Machine learning identifies high confidence variants in NGS

Gilad Mishne, Ph.D., Abdimalik Khalif, B.S., Daniel De Sloover, B.A.Sc., Kurt Smith, Ph.D., Annette Leon, Ph.D., FACMG, Jeroen Van Den Akker, Ph.D., MB (ASCP)

Introduction

Confirmation of next-generation sequencing (NGS) results by an alternative technology, such as Sanger sequencing, is currently recommended for all clinical tests in order to prevent false positive calls. Secondary confirmation adds significant time and cost, and impacts turn-around times and patient care. However, recent findings by Beck et al. [1] and Baudhuin et al. [2] indicate that the quality of modern NGS assays is on par with Sanger sequencing, and that there is limited utility in using it for confirmation. In addition, studies by Strom et al. [3] and Mu et al. [4] have demonstrated that the need for Sanger confirmation can be reduced by restricting it to low-quality variant calls, although the use of a single quality value to determine the quality of the call cannot rule out the need for confirmation. The analysis described here assesses the technical feasibility of combining multiple quality signals of a variant call using machine learning to reliably identify variant calls of high confidence, that are therefore expected to confirm.

Methods

A set of 5,318 Single Nucleotide Variants (SNVs) and Indels was detected by the Color Test, a 30-gene panel NGS genetic test for hereditary cancer risk [5], and subsequently re-assessed by Sanger sequencing. The bioinformatics analysis pipeline aligned reads against GRCh37.p12 with the Burrows-Wheeler Aligner [BWA-MEM], and called variants using the GATK3 HaplotypeCaller module. Coverage requirements for variant calling were a minimum of 20 unique reads (20X) for each base of the reportable range, and at least 50X for 99% for the reportable range. Median coverage was in the 200-300X range. All variants included in this analysis were classified according to ACMG guidelines as VUS, likely pathogenic, or pathogenic.

For each variant, multiple quality signals of the call and of the genomic position were collected; the features are summarized in Table 1. The data were then used to train a logistic regression model that estimates the probability that a given variant called as detected by NGS will subsequently be confirmed as present using Sanger sequencing.

Results

In our context, a *false positive prediction* is a variant detected by NGS that was predicted to be high-quality, but was not detected by Sanger sequencing. A *false negative prediction* is an NGS variant that was predicted to require Sanger confirmation due to its low-quality, but for which the Sanger actually identified the variant as present. With the task at hand, false positive predictions are significantly more costly than false negative predictions: while a false negative introduces some delay to completing the analysis, a false positive can lead to an incorrect result reported to a patient, if Sanger sequencing was not used to test and remove the low-quality NGS variant. As such, our model was tuned to eliminate false positive predictions. Using 10-fold cross-validation, the model achieved 99.1% accuracy (95% confidence interval: +/- 0.6%), see Figure 2. More importantly a 0% false positive prediction rate; a confusion matrix is shown in Table 2.

Conclusions

Our results demonstrate that a model defining the quality of NGS calls using features of the call site and the sequencing process can be effective in identifying high-confidence variants for which confirmation is not necessary. This approach can help to differentiate real variants from potential noise in many high-throughput NGS workflows as well as reduce cost and TAT of NGS based clinical tests.

*Color continues to use secondary confirmation of all clinically significant variants.

References

[1] Beck, Tyler F., James C. Mullikin, NISC Comparative Sequencing Program, and Leslie G. Biesecker. 2016. "Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants." *Clinical Chemistry* 62 (4): 647–54.

[2] Baudhuin, Linnea M., Susan A. Lagerstedt, Eric W. Klee, Numrah Fadra, Devin Oglesbee, and Matthew J. Ferber. 2015. "Confirming Variants in Next-Generation Sequencing Panel Testing by Sanger Sequencing." *The Journal of Molecular Diagnostics*: *JMD* 17 (4): 456–61.

[3] Strom, Samuel P., Hane Lee, Kingshuk Das, Eric Vilain, Stanley F. Nelson, Wayne W. Grody, and Joshua L. Deignan. 2014. "Assessing the Necessity of Confirmatory Testing for Exome-Sequencing Results in a Clinical Molecular Diagnostic Laboratory." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 16 (7): 510–15.

[4] Mu, Wenbo, Hsiao-Mei Lu, Jefferey Chen, Shuwei Li, and Aaron M. Elliott. 2016. "Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing." *The Journal of Molecular Diagnostics: JMD* 18 (6): 923–32.

