Supplemental Evaluation of the Bioinformatics Pipeline

1. Summary

The validation work focused on ensuring the reliability and accuracy of bioinformatics methods used. This involved:

1. A comparative analysis between the Sentieon aligner and the existing BWA-MEM aligner.
2. An examination of synthetic mixture samples to determine the mixture ratios, at varying coverage depths, that can produce usable genotype files.
3. A comparison of different filtering thresholds for SNP filtering to optimize the creation of genotype files.

Conclusion:

The supplemental bioinformatics validation has demonstrated that:

1. The Sentieon aligner is comparable to the existing BWA-MEM aligner in terms of endogenous DNA percentage, duplicate rate, and the identification of FTDNA SNPs.
2. The pipeline is capable of accurately producing genotype files for mixture ratios up to 9:1, across various coverage levels.
3. While usable files can be produced down to 0.25X coverage, filtering thresholds have a significant impact on matching performance, especially at lower coverage levels.

The following observations were made on these methods:

Comparison of Sentieon and BWA-MEM

The comparison involved 8 samples and focused on three key metrics: the average endogenous DNA percentage, duplicate rate, and the number of FTDNA SNPs with a phred-scaled genotype posterior probability (GP) of at least 0.90.

- Average Endogenous DNA Percentage: Both Sentieon and BWA-MEM showed an identical average of 41.1%.
- Average Duplicate Rate: Both aligners also had an identical average duplicate rate of 48.3%.
- Average Number of FTDNA SNPs (GP >= 0.90): The average number of SNPs was very close, with BWA-MEM at 528,025 and Sentieon at 528,023.

These results indicate that Sentieon and BWA-MEM are comparable in performance across the evaluated metrics.

Determination of Usable Mixture Ratios

The analysis of synthetic mixture samples revealed that mixtures up to a ratio of 9:1, at all sequencing depths, provided usable uploads and adequate matching results. This was supported by data showing that even at a 9:1 mixture ratio, there were no significant false negatives or positives at the top 25 level. False negatives were defined as top 10 control matches that were not found in the top 25 sample matches. False positives were defined as top 10 sample matches that were not found in the top 25 control matches. The top 10 total cM (centimorgans) difference was also minimal, indicating reliable files were generated for various coverage levels, up to this 9:1 mixture ratio.

Analysis of Filtering Thresholds

The validation of filtering thresholds showed that usable files were produced with all combinations of filtering thresholds down to 0.5X coverage. However, at the 0.25X coverage level, not all files succeeded, and using more

strict GP filtering resulted in files that were deemed too "matchy," indicating an excessive number of matches. At 0.5X coverage, greater differences in the shared cM of matches (versus control) were observed, which became even more pronounced at the 0.25X coverage level. The following patterns were observed:

1. As the GP filtering threshold decreased, the total cM and average cM differences between control and sample increased. This indicates the lower the GP threshold, the greater the matching performance issues.
2. The lower the RR threshold, the allele frequency at which homozygous reference SNPs are included/excluded, the higher the number of SNPs. However, total matches also increased, indicating excessive matchiness.
3. For each RR threshold, there was an approximate # of SNPs below which samples failed batching and were labeled “matchy” by GEDmatch

To limit the number of false matching segments, 0.10 was determined to be the optimal RR filtering value. It was determined the GP filtering threshold should be done in a stepwise manner.

1. If a file with >500k SNPs can be produced using a GP threshold of 0.90 or higher, the highest value satisfying this requirement should be used.
2. If this is not possible, the highest GP threshold value, 0.70 or above, that produces a file of >470k SNPs should be used.
3. If this is not possible, the highest GP threshold value, 0.70 or above, that produces a file of >460k SNPs should be used.

It was also discovered that a number of kits were failing to upload due to GEDmatch due to an “HTZ string too long” error. To meet the requirements for upload, the datafile was optimized to meet the database’s needs by first removing all SNPs on ClinVar that were not labeled benign, before further downstream processing. These SNPs are potentially medically relevant SNPs that have been added to ClinVar, a “public archive of reports of human variations classified for diseases and drug responses”. The results for a subset of samples that did succeed in uploading were compared before and after this change. The total number of matches decreased by 1.1%, and each of the top 10 matches showed the same total cM shared before and after the change. Therefore, the change was determined to not have a significant effect on matching performance.

3. Validation Components

Bioinformatics Pipeline (Data Analysis)