BLG 348E Term Project

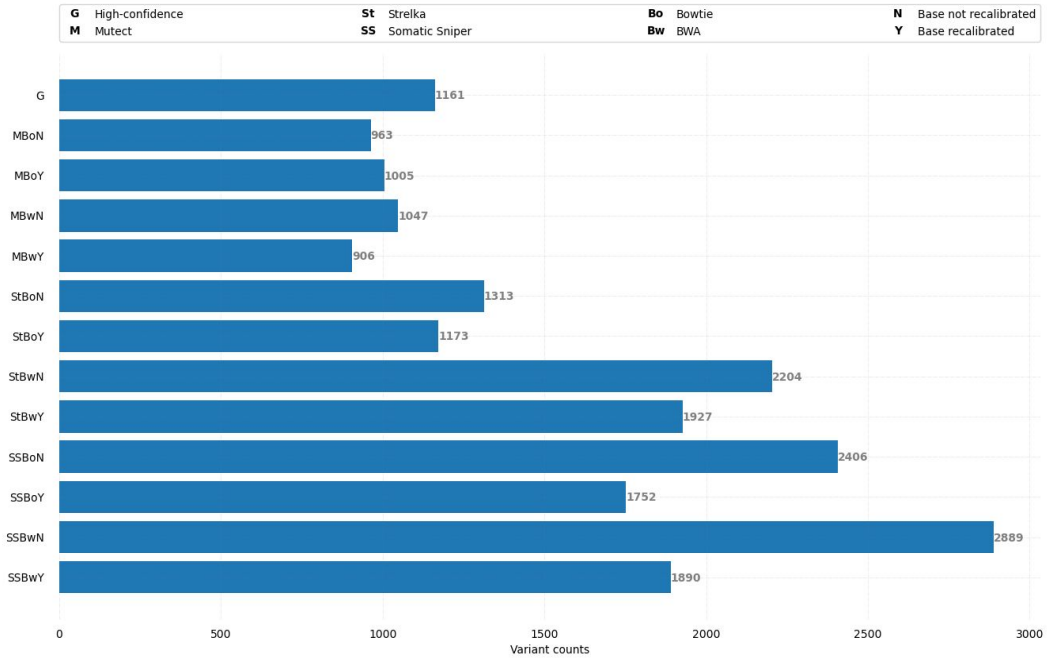
150190802 : ÖZCAN ANBALAY
150210921 : DUC QUANG NGUYEN

SUMMARY

- **Objective:** To evaluate the effects of different aligners and variant calling algorithms on variant detection.
- **Genome sets:** SRR7890850 and SRR7890851, sourced from the breast cancer cell line HCC1395 and its B lymphocyte-derived normal counterpart HCC1395BL.
- **Mappers:** Bowtie/BWA
- **Variant callers:** Mutect/Strelka/Somatic Sniper
- **Base recalibration:** Yes/No
- **Total:** 12 distinct pipelines
- **Machine stats:** Windows 11, 32GB RAM, >200GB storage space.
- **Completion time:** 5 days

