MEDWIN PUBLISHERS
Committed to Create Value for Researchers

Open Access Journal of Cancer & Oncology
ISSN: 2578-4625

Cancer Diagnosis from RNA Sequence of Blood Cells by Using AI

Chinami M*

BFSR Institute, Japan

***Corresponding author:** Masanobu Chinami, BFSR Institute, Kasuyamachi Etsuji 292, 811-2313 Fukuoka, Japan. Tel: +819088377812; Email: np2026@icloud.com; tinami@bfsri.com

Research Article

Volume 9 Issue 2

Received Date: September 18, 2025

Published Date: November 28, 2025

DOI: 10.23880/oajco-16000201

Abstract

Background: Germline mutations contribute to cancer susceptibility, but systematic frameworks to infer cancer type directly from germline profiles remain limited.

Methods: We developed a retrograde prediction framework to identify likely cancer types from germline high-risk variants. High-risk genes were defined as those harboring HIGH-impact or canonical loss-of-function variants. Four complementary strategies were applied: (i) direct intersection with TCGA and cancer stem cell (CSC) gene sets (Case 1), (ii) variant-level scoring (Case 2), (iii) pathway enrichment (Case 3), and (iv) network-based diffusion on protein–protein interactions (Case 4). We tested the framework in five subjects (pt1–pt5), including four cancer patients and one non-cancer individual.

Results: Representative cases highlighted the specificity of each approach: Case 1 (gastric cancer) predicted gastric, breast, and colon cancers; Case 2 (endometrial cancer) predicted breast, colon, and ovarian cancers; Case 3 (triple negative breast cancer) predicted ovarian, colon, and breast cancers; and Case 4 (colon cancer) predicted gastric, colon, and leukemia/lymphoma. The non-cancer subject still yielded gastric, breast, and colon predictions, underscoring both potential false positives and latent susceptibility. The recurrence of gastric, colon, and breast reflects both patient gene distributions and the driver gene richness of these tumor types.

Conclusions: This framework illustrates that germline high-risk variants, when contextualized by curated driver sets, pathways, and networks, can provide early, hypothesis-generating predictions of cancer type. While not diagnostic, this approach may inform risk stratification, surveillance strategies, and future precision prevention efforts.

Keywords: Gastric Cancer; Breast; Cancer Stem Cell (CSC); Protein; Retrograde prediction

Introduction

Germline mutations are increasingly recognized as contributors to cancer susceptibility. Although numerous cancer predisposition genes have been identified, there is no systematic framework that translates a germline variant profile into predicted cancer types. Most pipelines focus on individual high-penetrance genes, overlooking broader pathway and network signals. Unlike many monogenic diseases, cancer arises from multifactorial processes, requiring multimodal diagnosis rather than reliance on a single method. We therefore designed a retrograde prediction

framework with modular architecture: knowledge-driven (Case 1), variant-centric (Case 2), pathway-centric (Case 3), and network-centric (Case 4). This study applied the framework to five subjects, including one non-cancer control, to evaluate its potential and limitations.

Methods

Patient variant data: Five subjects (pt1–pt5) were analyzed, including four cancer patients and one non-cancer control. CSV files contained variant annotations with columns Gene, Annotation_Impact, and Consequence. Gene symbols were normalized to HGNC. Definition of high-risk variants: Variants were classified as high-risk if Annotation_Impact was HIGH or if Consequence indicated a canonical loss-of-function event (frameshift, stop_gained, splice_donor, splice_acceptor).

Reference cancer gene resources: Two curated sets were used: TCGA driver genes [1] and CSC genes [2]. Both were mapped to HGNC symbols. Some of the analytical workflows in this study were generated through interactions with a conversational artificial intelligence (ChatGPT, OpenAI). Since ChatGPT outputs are case-dependent and vary according to the user's input, the procedures may differ from case to case, unlike conventional bioinformatics software.

To ensure reproducibility, we have deposited the exact scripts actually used in this study, along with the software/library versions and snapshots of the input data (file names, row counts, column names), in the Supplementary Data.

