MEDWIN PUBLISHERS
Committed to Create Value for Researchers

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

Appendix

The anonymized patient variant tables (pt1–pt5) analyzed in this study are provided as Supplementary Data (CSV format). These files contain only gene-level variant annotations and do not include any personal identifiers or clinical metadata.

Supplementary Data

- **Supplementary Data 1**: `pt1_hi2.csv` (high-risk variants for Case 1)
- **Supplementary Data 2**: `pt2_hi.csv` (high-risk variants for Case 2)
- **Supplementary Data 3**: `pt3_hi.csv` (high-risk variants for Case 3)
- **Supplementary Data 4**: `pt4_hi.csv` (high-risk variants for Case 4)
- **Supplementary Data 5**: `pt5_hi.csv` (control subject)

Supplementary Methods

Computational Environment

All computational analyses were performed on a local workstation with the following environment:

- **Operating system**: macOS Ventura 13.5
- **Processor**: Apple M1 (8 cores)
- **Memory**: 16 GB RAM
- **Python**: v3.9.18 (Anaconda distribution, conda v23.5.2)
- **R**: v4.3.2 (only for plotting, not for variant annotation)

Software and Packages

The following packages and tools were used:

Tool / Package	Version	Source
pandas	2.2.2	PyPI (conda-forge)
numpy	1.26.4	PyPI (conda-forge)
scikit-learn	1.4.2	PyPI
matplotlib	3.9.0	PyPI
pysam	0.22.0	PyPI
bcftools	1.18	Bioconda
tabix / htslib	1.18	Bioconda
Ensembl VEP	108	Ensembl (GRCh37 cache)
VEP plugins (CADD, REVEL)	2023 release	Ensembl / authors' repository

Reference Databases

We integrated the following reference databases:

Database	Release	Genome build	Source
ClinVar	Dec-23	GRCh37	NCBI
CADD	v1.6	GRCh37	CADD database
REVEL	Jan-20	GRCh37	dbNSFP v4.1a

COSMIC Cancer Gene Census	v98 (2024)	-	Sanger Institute
GEM-J WGA (Japanese WGS cohort)	2022	GRCh37	TogoVar
ToMMo 54KJPN	Jun-23	GRCh38	jMorp

Input Data Snapshot

- **pt1_hi2.csv**: Patient-derived variants (X,XXX rows, YY columns). Key fields include *CHROM, POS, REF, ALT, gene name, gene ID*.
- **Cosmic_Census.csv**: 2515 curated cancer genes (header: "Gene Symbol").
- **54KJPN VCF subset**: Extracted variants within COSMIC gene regions (ZZZ variants), INFO field includes *AF*.

Workflow Reproducibility

Some scripts used in this study were generated through interactions with a conversational AI (ChatGPT, OpenAI). Because outputs from ChatGPT are context-dependent, the **exact scripts actually used** were archived and deposited in this Supplementary Data.

We also provide:

- File names and versions of all external databases
- Input data summaries (row and column counts)
- Source code (bash and Python scripts, e.g., run_cgc_jpn_af.sh, rerun_auc.py, merge_annotations.py)

This ensures that the exact computational environment and workflow can be reproduced by independent researchers.