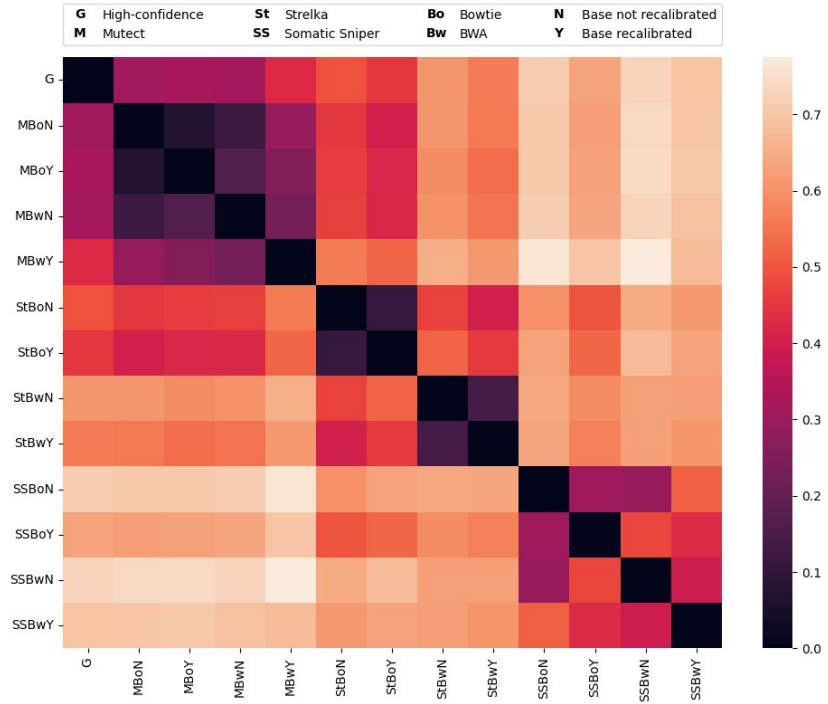
Findings

Variant counts among pipelines



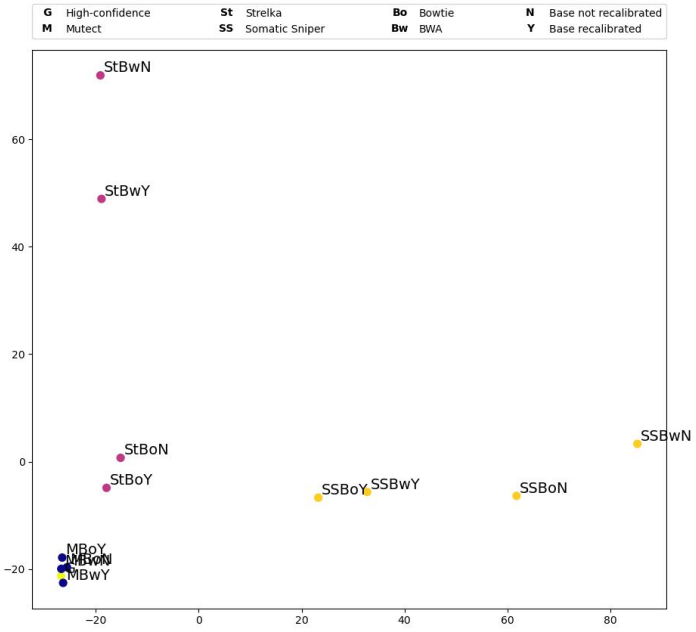
Variations in variant counts among 12 pipelines and the gold standard

Convergence of pipelines in detected variants



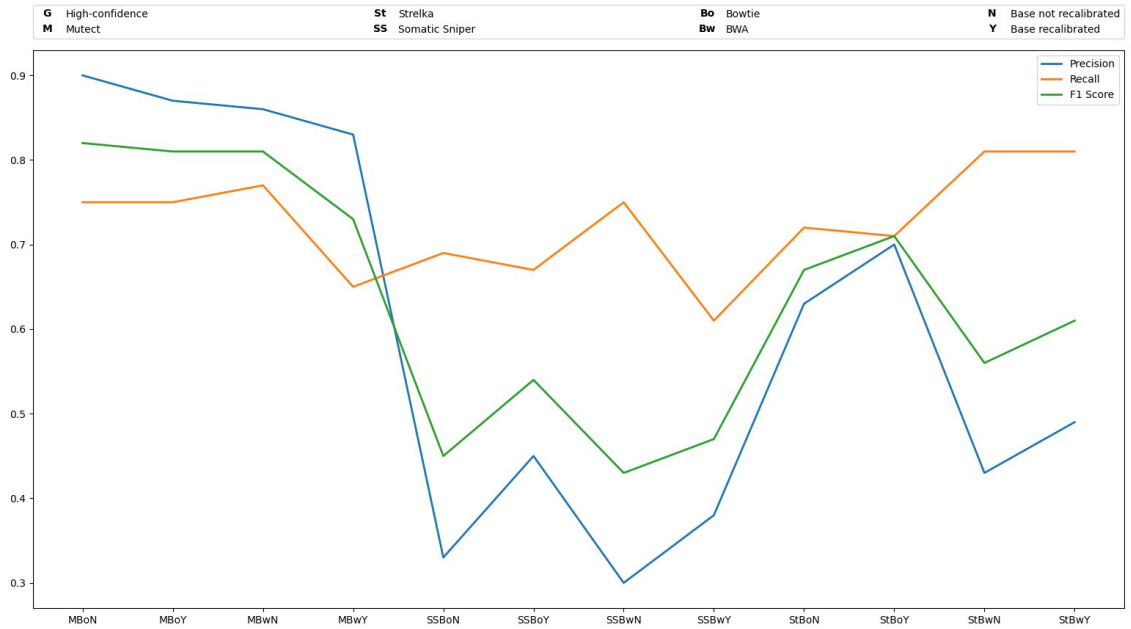
Convergence of pipelines with respect to detected variants, measured by Jaccard distances between any two pipelines

