

Optimizing clinical next-generation sequencing workflows through evidence-based expansion of statistical confirmation criteria

Jared Evans MS, Rebecca Mar-Heyming PhD, Marcus Mahar PhD, Katherine Shortt PhD



Background

Variant confirmation by statistics is an established method to optimize clinical next-generation sequencing (NGS) workflows, but criteria are often conservatively defined. Strategically routing a subset of variants for orthogonal confirmation allows the primary variant calling pipeline to remain highly sensitive and specific, while providing faster results to patients through a reduction in confirmatory testing.

We evaluated clinical confirmation data and ground truth samples from hereditary cancer multi-gene panels (CancerNext-Expanded), exomes (ExomeNext), and whole-genome sequencing (GenomeNext) from a credentialed diagnostic laboratory to build upon our existing validation framework and determine the accuracy of expanding statistical thresholds across various variant types and quality metrics.

Methods

Benchmark Data

A combination of 66K Sanger confirmed clinical variants and 122K published benchmark variants were used to evaluate the performance of confirming variants by statistics. The published benchmark data came from 7 Genome In A Bottle (GIAB) samples (HG001-HG007). Autosomal variant calls were compared to GIAB v4.2.1[1] variants for all samples and HG002 hemizygous variants on chromosome X were compared to a separate GIAB XY v1.0 benchmark dataset [2]. Homozygous reference genotypes were generated by performing joint-genotyping of the HG002 trio.

Variant concordance was performed in the exonic (+/- 20bps) region for 3,614 genes of interest. The genes were intersected with the GIAB benchmark regions, which covered 99% of the gene regions. Pseudoautosomal region (PAR) genes were excluded from Hemizygous analysis. Concordance was calculated with RTG Tools vcfeval (v3.13).

Machine Learning

Machine Learning models were trained and evaluated for their ability to classify TP/FP variants. ML results were used for identifying top quality metrics, as well as finding the upper limit of filtering performance for a set of quality metrics.

Variants were split using an 80/20 train/test approach. Due to the large number of recurring variant sites, the train and test datasets were created to include different unique variant sites to prevent data leakage. SNVs and INDELs were trained separately because of their differing quality score distributions. Decision thresholds were adjusted to maximize precision to reduce potential false positives being statistically confirmed. Feature Importance analysis was performed on the ML models to help inform whether new quality scores should be considered. The scikit-learn Python package was used for model training and evaluation.

Results from two different ML methods were compared:

- XGBoost: A highly accurate, fast implementation of gradient boosting for decision trees. More complex to interpret feature importance.
- Logistic Regression: Simpler binary classification method based on linear relationship between features and log-odds. Lower accuracy, but can be more straightforward to interpret feature importance.

Hard Thresholds

Quality metrics used for the hard thresholds were selected from the feature analysis of the machine learning models. Specifically, the logistic regression models were used due to the more straightforward interpretation of the coefficient values compared to the xgboost decision points. Each hard threshold value was maximized for precision and then evaluated in combination with other metrics.

References

1. Wagner J, Olson ND, Harris L, et al. Benchmarking challenging small variants with linked and long reads. Cell Genom. 2022;2(5):100128.
2. Wagner J, Olson ND, McDaniel J, et al. Small variant benchmark from a complete assembly of X and Y chromosomes. Nat Commun. 2025;16(1):497.

Results

The following figures and tables were generated from hereditary Cancer Panel data using Sanger confirmed variants. A similar analysis was performed for Exome and WGS data.