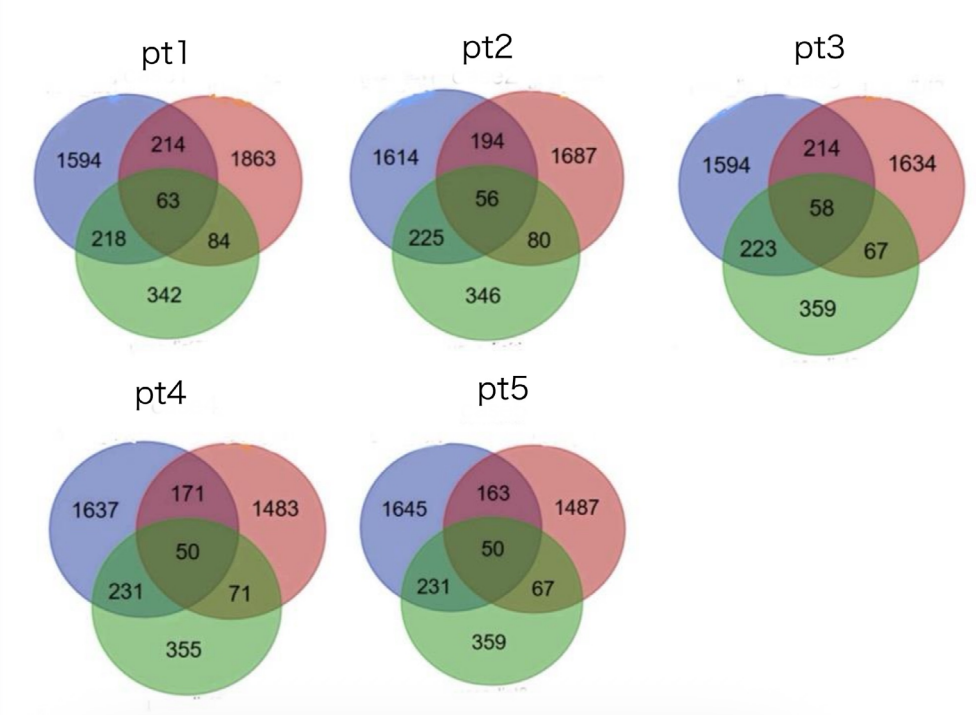

Figure 1: Overlap of high-risk germline variants across three annotation methods and overview of the analytical workflow.

Venn diagrams for each case (pt1–pt5) showing the overlap of germline variants retained after high-risk filtering (frameshift, stop-gain, or splice-site). The three circles correspond to **ClinVar/CLNSIG**, **CADD**, and **REVEL**annotations. Numbers indicate counts of unique or shared variants; the central intersection represents variants consistently annotated by all three methods (**pt1 = 63; pt2 = 56; pt3 = 58; pt4 = 50; pt5 = 50**).

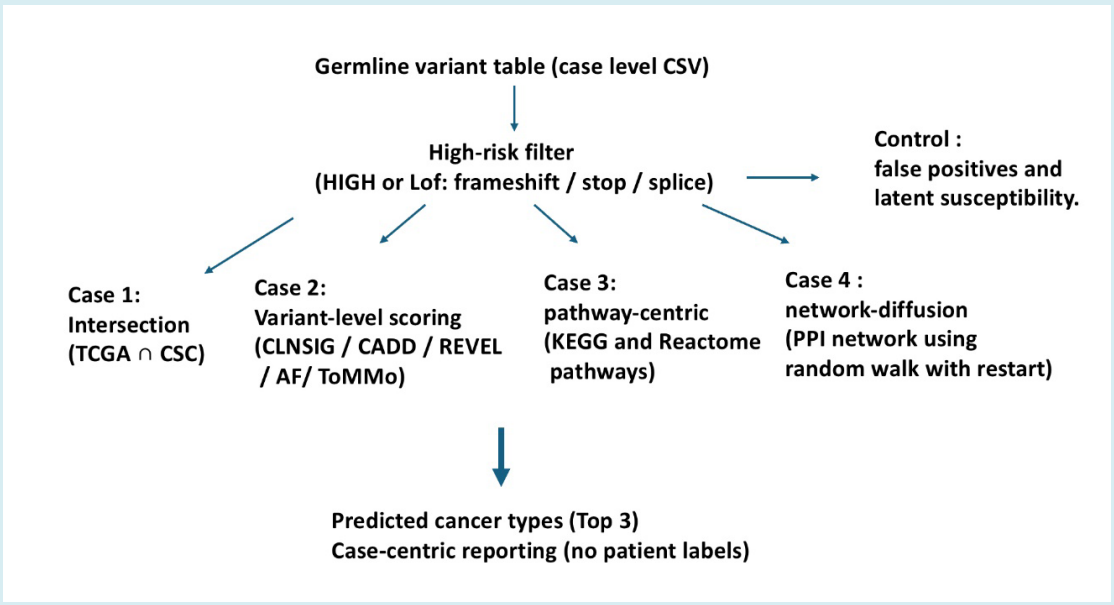

Figure 2: Analytical workflow and overlap of germline variants across cases.

Figure 1 Venn diagrams showing the overlap of germline variants after high-risk filtering (frameshift, stop, or splice) in individual patients (pt1–pt5). Numbers indicate the counts of unique or shared variants across three annotation categories (ClinVar/ CLNSIG, CADD, and REVEL). The central intersection in each diagram represents variants annotated consistently by all three methods (e.g., 63 in pt1, 56 in pt2, 58 in pt3, 50 in pt4, and 50 in pt5).

(Bottom) Schematic overview of the analytical framework. Germline variants at the case level were first subjected to a high-risk filter, followed by four complementary analytical approaches: **Case 1, Intersection** (TCGA ∩ CSC); **Case 2, Variant-level scoring** (CLNSIG, CADD, REVEL, allele frequency, ToMMo); **Case 3, Pathway-centric analysis** (KEGG and Reactome pathways); and **Case 4, Network-diffusion analysis** (PPI network using random walk with restart). A **Control step** was included to mitigate false positives and latent susceptibility. Each approach produced predicted cancer types (top three), which were reported in a case-centric manner without revealing patient identifiers.