Convergence of pipelines in detected variants



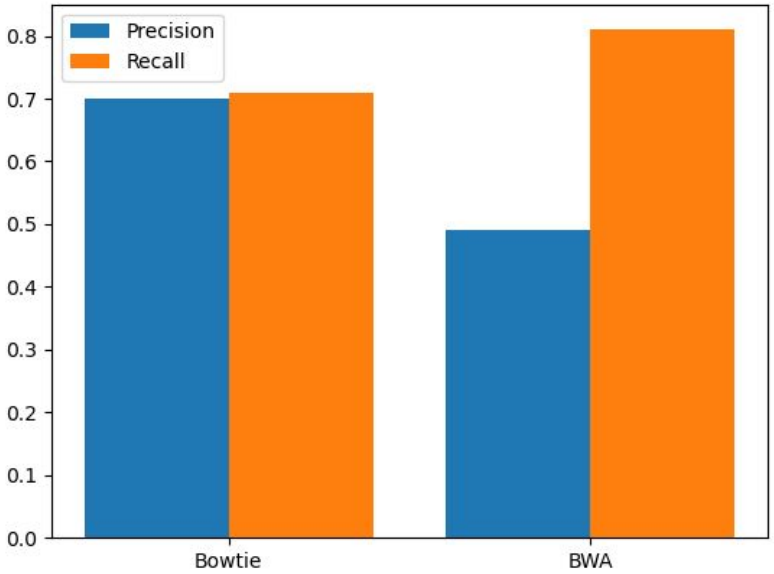
PCA of all pipelines with respect to variant calls