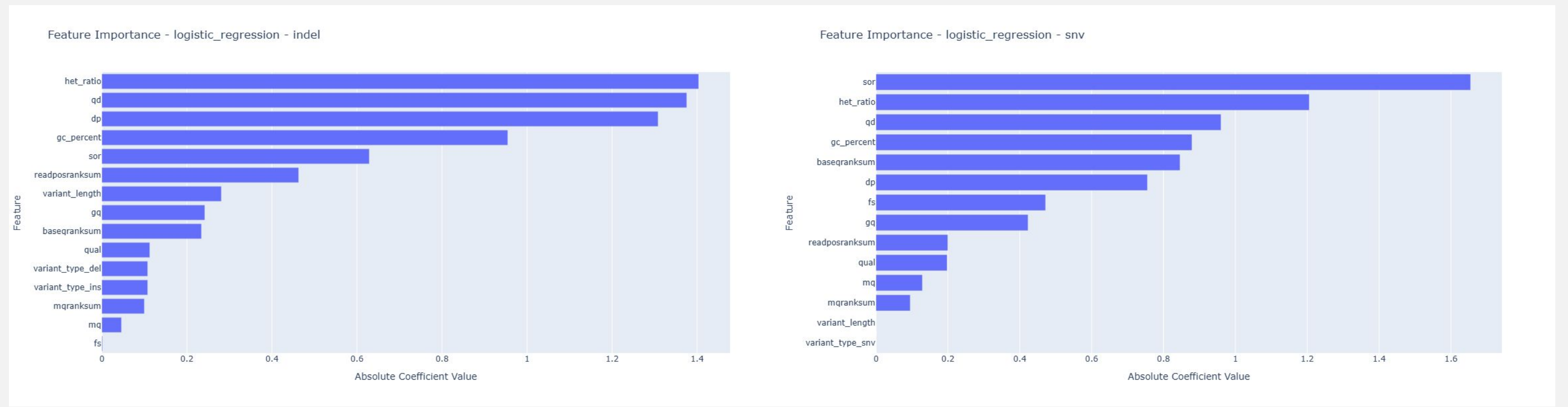


Figure 1. Feature importance analysis from the Logistic Regression models for SNVs (left) and INDELs (right). Top metrics were selected for incorporating into statistical confirmation criteria.

