read quality at each position.

*Read-to-Reference Mapping*

BWA mem was used to map the trimmed cryptic *M. tuberculosis* reads to the H37Rv reference genome. The BWA-generated .sam file was converted to .bam formatting, sorted, and then indexed into a .bai file using samtools.

*Call variant processing*

Bcftools mpileup was used to call positional information from the sorted .bam output for read data and H37Rv reference. The Bcftools call program was used to call variants with single ploidy using the multiallelic caller function (versus the previous version’s consensus caller). This output was then sorted for biological function using a perl script provided by Dr. Driscoll. A total of 33 variants were called using bcftools (Table 1; see Supplementary Data for full output table). Our pipeline only retrieved indels; no SNPs were called during variant call processing. Two variants occurred within repeat regions (RR) and an additional four within intergenic regions (IGR) (Fig 4). Of coding sequence variants, two were present within pseudogenes (Rv1667c and Rv1887a) and three occurred within hypothetical proteins; biological function is unknown. Variations within the coding regions (CDS) of PPE/PE-PGRS family proteins were the most abundant variant found in our analysis when parsed by biological function (11 of 33 total variants). All indels were examined for known variants at the identified position using the PolyTB web tool of archived genomic variations.



*Figure 4: Variant Type distribution between genomic features for cryptically-resistant M. tuberculosis indels.* Indels retrieved from the bcftools variant calling pipeline were present in three types of genomic features; Coding sequences (CDS), intergenic regions (IGR), and repeat regions (RR).

*PolyTB Analysis*

Positional information from the variants table was used to manually search the PolyTB web tool to investigate whether our called indels had been previously recorded within the database or if they had the potential to be a new variant. Results of this search have been recorded in Table 1. For those variants that were not recorded in PolyTB, we examined the biological function of CDS variants to estimate their potential to confer antibiotic resistance.

*Table 1. List of structural variants for cryptic M. tuberculosis reads to H37Rv reference variant calling.* Location and type of variant for bcftools variant calling output. Predicted products with the “|” character denotes the predicted flanking products of an intergenic variant. Feature corresponds to the feature present at the called position (Intergenic Region, Coding Sequence, or Repeat Region). Status of each variant (presence/absence) within the PolyTB tool noted in the final column.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variant: Predicted Product** | **Genomic Position** | **Feature** | **Variant Type** | **Quality** | **Depth**  **Score** | **PolyTB** |
| transmembrane protein | antitoxin VapB1 | 71584 | IGR | insertion | 228 | 46 | Not Recorded |
| maturase | 79504 | CDS | insertion | 228 | 62 | Not Recorded |
| oxidoreductase | 90888 | CDS | deletion | 129 | 39 | Not Recorded |
| PPE family protein (PPE6) | 374786 | CDS | deletion | 42.7581 | 64 | Not Recorded |
| 77bp Mycobacterial Interspersed Repetitive Unit | 580773 | RR | deletion | 228 | 38 | Known Var |
| PE-PGRS family protein (PE-PGRS7) | 672490 | CDS | deletion | 52 | 6 | Not Recorded |
| PE-PGRS family protein (PE-PGRS14) | 928158 | CDS | deletion | 14.7883 | 22 | Known Var |
| PE-PGRS family protein (PE-PGRS15) | 968426 | CDS | insertion | 130 | 50 | Known Var |
| PE-PGRS family protein (PE-PGRS21) | 1213679 | CDS | insertion | 123 | 15 | Not Recorded |
| hypothetical protein | 1428312 | CDS | deletion | 108 | 52 | Not Recorded |
| esterase LipO | 1602083 | CDS | deletion | 228 | 87 | Not Recorded |
| hypothetical protein | 1612278 | CDS | insertion | 64 | 112 | Known Var |
| 21 bp imperfect direct repeat | 1612624 | RR | insertion | 228 | 70 | Not Recorded |
| Rv1667c (pseudogene) | 1894300 | CDS | insertion | 163 | 66 | Known Var |
| hypothetical protein | 56 bp direct repeat | 1907459 | IGR | deletion | 228 | 100 | Not Recorded |
| Sec system translocase (SecA2) | CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase | 2068975 | IGR | deletion | 78 | 93 | Not Recorded |
| hypothetical protein | 2133468 | CDS | insertion | 228 | 69 | Known Var |
| Rv1887a (pseudogene) | 2137521 | CDS | insertion | 177 | 90 | Known Var |
| short-chain type dehydrogenase/reductase | 2180796 | CDS | deletion | 218 | 91 | Not Recorded |
| ubiquinol-cytochrome C reductase cytochrome subunit B | 2461325 | CDS | deletion | 89 | 45 | Not Recorded |
| ribonuclease Z | 2704884 | CDS | insertion | 228 | 80 | Not Recorded |
| hypothetical protein | 3096286 | CDS | deletion | 80 | 96 | Not Recorded |
| Possible MT-complex-specific genomic island | 3120523 | CDS | insertion | 22.1235 | 137 | Known Var |
| CRISPR-associated protein Cas10/Csm1 | 3131469 | CDS | insertion | 108 | 58 | Known Var |
| PE-PGRS family protein (PE-PGRS48) | 3163488 | CDS | deletion | 43.3335 | 18 | Not Recorded |
| transposase | 3194705 | CDS | deletion | 34.5857 | 76 | Not Recorded |
| glutaredoxin electron transport protein NrdH | hypothetical protein | 3415180 | IGR | deletion | 112 | 104 | Known Var |
| PPE family protein (PPE51) | 3502013 | CDS | deletion | 5.76602 | 100 | Not Recorded |
| PE-PGRS family protein (PE-PGRS50) | 3738212 | CDS | deletion | 103 | 12 | Not Recorded |
| PE-PGRS family protein (PE-PGRS50) | 3738516 | CDS | insertion | 125 | 9 | Not Recorded |
| PE-PGRS family protein (PE-PGRS51) | 3779671 | CDS | insertion | 16.7562 | 15 | Not Recorded |
| PE-PGRS family protein (PE-PGRS53) | 3928784 | CDS | insertion | 16.849 | 7 | Not Recorded |
| ESX-1 secretion-associated protein EspK | 4359135 | CDS | insertion | 14.0201 | 4 | Not Recorded |

