Automation Of Time-Consuming WES Steps And Use Of LLMs

Anand Janakiraman, Atanu Pal, Soumit Sur, Ramesh Hariharan;
Strand Life Sciences, Bangalore, India

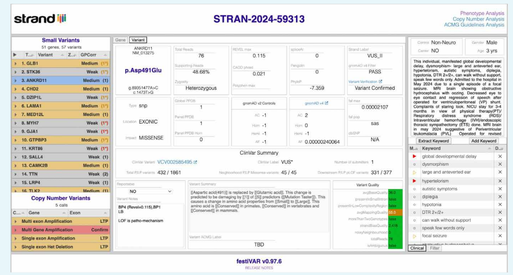
Introduction

In order to keep pace with increasing WES workloads for rare disease diagnosis in our CAP-accredited laboratory, we have implemented an effective variant prioritization algorithm named festiVAR (fast estimation of variants for automated reporting).

This poster describes 3 steps in the automation that we have undertaken - first, that uses a combination of gene level and variant level metrics to reduce the large number of variants to a manageable list; second, that uses LLMs to further prioritize the list based on gene-phenotype correlation; and third, that uses LLMs to perform literature search to aid variant classification.

Summary

The described optimizations were incorporated into an assessment interface that displays the shortlisted ranked gene list, with known phenotypic information displayed readily for efficient genotype-phenotype correlation, and all relevant variant information, enabling manual inspection of the entire automated workflow where needed. This enables our teams of variant scientists to process several hundreds of exomes every month.



Shortlisting Genes and Variants

The shortlisting algorithm needs to take factors associated with both genes and variants into account in coming up with the overall shortlist, carefully fine tuning and normalizing these factors to output an overall score.

The scoring of genes is based on the HPO terms derived from the clinical indications text provided by the treating physician for the case. We found that considering the number of HPO terms of the case that a gene is associated with is insufficient in getting a good rank for the genes.

- One of the challenges is that different HPO terms have each a varying number of genes associated with them, typically based on the level that the term occurs in the HPO ontology.
- The second challenge is that any given gene in the case may be associated with one or more common HPO terms or differentiated HPO terms.

We designed a normalization scheme to downgrade terms that are common to many genes in the case, at the same time upgrading genes that are associated with larger number of differentiated HPO terms.

In addition, we have a repository of curated disease specific gene lists and the presence of the gene in these relevant genelist is also used in the shortlisting.

The variant level scoring is based on many of the standard factors including

- Presence of a P/LP submission in Clinvar
- Allele frequency and hom counts in gnomAD with cutoffs based on the age of onset.
- Pathogenicity predictor scores for the variant
- The known inheritance patterns of the gene and zygosity of the detected variant.

In addition, we have built up a database of variants observed in disease individuals and a cutoff on that is also factored into the scoring metrics above to discard false

Correlating Genes To Clinical Indications

A subsequent round of automation using LLMs was developed to further shortlist the prioritized genes based on their correlation to the clinical indications text.

The phenotypes associated with a gene are derived from publicly available data sources, and a training set of manually assigned correlation as Strong / Medium / Weak to the clinical indications was used concurrently to fine-tune OpenAI's GPT model.

A separate test set was used to determine the accuracy of the fine-tuned model in its ability to assist a variant scientist in shortlisting genes with Strong / Medium correlation and save time in looking at the variants in Weak-ly correlated genes.

As is well known in the reporting of clinical cases in hereditary disorders based on WES, the challenge of variants classified as VUS variants is one that is the most time consuming hindering a fast turnaround of reports. In our experience, we found that cases with multiple Medium ranked genes, which translated to the fact that the clinical indications of the case were not strongly correlated any particular gene, were the ones that were most often reported with VUSes.

Automating Literature Search of Variants

The last step in the reporting of variants in WES cases is the assessment of the variant pathogenicity to assign a label to the variant as per the ACMG guidelines.

Out of the 28 ACMG criteria for the assessment of variants, 19 of them can be automated using a variety of published rules and guidelines. However, the remaining 9 of them are hard to automate and require a manual assessment, often involving a literature search for novel variants. We picked 4 of these criteria, two of them related to segregation studies namely PP1 and BS4, and two of them related to functional studies namely PS3 and BS3, and employed LLMs to assist with the assessment of these criteria.

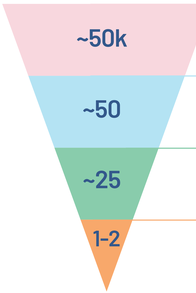
We have implemented a LangChain application to automate the search for scientific literature from PubMed and PubMed Central using synonyms of the gene and variant, then augmented OpenAI's GPT model with these documents to enable Retrieval Augmented Generation (RAG) of the above ACMG criteria.

While the generated text still needed a human eye to assess the relevance of the generated recommendation for these criteria, we consistently found that the automation helped save a fair amount of time in performing the literature searches and assessment manually.

Results

The results from the different automation steps were as follows:

- On a set of 996 cases that were processed using a previous algorithm that on average shortlisted ~60 genes per case, in 99.93% of these cases, the reported variant was within the top 25 genes.
- For fine tuning the LLM, we used clinical notes as well as 1418 manually assessed gene-phenotype correlations (including 459 Strong and 959 Weak correlations). A separate test set of 1590 genes assessed using this GPT model, resulted in an accuracy of 98.43% (498/506 Strong correlations labeled correctly, and 1067/1084 Weak correlations labeled correctly), indicating significant improvements compared to the base GPT-4 LLM which yielded an accuracy of 92%.



WES variants	A typical WES case generates about 45,000 variants, which when annotated with RefSeq genes and transcripts results in about 65,000 unique cdot variants, and 90,000 unique transcript overlaps.
Shortlisted variants	The shortlisting of genes and variants based on the factors explained hitherto resulted in a small list of about 50 genes and variants that were candidates for further inspection.
Prioritized variants	Within the shortlisted variants, we were able to prioritize the top 25 genes and variants further limiting the number that needed manual inspection using LLMs to perform G-P correlation.
Reported variants	The final step of assigning the variant label as per ACMG guidelines requiring a literature search was automated using LLMs to enable faster turnaround time.

Contact

Anand Janakiraman
Chief Operating Officer
& Chief Financial Officer

janaki@strandls.com