Assessing Confidence in Genetic Variants Identified in RNA-Seq Data

Matjaž Žganec[1], Marcel Levstek[2], Roman Luštrik[2], Janez Kokošar[1], Anže Lovše[2,3], Luka Ausec[1], Kristian Urh[2]

1 Genialis, Inc. | 2 Genialis d.o.o. | 3 University of Ljubljana, Faculty of Computer and Information Science

Submission **1076** / Presentation **1129**

INTRODUCTION

RNA-seq data hold significant potential for identifying genetic variants. However, interpreting genomic sites where no variant is detected remains a challenge. These sites may be true negatives (non-events) or may result from insufficient sequencing coverage, making their interpretation uncertain. Resolving this ambiguity is crucial for improving the accuracy and utility of RNA-seq-based variant detection.

Here, we introduce an innovative method for predicting a robust confidence metric, pQUAL, which quantifies the confidence in non-events that are not reported in traditional VCF files. Our metric leverages information from existing VCF files (e.g., coverage, alternative allele depth, and QUAL scores) and coverage files, reporting per-base coverage for all sites to model confidence scores for previously unreported sites (non-event sites). This approach enables accurate classification of sites into likely events, likely non-events, or inconclusive cases based on user-defined thresholds.

CONCLUSION

- **The presented RNA-seq variant calling pipeline accurately detected the KRAS G12C variant in lung and skin cancer samples.**
- **The proposed pQUAL metric provides a robust confidence score for sites not reported by established variant-calling tools.**
- **The presented method was protected by a provisional patent application.**

1 METHODS

Data

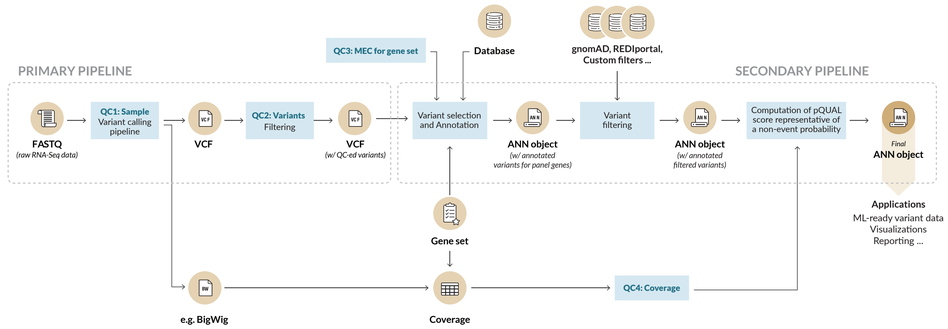
Cancer Cell Line Encyclopedia (CCLE) RNA-Seq samples for lung (n=168) and skin (n=45) cancer. KRAS G12C variant was investigated, given its clinical relevance in lung cancer.

Pipeline

The RNA-Seq reads were processed through a variant calling pipeline to produce standard VCF files. We extracted coverage information and VCF-derived metrics (alternative allele depth, QUAL scores) for all genomic sites. Using these data, we trained a model to compute pQUAL scores for non-event sites-sites for which no variant was yet reported in the VCF files. The model-derived pQUAL values were then used for model validation (by comparing to actual QUAL scores at event sites) and non-event interpretation (whether the sites not reported in the VCF files are true negatives or missed events due to insufficient coverage).

KRAS case study

We applied pQUAL to assess KRAS G12C detection in lung cancer and melanoma samples (case and control, respectively). Coverage plots and confusion matrices were generated to evaluate performance against the ground truth as measured by DNA-Seq.



▶ *Figure 1. Analytical pipeline.*
Primary analysis consisted of calling variants from raw RNA sequencing data and filtering these variants based on variant quality scores, producing a VCF file of variants.
Secondary analysis included additional filtering and selection steps using internal and external sources of knowledge. It produced annotation objects (ANN) complete with variant metadata annotations and our confidence score.
This data can be used for downstream applications such as reporting, visualization, or machine-learning modeling.

2 RESULTS

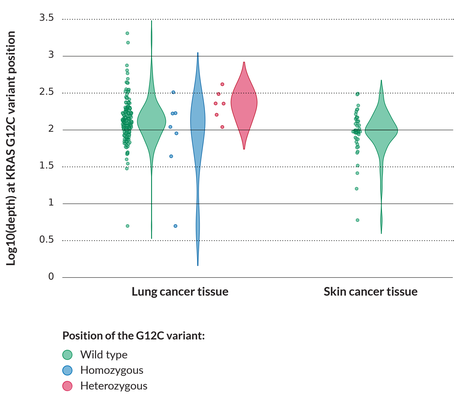
Variant calls

The sequencing coverage at the KRAS G12C mutation site was comparable between lung and skin cancer samples ***(Figure 2)***. However, lung cancer samples are frequently mutated in KRAS at position G12C (10-30 % in different populations), while skin cancer samples are typically wild type at this position ***(Table 1)***.