***Discussion***

Only ten of our 33 called indels were recorded in the PolyTB tool database. The “cryptic” designation of our resistant strain implied that the mechanism behind the resistance mis-screen would not result from a known variation. Some of the potentially-novel variants within our sample occurred in repeat or intergenic regions and were less likely to cause functional effects, save for regulatory regions. Only 19 of the 33 indels called were both located in CDS regions and also not recorded within the PolyTB database. As the goal of this project was to identify potential sources of cryptic resistance and improve clinical diagnostic markers, we proceeded with biological investigation of potentially novel CDS indels.

Roughly 47% (9 of 19) of potentially novel coding-region indels were found in PPE/PE-PRGS family proteins. This family forms antigen proteins implicated in pathogenicity and antibody recognition.1 This family is also thought to play a role in fibronectin-binding, a biologically important consideration for cell wall permeability.2 By reducing permeability, fibronectin-binding proteins inhibit the ability of antibiotics to penetrate the cell wall. When antibiotics fail to accumulate within the cytosol quickly enough, division can proliferate the bacterial infection at a rate faster than drug inhibition. Several *M. tuberculosis* fibronectin-binding proteins (FpbA, FpbB, FpbC) are known to play a role in reduced antibiotic sensitivity and are regarded as important to- but not solely responsible for- resistance mechanisms.3 Insertions and deletions within these gene coding regions could be deleterious for the protein product. Disruption of fibronectin-binding activity should, in theory, increase cell wall permeability and reduce any antibiotic resistance conferred through this mechanism. However, changes to genes in this protein family that code for antigens and other external morphology could increase antigen diversity and reduce potential for external binding; this might serve as a mechanism to increase resistance and confound tuberculosis detection.

PolyTB already recorded potentially deleterious mutations such as non-synonymous SNPs, insertions, and deletions within the PPE/PE-PRGS genes of various *M. tuberculosis* strains including 2 out of the 11 total PPE/PE-PRGS variants called by our analysis. While disruption of these genes might not provide novelty in mechanism of action, some variants may provide positional novelty. When using sequence-based markers for resistance screening, positional novelty could create false negatives and confound treatment, as observed in the clinical case that produced this WGS Illumina data. Whether these positional novelties are contributors to antibiotic resistance itself needs to be confirmed in future functional analyses.

Other CDS indels did not seem to be likely candidates for conferring resistance. While *espK* is associated with the ESX-1 secretion system and *M. tuberculosis* virulence, we did not find associations to antibiotic mode of action or resistance.4 The potentially novel PPE/PE-PRGS variants detected by our analysis must be investigated further before we can comment on their relevance as antibiotic resistance markers. Additional sequencing will be conducted at these loci to ensure variants were called correctly, especially for variants with low quality and depth scores. Furthermore, we will examine these locations within other assembled antibiotic-resistant *M. tuberculosis* genomes to determine whether our unrecorded variants are truly novel or were simply not yet recorded in the PolyTB resource. If contribution to resistance cannot be confirmed through other resources, functional analyses and complementation assays will be conducted to examine their impact upon antibiotic susceptibility. If the biologically-relevant variations are novel and rare, screening incorporation may not be informative unless the strain proliferates in the population. If this variation is found to persist in many antibiotic-resistant strains (and not susceptible lines), consideration as an additional marker may prevent additional false negatives in clinical environments.

It is possible that rare structural variants and SNPs remained undiscovered in our analysis due to the low coverage of our trimmed read set. Future work will utilize the serial isolate sequencing generated within this BioProject to evaluate the genomic variations within the population of cryptic *M. tuberculosis* samples. Increasing total read-to-reference coverage by mapping serial read sets should improve the power of variant detection and further reduce any chance of confounding polymerase-generated errors, while also allowing for population-level analyses in follow-up studies.

