An Automated Variant Verification System To Improve Reporting Efficiency

S.Katragadda, S. Ghosh, S. NSN, 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 being 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 the 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 the WES workflow so that all shortlisted variants are automatically processed.

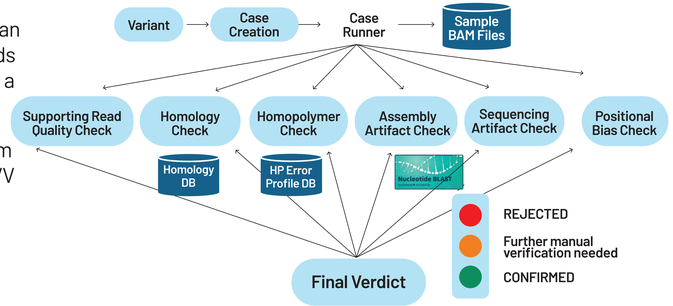
The system resulted in **an overall reduction of VV effort by ~80%.**

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.



Variant Cause: Poor Supporting Read Quality

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 considered as passed.

Variant Cause: Assembly Artifact

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.

Variant Cause: Homopolymer Region

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 as a homopolymeric artifact.

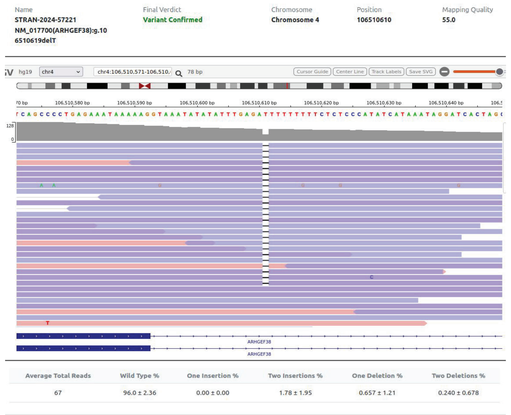
Variant Cause: Homology Artifact

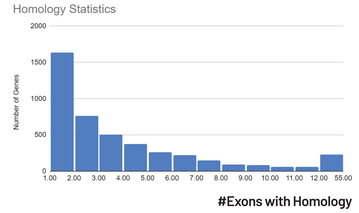
Homology artifact checker assesses if the variant reads originate from a homologous region. It uses a pre-computed database that contains all differentiating bases between regions in the genome having > 90% similarity with each other. This databases covers 16,591 regions from 4,442 genes.

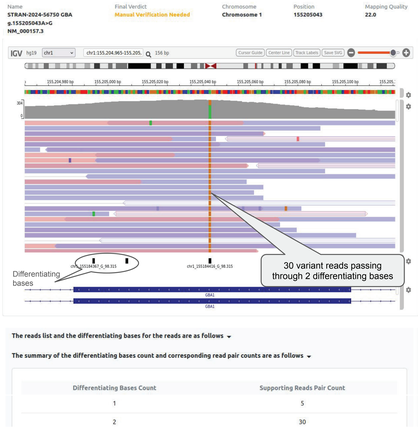
If the variant has at least 5 supporting read pairs that pass through at least 5 differentiating base loci with REF bases at those loci, the check is considered to be passed.

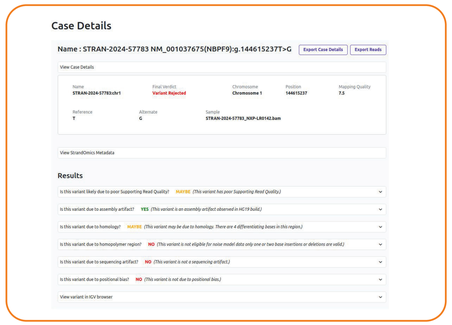
Variant Cause: Positional Bias

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









Contact

Shanmukh Katragadda
Chief Technology Officer - Bioinformatics
(PhD in Neural Networks from IISC, India)

+91-9880598107 shanmukh@strandls.com