Precision, Recall, and F1 Score

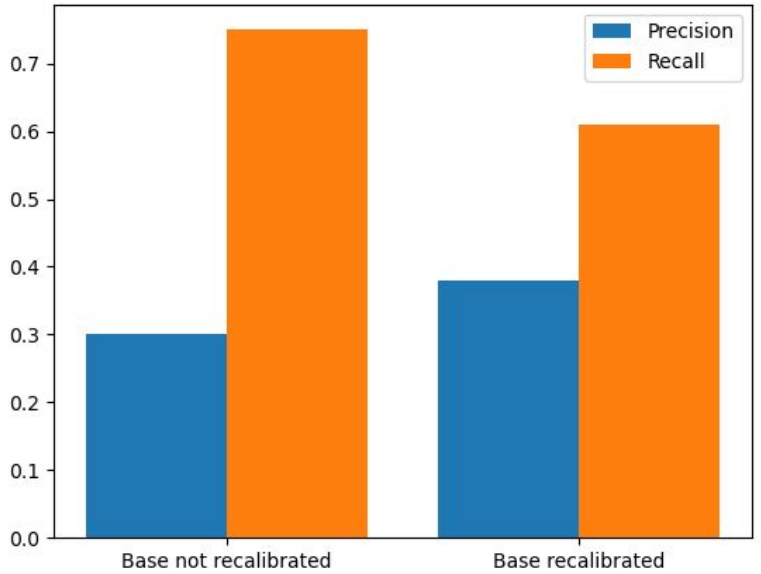


Performance with respect to precision, recall, and F1 score

Precision - Recall tradeoff



Two base recalibrated Strelka pipelines

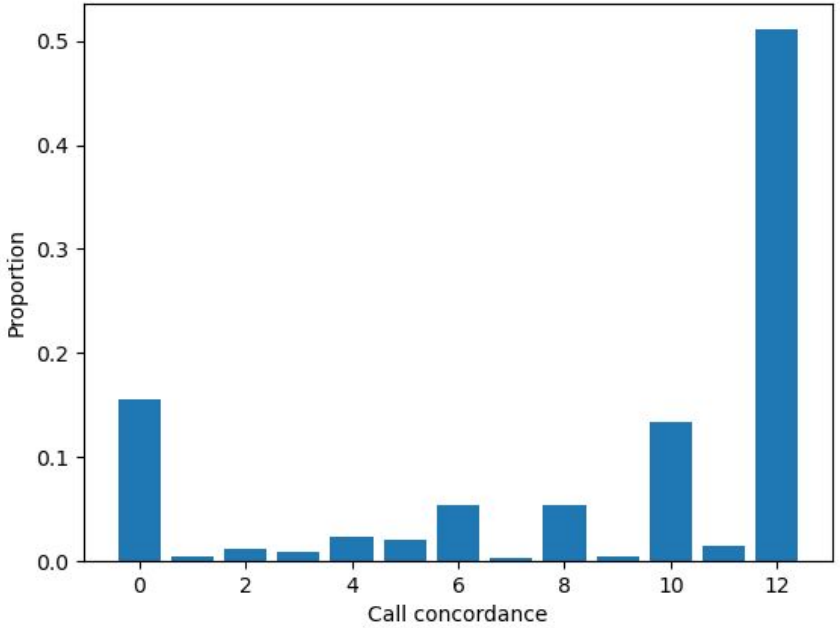


Two BWA-Somatic Sniper pipelines

Precision - Recall tradeoff

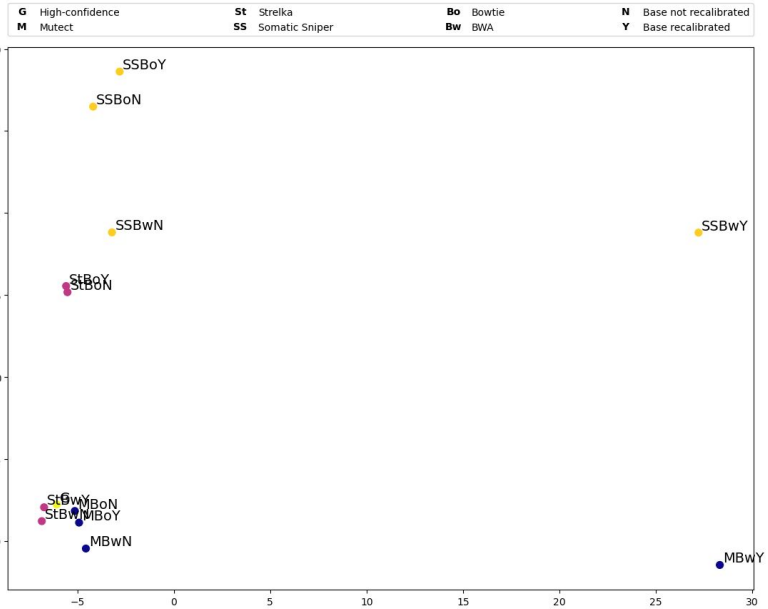
Precision - Recall curve with respect to variant counts of pipelines