*Troubleshooting*

Difficulty was encountered in converting the sam output to bam using samtools; an error kept occurring during file conversion that suggested a truncated sam file or improper/missing header. This error was a result of the need to specify a sam input; without the ‘S’ addition to the ‘-b’ option argument, samtools was interpreting the input .sam as a bam file, ignoring file ending, and thus couldn’t reconcile the header. Syntax issues were encountered during sorting and indexing of the resulting bam file; reducing the arguments and options to the default functions for samtools v0.1.19 corrected these problems.

The variant calling portion of our pipeline was modified due to errors encountered during GATK script coding and submission. An array of errors were encountered; most commonly was a memory-associated VM error, a missing -jar file, and an inability to activate the Java environment needed to run the program despite an up to date version on the bioconda environment. GATK is also more suited to analysis of eukaryotic systems; we attempted to use gatk HaplotypeCaller with ploidy specification to circumvent this but in light of complications, we decided to make better use of our time and analysis environment by selecting an alternate variant calling program. Bcftools was able to call variants in a prokaryotic genomic environment with fewer alterations and less troubleshooting steps. The only issue encountered was in variant options during bcftools call; without the ‘-v’ argument, bcftools call produced a shortened table of non-variants.

*Learning Outcomes*

Coding difficulty increased this project compared to the last during the samtools and bcftools pipeline steps. Our group faced very few issues with syntax and arguments in Project 1; Project 2 made us troubleshoot code at multiple steps to overcome pipeline issues. This was an important exercise in triple checking version-specific arguments and options, especially with Samtools v0.1.19, where arguments had changed dramatically between versions. In future projects, we will rely more heavily on version-specific manuals rather than trusting code structure from other pipelines. Heavy queue traffic also required us to optimize wall time and queue choice to complete jobs in a timely matter. Familiarizing ourselves with the Spruce HPC helped us improve our efficiency in working through our analysis pipeline.

***Pipeline Code***

**##FASTQC**

#! /bin/bash

module load genomics/bioconda

source /shared/software/miniconda3/etc/profile.d/conda.sh

conda activate tpd0001

cd $SCRATCH

# Initial QC

#

fastqc TBreads\_1.fastq.gz TBreads\_2.fastq.gz

fastqc TBinterleaved.fastq.gz

#

conda deactivate

**##TRIMMOMATIC**

#! /bin/bash

module load genomics/bioconda

source /shared/software/miniconda3/etc/profile.d/conda.sh

conda activate tpd0001

cd $SCRATCH

#Trimming and FastQC

#

trimmomatic PE -threads 6 -trimlog TBtrim TBreads\_1.fastq.gz TBreads\_2.fastq.gz -baseout TBtrim.fastq.gz CROP:200 MINLEN:200

fastqc TBtrim\_1P.fastq.gz TBtrim\_2P.fastq.gz

#

conda deactivate

**##BWAmem1**

#! /bin/bash

module load genomics/bioconda

source /shared/software/miniconda3/etc/profile.d/conda.sh

conda activate tpd0001

cd $SCRATCH

# BWA mem attempt one

#

# indexing reference genome

bwa index H37Rv.fna

#mapping reads to reference

bwa mem -t 12 -c 50000 -P -B 3 H37Rv.fna TBtrim\_1P.fastq.gz TBtrim\_2P.fastq.gz > TB3.sam

#sam to bam conversion and output sort

samtools view -bS -o TB3.bam TB3.sam

samtools sort TB3.bam TB3.sort

#indexing sorted bam file

samtools index TB3.sort.bam

conda deactivate

**##bcftools variant calling**

#

#! /bin/bash

module load genomics/bioconda

source /shared/software/miniconda3/etc/profile.d/conda.sh

conda activate tpd0001

cd $SCRATCH

#variant calling and file conversion

bcftools mpileup --threads 12 -6 -A -Ob -o TB3.bcf -f H37Rv.fna TB3.sort.bam

bcftools call -v --ploidy 1 --threads 12 -A -m -o TB3v.vcf TB3.bcf

#sorting vcf output with genomic information

./vcf2table.pl H37Rv.gff -g H37Rv.fna < TB3v.vcf > TB3v.var\_table.txt

conda deactivate

————————————————-

***Supplementary Data***

See the class github repository at /wvu\_gda/gitgud/TB3v.var\_table.txt for the full called variant table.

**Additional References**

1. Brennan, M. J. (2017). The enigmatic PE/PPE multigene family of mycobacteria and tuberculosis vaccination. *Infection and immunity*, *85*(6), e00969-16.
2. Espitia, C., Laclette, J. P., Mondragón-Palomino, M., Amador, A., Campuzano, J., Martens, A., ... & Moreno, C. (1999). The PE-PGRS glycine-rich proteins of Mycobacterium tuberculosis: a new family of fibronectin-binding proteins?. *Microbiology*, *145*(12), 3487-3495.
3. Nguyen, L. (2016). Antibiotic resistance mechanisms in M. tuberculosis: an update. *Archives of toxicology*, *90*(7), 1585-1604.
4. McLaughlin, B., Chon, J. S., MacGurn, J. A., Carlsson, F., Cheng, T. L., Cox, J. S., & Brown, E. J. (2007). A mycobacterium ESX-1–secreted virulence factor with unique requirements for export. *PLOS pathogens*, *3*(8), e105.