An Automated Variant Verification System To Improve Reporting Efficiency

S.Katragadda, S. Ghosh, S. Nayanala, A. Das Mahapatra, S. Aliya Afreen, A. Singh, A. Janakiraman, V. Veeramachaneni; **Strand Life Sciences, Bangalore, India**

Introduction

As NGS laboratories move to whole exome sequencing (WES), a larger number of SNVs/Indels are shortlisted in each case.

Before these variants can be included in a report, a final check on variants with borderline quality indicators is carried out by experienced bioinformaticians to ensure the variants are not artifacts arising from genome assembly differences, homology, sequencing errors, or pipeline parameters.

This variant verification (VV) process can be time consuming since it often involves the use of external tools and databases. It can also be subjective since it depends on experts reviewing the reads in a genome browser.

Our modular and pluggable system aims to significantly reduce VV time, by computationally determining if the variant in question is an artifact or not.

Achievements

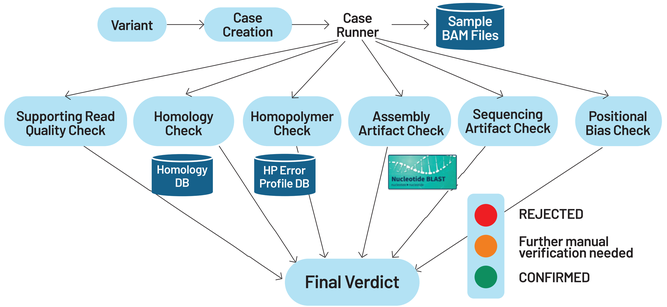
- Developed a modular and extensible system for verifying SNVs/Indels for clinical reporting
- Integrated the system into WES workflow so that all shortlisted variants are automatically processed.
- **Overall reduction of VV effort by ~80%.**

Approach

The system consists of multiple tools which perform different types of quality checks.

When a VV request is made for a specific variant in a sample, each tool assesses if the variant could be a specific type of artifact and returns a verdict.

The system then aggregates the results from all the individual tools and delivers a final verdict.



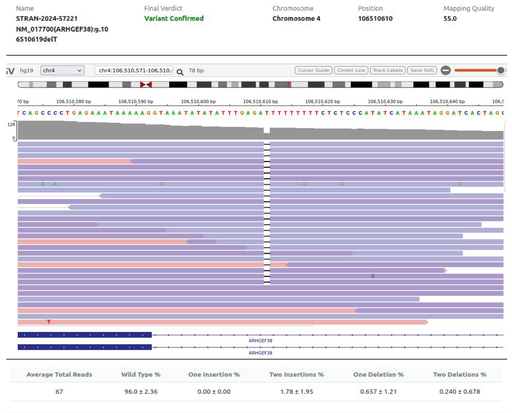
Supporting Read Quality Check

If there are at least 30 reads with length > 120 bp and mapping quality = 60 at the variant locus, and the variant is supported by at least 30% of such reads, the check is passed.

Homopolymeric Artifact Check

Homopolymer checker assesses if an indel in a homopolymer stretch is likely to be genuine. It uses a pre-computed error profile containing the mean (μ) and standard deviation (σ) of the supporting reads % for indels of different lengths.

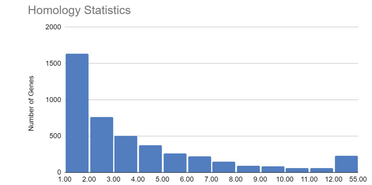
If the supporting reads % of the variant is less than μ + 2 * σ, the variant is labelled a homopolymeric artifact.

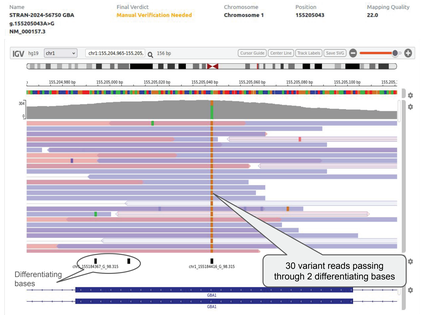


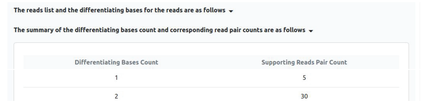
Homology Artifact Check

Homology artifact checker assesses if the variant reads originate from a homologous region. It uses a pre-computed database that contains all differentiating bases between regions in the genome having > 90% similarity with each other. This database covers 16,591 regions from 4,442 genes.

If the variant has at least 5 supporting read pairs that pass through at least 5 differentiating base loci with REF bases at those loci, the check is considered to be passed.







Assembly Artifact Check

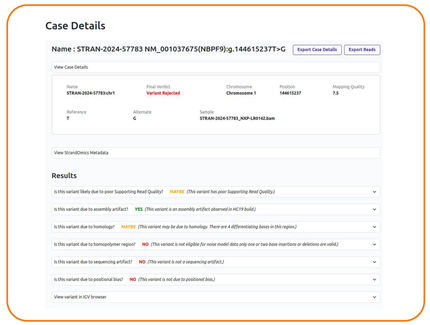
Assembly artifact checker extracts a representative read with the variant allele, aligns it against GRCh38 and T2T-CHM13, and analyzes the alignments to determine if the variant call is a result of build differences.

Positional Bias Check

Positional bias checker detects if the variant is a result of noisy alignment towards the end of the reads.

Sequencing Artifact Check

Sequencing artifact checker flags variants which are supported primarily by mates having disagreeing bases at the variant location.



Contact

Shanmukh Katragadda
Chief Technology Officer - Bioinformatics
+91-9880598107
shanmukh@strandls.com