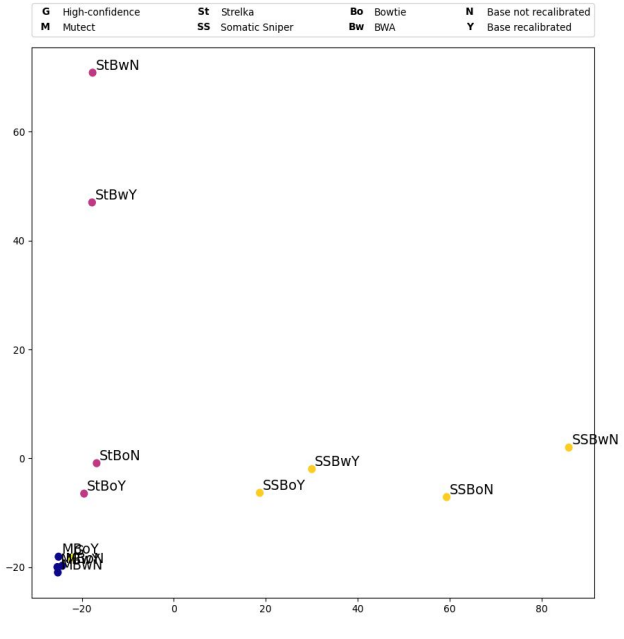
False negatives - False positives



Call concordance of pipelines

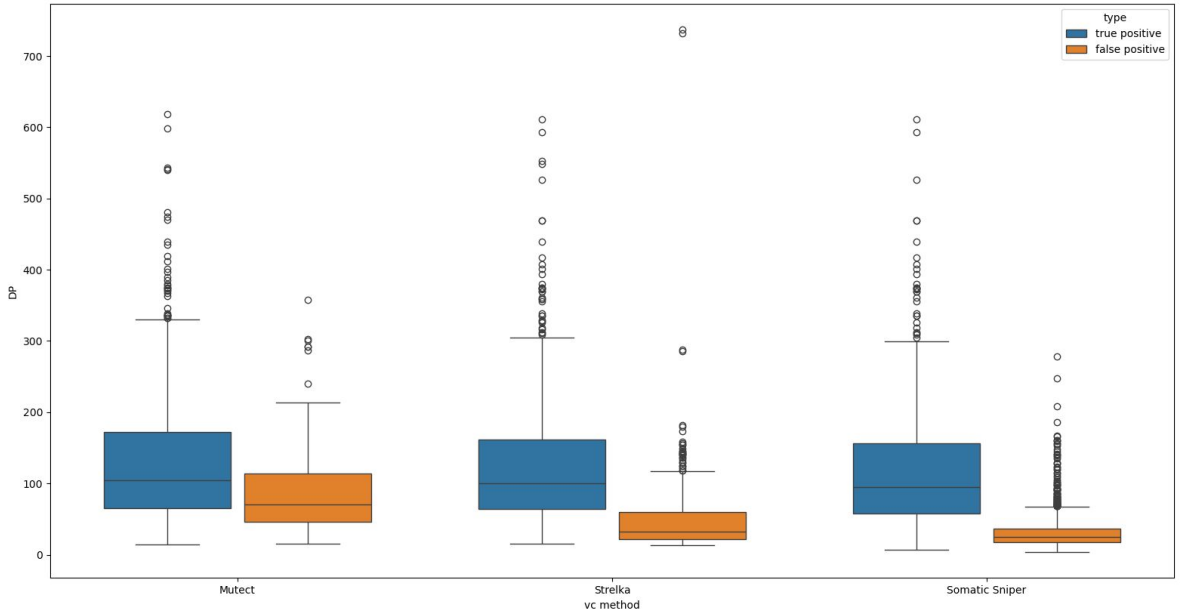
False negatives - False positives





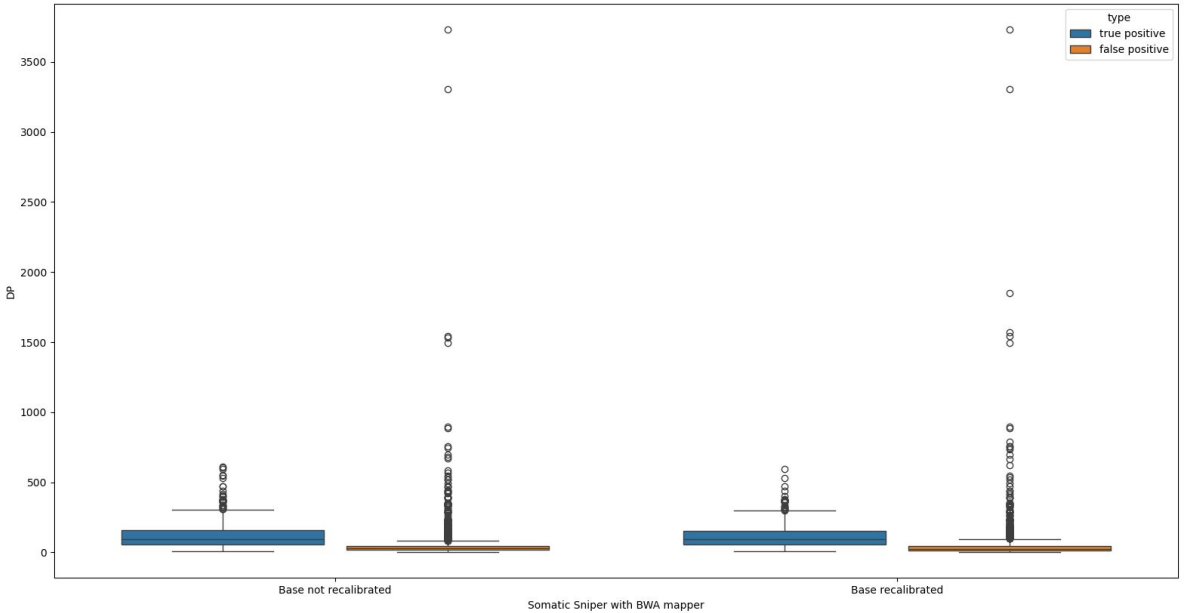
PCA of pipelines with respect to (left) false negatives and (right) false positives

Read depth and detection accuracy



Comparing read depth between true positives and false positives across three pipelines of different callers

Read depth and detection accuracy



Presence of variants with exceptional high read depth in pipelines running Somatic Sniper and BWA mapper

Key takeaways

- Most variations in variant outputs, precision, recall, and F1 score can be attributed to variant callers. Mutect has the best performance among the three.
- The choice of aligners and base recalibration affects precision and recall to certain extent.
- There is precision - recall tradeoff with respect to the number of variant outputs.
- Over half of gold standard variants are detected by all 12 pipelines. However, 15% of gold standard variants are detected by no pipelines.
- Higher read depth has a positive impact on detection accuracy. Nevertheless, exceptionally high read depth seems to be associated with false positives.