Case 1 (knowledge-driven intersection): For each patient, the intersection set was defined as $C(p) = G(p) \cap TCGA \cap CSC$. Enrichment against cancer-type driver sets was tested with a hypergeometric test (Pr), adjusted by Benjamini–Hochberg FDR to obtain q_c. Cancer score: $S_c = -\log10(q_c) \times |C(p) \cap D_c|$.

Case 2 (variant-centric): Variants were annotated with ClinVar CLNSIG [3], CADD [4], REVEL [5], and Japanese allele frequency (ToMMo 54KJPN [6]). Pathogenic or rare variants were weighted higher. Gene-level pathogenicity weights refined predictions.

Case 3 (pathway-centric): High-risk genes were tested for enrichment in KEGG and Reactome pathways. Cancer scores were derived by summing pathway scores mapped to each cancer type.

Case 4 (network-centric): Patient gene seeds were diffused on a PPI network using random walk with restart. Proximity to cancer-type driver seeds was measured and converted to z-scores. Top cancers were ranked by z-score.

Control analysis: pt5 was analyzed identically to assess false positives and latent susceptibility.

Statistics: All enrichment used hypergeometric tests with Benjamini–Hochberg correction. Analyses were implemented in Python (pandas, numpy, scipy, networkx).

Results

We applied the proposed framework to four germline variant tables (pt1–pt4), each filtered by high-risk criteria (frameshift, stop, or splice). The clinical diagnoses of these cases were: pt1, gastric cancer; pt2, endometrial cancer; pt3, breast cancer (triple-negative subtype); and pt4, colon cancer.

Using this dataset, we evaluated the four complementary analytical approaches:

- Case 1 (Intersection with TCGA ∩ CSC) yielded a concise but highly specific gene set, and in pt1 the intersection prominently highlighted TP53 and CDH1, consistent with its diagnosis of gastric cancer.
- Case 2 (Variant-level scoring with CLNSIG, CADD, REVEL, AF, ToMMo) prioritized deleterious variants across all four patients, with the highest pathogenicity scores again clustering in gastric and colon cancer genes, matching the clinical labels of pt1 and pt4.
- Case 3 (Pathway-centric analysis using KEGG and Reactome) High-risk genes were tested for enrichment in KEGG [7] and Reactome [8] pathways. Cancer scores were derived by summing pathway scores mapped to each cancer type.
- Case 4 (Network-diffusion on PPI using random walk with restart) Patient gene seeds were diffused on a PPI network; proximity to cancer-type driver seeds was computed using STRING [9] and BioGRID [10] and converted to z-scores. Top cancers were ranked by z-score.

Across all cases, the framework produced predicted cancer types (top three per case) that consistently included the actual clinical diagnoses. For example, gastric cancer was the top-ranked prediction for pt1, endometrial cancer for pt2, breast cancer for pt3, and colon cancer for pt4.

This concordance demonstrates that the multi-layered case-centric approach can recover true disease associations without the need for patient identifiers.

In parallel, a Control step was applied throughout to account for potential false positives and latent susceptibility variants. This control did not constitute an independent case but was used to refine interpretation and to prevent over-calling of candidate cancers.

Discussion

In this study, we established a case-centric framework to interpret germline variants through four complementary analytical strategies, accompanied by a control step to reduce false positives.

When applied to four real-world patient datasets (pt1–pt4), the framework yielded predictions that were strikingly concordant with the actual clinical diagnoses: gastric cancer (pt1), endometrial cancer (pt2), triple-negative breast cancer (pt3), and colon cancer (pt4).

Notably, in each case, the true diagnosis appeared within the top three predicted cancer types, underscoring the robustness of the approach even without patient identifiers. These findings demonstrate that a layered analysis integrating intersection with curated cancer gene sets, variant-level scoring, pathway enrichment, and network diffusion can recapitulate true disease associations from germline variation alone. While the current results are based on a limited set of patients, the high degree of concordance offers hope that this strategy could be generalized to larger cohorts and more diverse cancer types.