[5] Zhou, Alicia Y., Jeroen Van Den Akker, Kurt. A. Smith, Robert O'Connor, Serra Kim, Daniel E. DeSloover, Tom Walsh, Elad Gil, Taylor Sittler. "Frequency of mutations in multi-gene panel testing of 3,600 individuals for hereditary breast and ovarian cancer risk." *ASHG 2016*.

Relevant sequence characteristics and quality signals in the variant confidence model

Similarly to previous observations on similar datasets [3, 4], we found that setting a threshold on a single quality score is insufficient for distinguishing high confidence from low confidence calls. Although some signals are more indicative than others, a combination is required to achieve a high-confidence model. A benefit of the features used in the developed model is their availability: the signals are directly accessible in the output of alignment (e.g., read depth), variant calling (e.g., quality scores), or can be computed directly from the reference genome (e.g., GC content). There are no secondary models required, and similar models can be developed for any sequencing assay in a straightforward fashion.

An examination of the features shows that call-specific signals (such as allele frequency and call quality) are more predictive than site-specific signals (such as GC content and presence of homopolymers): a breakdown for several features is shown in Figure 1. The top features that are most predictive of the quality of the call are the allele fraction, mapping quality, depth, call quality, and GC content; using just these features, a model that achieves over 96% accuracy can be constructed, but its false positive prediction rate is not 0%.

Feature	Description	Value range (5th-95th percentile, median)
DP	NGS read depth at the variant position.	91-433 (233.5)
AD	Number of reads that support the variant call.	31-390 (113)
AF	Fraction of reads that support the variant call, i.e. AD / DP.	0.23-0.56 (0.48)
GC @ 5, 20, 50	Fraction of GC content in the 5, 20, and 50 bases around the variant position.	0.18-0.72 (0.45) 0.29-0.69 (0.44) 0.30-0.67 (0.42)
MQ	Root Mean Square of the mapping quality of the call.	59.4-60 (60)
GQ	Genotype Quality of the call.	50-99 (99)
WHR	Weighted Homopolymer Rate in a window of 20 bases around the variant position: the sum of squares of the homopolymer lengths, divided by the number of homopolymers.	1.7-4.2 (2.4)
HPL-D	Distance to the longest homopolymer within 20 bases from the call position.	0-15 (5)
HPL-L	Length of the longest homopolymer within 20 bases from the call position.	2-6 (4)
QUAL	Quality score assigned by the variant caller to the call.	416-5448 (2675)
QD	QUAL, normalized by DP.	2.7-16.9 (11.2)
FS	Phred-scaled p-value using Fisher's exact test, to detect strand bias.	0-7.6 (1.5)

Table 1: Features used in the logistic regression model.

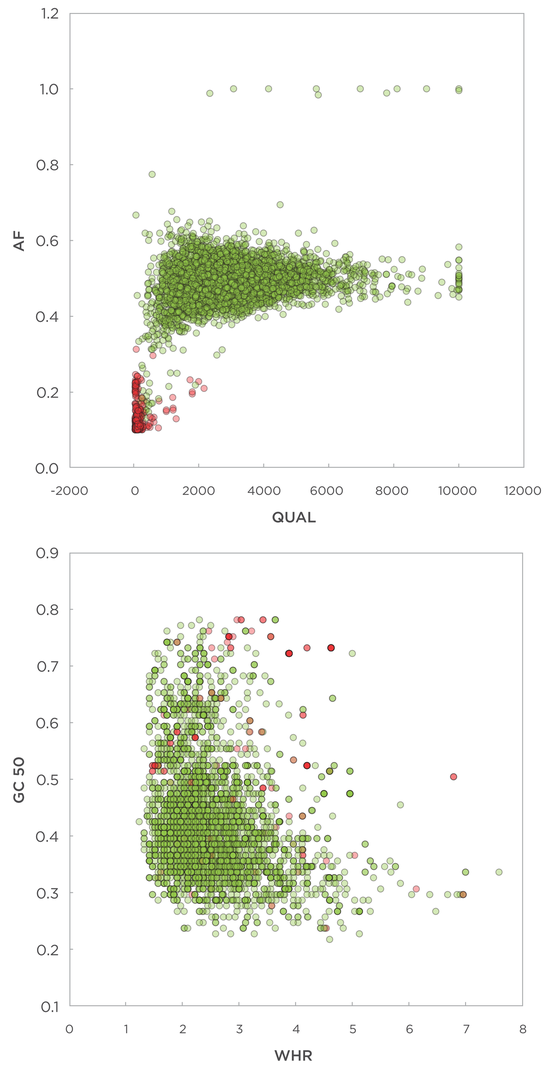

Figure 1: Predictive power of call-dependent and site-dependent signals.