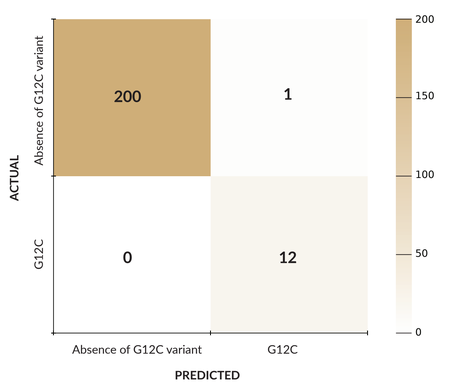
The accuracy of the RNA variant calling pipeline was evaluated against the ground truth ***(Table 1)*** as measured by the DNA-Seq variant calls. The pipeline correctly identified all 12 samples harboring the G12C mutation and 200 cases with the absence of the G12C variant ***(Figure 3)***. There was a single false positive. In this sample, the variant caller called two SNPs at adjacent positions. This indicated two separate amino acid changes at codon twelve (G12V and G12C) instead of a G12F change indicated in ground truth data. This issue can be mitigated downstream by investigating amino acid changes within a codon of a gene or combining adjacent SNPs into MNPs before annotating the variants.

▼ *Table 1. Number of samples with ground truth annotated for lung and skin samples in the CCLE dataset.*

	Lung cancer samples	Skin cancer samples	Total
G12C variant	12	0	12
Without G12C variant	156	45	201
Total	168	45	213



▲ *Figure 2. Violin plot displaying the depth of coverage at the KRAS G12C variant position across lung cancer tissue (n = 168) and skin cancer tissue (n = 45) types and alteration types (absence of G12C variant, homozygous, heterozygous). The depth distribution for each tissue and alteration type is shown, with kernel density estimates (KDE) used to depict the distribution of the data.*



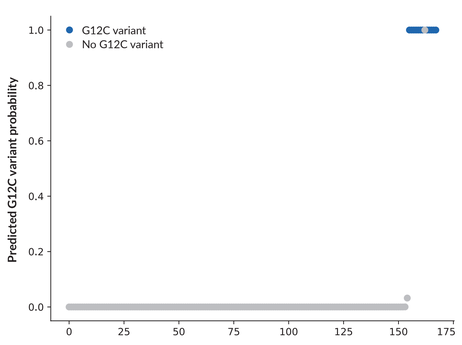
▲ *Figure 3. Confusion matrix showing the comparative RNA vs DNA variant calling results for identifying the presence or absence of KRAS G12C variant. The RNA variant calling pipeline correctly identified 12 G12C samples and 200 samples where it is not present. A G12F sample was misclassified as G12C-mutated, resulting in a single false positive. The color intensity reflects the number of samples in each category, with darker shades indicating a higher count.*

Confidence scores

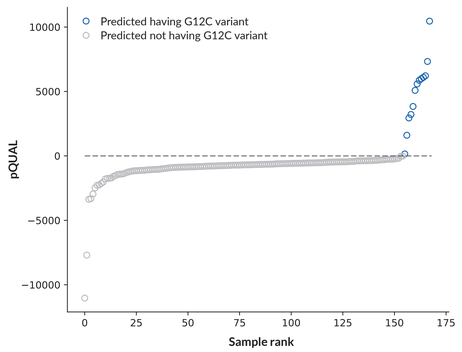
Probabilities of mutations for the G12C variant site for all 213 samples were computed and plotted ***(Figure 4)***. The model was able to differentiate between G12C-positive and G12C-negative samples. All but one sample without the variant (gray points) were correctly predicted with probabilities near zero, while the G12C-positive samples (blue points) were all correctly predicted as mutated with probabilities close to one, showing excellent separation. However, the model misclassified one sample due to a double mutation as discussed above.

The predicted QUAL (pQUAL) scores for lung cancer RNA-seq samples ***(Figure 5)*** provided insight into the model's classification. Samples predicted as G12C-positive (blue outline) generally exhibited high positive pQUAL scores, whereas G12C-negative samples (grey outline) mostly clustered around or below zero. This pattern underscored the model's strong capacity to separate samples based on the probability of the G12C variant.

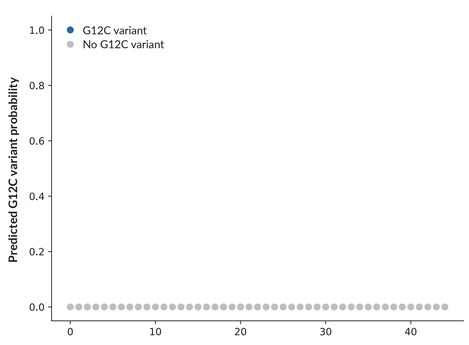
All skin cancer samples were predicted with probabilities near zero ***(Figure 6)***, consistent with the absence of the G12C variant in the ground truth data. One sample had a higher probability (0.00001) due to low coverage at the KRAS G12C site (six reads), but was nonetheless classified correctly.



▲ *Figure 4. Rank plot of predicted G12C variant probabilities for lung cancer samples. The model clearly separated the mutated and wild-type samples. One false positive was explained by the double mutation at adjacent positions, as explained above.*



▲ *Figure 5. A rank plot of pQUAL scores in lung cancer RNA-seq samples. The dashed horizontal line is a classification threshold set at zero. The model effectively differentiated between the two classes.*



▲ *Figure 6. Rank plot of predicted G12C variant probability for skin cancer samples. The probabilities were all negligible, consistent with the ground truth that all skin cancer samples were wild type.*