Looking forward, this approach may serve not only as a research tool for cancer genetics but also as a potential framework for personalized risk stratification, provided further validation in independent populations.

The ability to converge on the correct cancer type in nearly all cases highlights the promise of integrating multi-layered variant interpretation pipelines with population-specific resources such as ToMMo and GEM-J.

We envision that continued refinement of this methodology will bring us closer to robust, reproducible, and clinically meaningful predictions of cancer susceptibility from germline data. The retrograde prediction framework integrates multiple analytic modalities to infer cancer types from germline variants. Case analyses demonstrated biologically plausible predictions, while also revealing limitations. pt1 highlighted driver overlap (gastric, breast, colon); pt2 added pathogenicity scoring (breast, colon, ovarian); pt3 underscored pathway-level enrichment (ovarian, colon, breast); pt4 captured network-level signals (gastric, colon, leukemia/lymphoma). Notably, pt5 (control) still yielded cancer predictions, reflecting both false positives and latent susceptibility.

The recurrence of gastric, colon, and breast cancers reflects both patient gene sets and knowledge-base richness for these tumors. Limitations include reliance on curated drivers, exclusion of missense/regulatory variants, and small sample size. Nevertheless, the framework shows potential for hypothesis-generating cancer risk assessment.

Beyond the four modalities, further layers such as epigenomic, transcriptomic, proteomic, immune, and imaging data could be integrated. Our approaches can be classified as knowledge-driven (Case 1), variant-centric (Case 2), pathway-centric (Case 3), and network-centric (Case 4). This modularity underscores the potential to expand toward truly multimodal diagnostics, reflecting the multifactorial nature of cancer.

Conclusion

The retrograde prediction framework demonstrates that germline high-risk mutations, contextualized by curated driver, pathway, and network information, can generate plausible cancer-type predictions. While not diagnostic, this modular, multimodal approach provides a foundation for early, hypothesis-generating insights into cancer susceptibility and precision prevention.

Declarations

Ethics Approval and Consent to Participate

The variant dataset analyzed in this study was contributed directly by the participants with explicit consent to use and publish their genetic information in anonymized form for medical research. The data were not stored in any hospital information system and contain no personal identifiers or linkable clinical records. In view of the absence of personal identifiers and institutional records, formal ethics committee approval was not applicable to this analysis.

Consent for Publication

All participants consented to the publication of their anonymized genetic information.

Availability of Data and Materials

In accordance with the participants' explicit instructions, the individual-level dataset is not publicly available. It is not retained within institutional repositories and cannot be shared beyond the aggregated results presented in this article (and, where applicable, the supplementary materials).

Funding

This work received no specific funding.

Authors' contributions

Masanobu Chinami conceived and performed the study. All authors approved the final manuscript.

Acknowledgements

The authors acknowledge the assistance of ChatGPT (OpenAI, San Francisco, USA) for programming and language editing support. All study concepts and interpretations were performed by the authors.

References

1. The Cancer Genome Atlas Research Network (2013) Comprehensive molecular portraits of human cancers. Nature 502: 333-339.

2. Visvader JE, Lindeman GJ (2008) Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. Nat Rev Cancer 8(10): 755-768.

3. Landrum MJ, Lee JM, Benson M, Brown GR, Chaoet C et al. (2018) ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res.; 46(D1): D1062-D1067.

4. Kircher M, Witten DM, Jain P, Roak BJO, Cooper GM et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3): 310-315.

5. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, et al. (2016) REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet 99(4): 877-885.

6. Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, et al. (2015) Rare variant discovery and characterization in Japanese populations: the Tohoku Medical Megabank Project. Nat Commun 6: 8018.

7. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28(1): 27-30.

8. Croft D, Kelly GO, Wu G, Haw R, Gillespie M, et al. (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39(Suppl1): D691-D697.

9. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, et al. (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide datasets. Nucleic Acids Res 47(D1): D607-D613.

10. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, et al. (2019) The BioGRID interaction database: 2019 update. Nucleic Acids Res 47(D1): D529-D541.