Top: The values of the two strongest features associated with the call: allele frequency (AF) and call quality (QUAL), in all instances in the dataset. Instances marked in green are NGS calls that were confirmed with Sanger, and red ones are cases that did not confirm.

Bottom: The values of the two strongest features associated with the site: GC content in the 50 positions around the variant (GC 50) and weighted homopolymer rate (WHR).

Performance of the variant confidence model vs Sanger sequencing

The rate of low confidence variants that were confirmed by Sanger sequencing was low (17%). However, we wanted to understand the reason for the low-quality score estimated by the model in these cases. We examined the subset of variants that the model predicts as low-quality ones requiring confirmation, and that were confirmed using Sanger sequencing. Over half of these variants were called in less than 30% of the NGS reads; other recurring variant types in this set included those occurring in regions that are typically difficult for NGS sequencing, i.e. having high GC content or in close proximity to long homopolymers. An example of such a variant is shown in Figure 3. In these cases, although the model determined the call to be of low quality, the NGS assay was actually accurate in its calls, further strengthening the observations in [1-4].

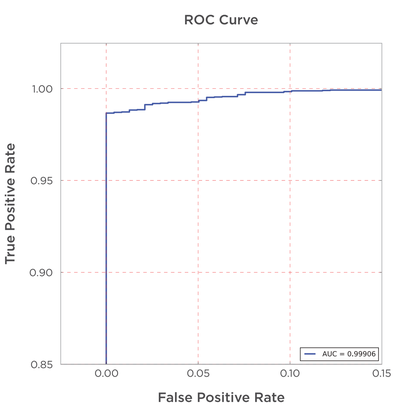

Figure 2: ROC Curve for the variant confidence model, showing an Area Under the Curve (AUC) of almost 1.

	Present	Not Present
High confidence variant	5031 / 5031 (100.0%) *True positive prediction*	0 / 5031 (0%) *False positive prediction*
Low confidence variant	49 / 287 (17.1%) *False negative prediction*	238 / 287 (82.9%) *True negative prediction*

Table 2: A breakdown of the performance of the prediction model on the subset of variants confirmed with Sanger ("Present") and the subset of variant not confirmed ("Not Present").

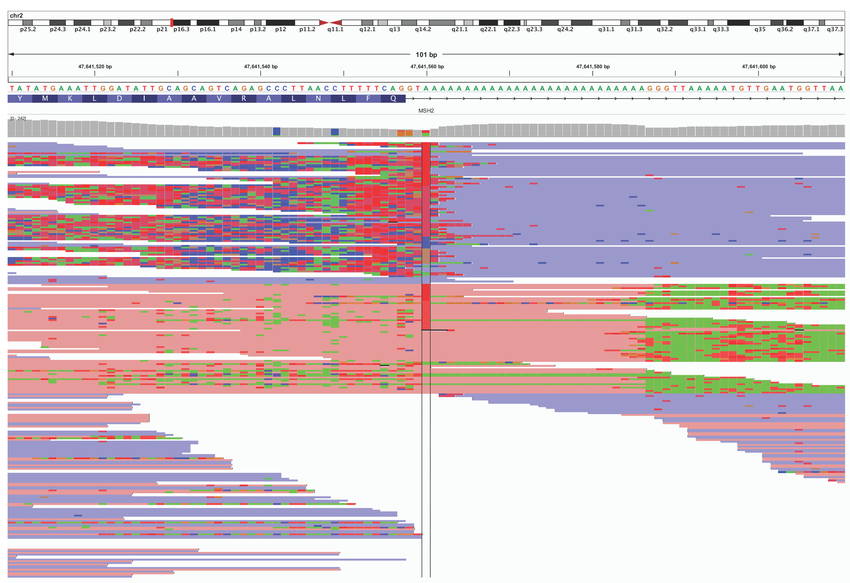

Figure 3: An example of a false negative prediction: the pathogenic variant 'MSH2, c.942+3A>T' is challenging to detect in NGS due to the presence of a long homopolymer. Too stringent filtering would reduce the sensitivity to call such variants.