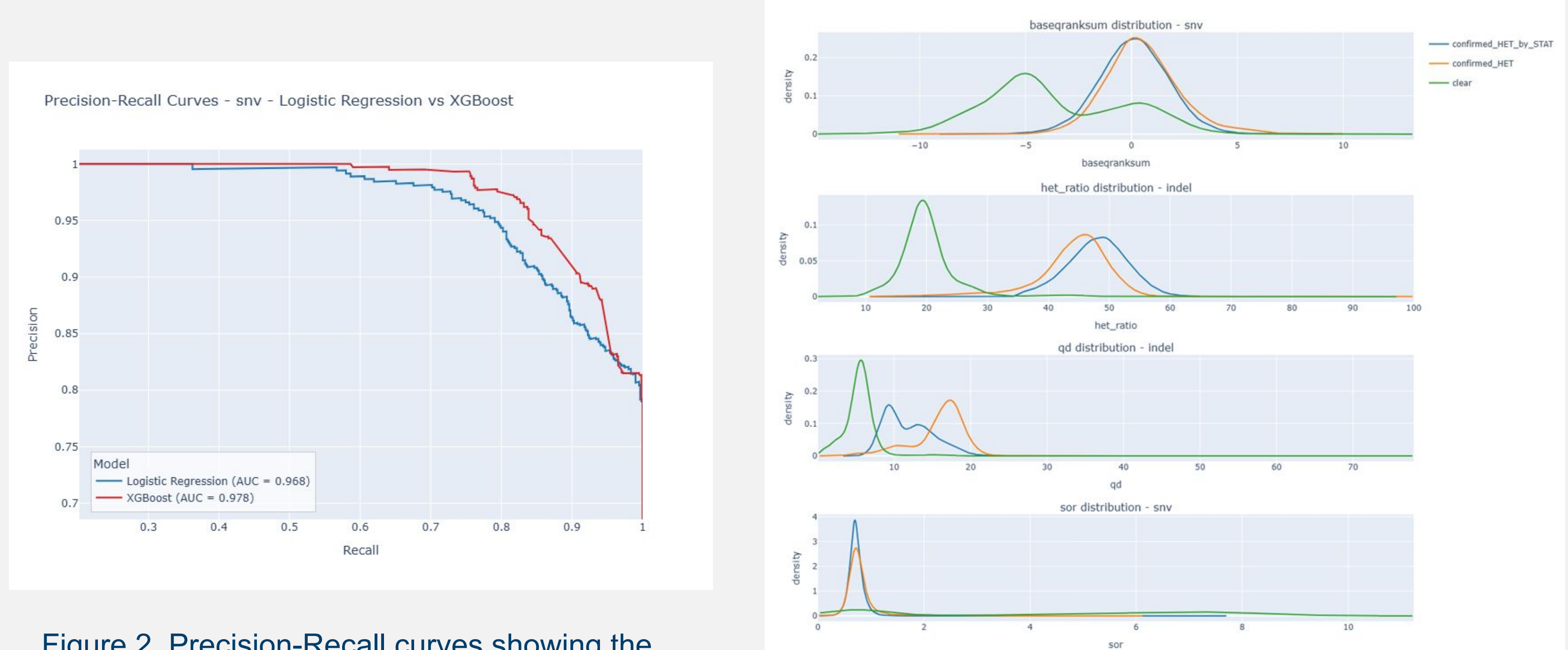


Figure 2. Precision-Recall curves showing the performance of the XGBoost and Logistic Regression machine learning models on a hold-out SNV test dataset.

Figure 3. Distribution of top quality scores identified from the machine learning feature analysis. Most have noticeable separation between the variants confirmed heterozygous (confirmed_HET) vs confirmed wildtype (clear) with Sanger sequencing.

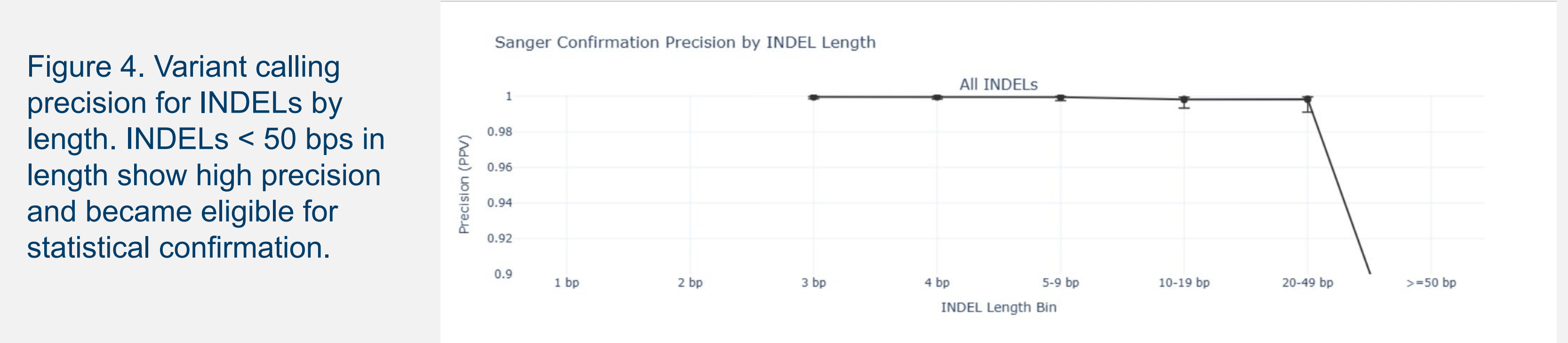


Figure 4. Variant calling precision for INDELs by length. INDELs < 50 bps in length show high precision and became eligible for statistical confirmation.

