PacBio Structural Variant (SV) Pipeline

🧬 Overview
This pipeline is designed for the annotation and clinical prioritization of structural variants (SVs) detected from PacBio HiFi long-read sequencing data. It is tailored for somatic cancer variant discovery and classification, particularly in leukemia cases.

✅ Key Features
- Germline filtering using tumor-normal comparison
- Functional annotation via known oncogenes, tumor suppressors, and cancer driver genes
- Clinical relevance scoring using real-time API queries to:
  - CIViC
  - OncoKB
  - COSMIC (local fallback)
- SV-specific scoring based on size, type (e.g., fusions, deletions, duplications)
- ACMG-style clinical classification: Pathogenic, Likely Pathogenic, VUS, Benign
- Excel output with scoring breakdown and logs for each sample

📁 Project Structure
pacbio-sv-pipeline/
├── data/
│ ├── oncogenes.txt
│ ├── tumor_suppressors.txt
│ └── cosmic_gene_hits.tsv
├── scripts/
│ └── structural_variant_finder.R
├── test_samples/
│ ├── test_tumor.tsv
│ └── test_normal.tsv
├── results/
├── test_run.sh
├── README.md
└── LICENSE

🚀 How to Run
Prerequisites
- R ≥ 4.0
- R packages: `dplyr`, `stringr`, `openxlsx`, `readr`, `httr`, `jsonlite`, `clusterProfiler`, `org.Hs.eg.db`, `ggplot2`
- API keys:
  - CIViC: https://civicdb.org
  - OncoKB: https://oncokb.org

Command
```bash
Rscript scripts/structural_variant_finder.R data/sample01_tumor.tsv
```

data/sample01_normal.tsv
📊 Outputs
sample01_clinical_significance.xlsx: Annotated SV table with gene, SV type, clinical scores, and classification

sample01_pipeline_log.txt: Full log of the run with error handling notes

🔧 Customization
Replace gene lists in data/ with custom panels

Replace COSMIC annotations or expand fallback sources

Adjust scoring thresholds inside the R script

🧪 Example Test Run
Run the test samples with:

bash test_run.sh