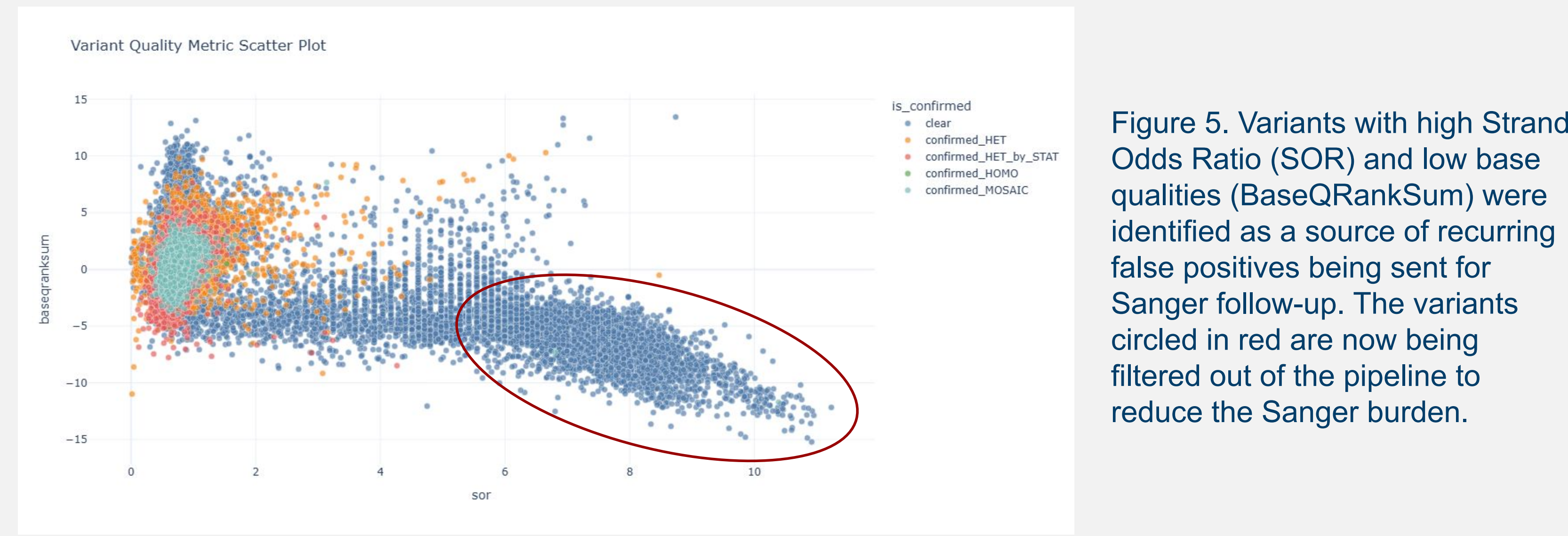


Figure 5. Variants with high Strand Odds Ratio (SOR) and low base qualities (BaseQRankSum) were identified as a source of recurring false positives being sent for Sanger follow-up. The variants circled in red are now being filtered out of the pipeline to reduce the Sanger burden.

Discussion

Machine Learning Informing Hard Thresholds

The machine learning models were used to inform hard threshold adjustments within our current statistical confirmation system. They were valuable in helping quickly identify top quality metrics and variant types to focus on for Sanger reduction. The feature importance analysis was able to quickly narrow down a large list of potential quality attributes. A precision exceeding 0.999 was achieved after using that information to expand the ranges of previous quality metrics and adding new metrics.

Quality Metrics

- The following quality metrics are being used in our expanded statistical confirmation criteria.
- Total Depth (DP)
 - Heterozygous Read Ratio (Het_Ratio)
 - Strand Odds Ratio (SOR)
 - Base Quality Rank Sum (BaseQRankSum)
 - Genotype Quality Score (GQ)
 - Mapping Quality (MQ)

Challenging Genes

Genes and regions of the genome that are difficult to sequence with short read technology can produce false positives that aren't easily filtered out with quality filters. This can be due to pseudogene interference or low complexity repeat sequences. Variants from the following regions were excluded from statistical confirmation due to the high risk of false positives, ensuring they will be confirmed with Sanger sequencing instead.

- PMS2 exons 11-15, FLG exon 3, SMAD4 INDELs
- Low complexity INDELs (homopolymer, dinucleotide, trinucleotide repeats)

Sanger Reduction

Our analysis demonstrated that refining multiple confirmation criteria safely and significantly reduces the Sanger sequencing burden. The following is a summary of the reduction in Sanger sequencing achieved from 3 different assays.

- Cancer Panel
 - 31% reduction by increasing the length of INDELs eligible for statistical confirmation from < 3 bps to < 50 bps
 - 9% reduction from filtering out recurring false positive variants and strand bias artifacts
- Exome
 - 17% reduction by allowing variants with an expanded range of heterozygous read ratio of 35-65% to be confirmed by statistics
 - 12% reduction by increasing the length of INDELs eligible for statistical confirmation from < 3 bps to < 50 bps
- WGS
 - A new assay, with modeling informed by clinical panel and exome data. Benchmark and validation data show the newly implemented criteria will statistically confirm >95% of variants in non-excluded regions.

Conclusions

- Our evidence-based expansion of statistical variant confirmation criteria was able to reduce previous Sanger confirmation rates by ~40% for hereditary cancer panels and ~30% for whole exome assays
- Allowing larger INDELs to be confirmed by statistics along with filtering recurring false positive and strand bias artifacts had the most impact on Sanger reduction
- The machine learning models were valuable in quickly identifying top metrics and high-precision variant subsets to target with expanded statistical confirmation thresholds
- Exceptions to statistical confirmation are still necessary for challenging regions, such as low complexity sites and genes with pseudogene interference