# 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

---

## How to Run

### Prerequisites

- **R ≥ 4.0**
- **R packages**: `dplyr`, `stringr`, `openxlsx`, `readr`, `httr`, `jsonlite`, `clusterProfiler`, `org.Hs.eg.db`, `ggplot2`

#### Input Files

- Tumor SV calls exported as a tab-delimited file (`.tsv` or `.txt`) with columns:

  ```
  Chr, Start, End, SV_Type, Gene_name, [other fields]
  ```

- Matched normal SV calls in the same format.
- Local COSMIC TSV (to annotate known cancer SVs). Place it under `data/` (see Folder Structure).
- **API keys** for querying CIViC and OncoKB.

### Obtain API Keys

1. Sign up (free) at **CIViC** and obtain a CIViC API key.
2. Sign up (free) at **OncoKB** and obtain an OncoKB API key.
3. Edit the top of `scripts/structural_variant_finder.R` and set:

   ```
   civic_api_key  <- "<YOUR_CIVIC_API_KEY>"
   oncokb_api_key <- "<YOUR_ONCOKB_API_KEY>"
   ```

### Prepare Input Files

1. Create a folder named `data/` (if it doesn't exist).
2. Under `data/`, place your `tumor_sample1.tsv` (tumor SV calls) and `normal_sample1.tsv` (matched normal SV calls).
3. Place `cosmic.tsv` (downloadable from **COSMIC → Structural Variants**) into `data/`.

### Run the Pipeline

From the repository root, execute:

```
Rscript scripts/structural_variant_finder.R \
    data/tumor_sample1.tsv \
    data/normal_sample1.tsv
```

**Outputs**

- `results/sample01_clinical_significance.xlsx`: Annotated SV table with gene, SV type, clinical scores, and classification.
- `results/sample01_pipeline_log.txt`: Full log of the run with error handling notes.

### Inspect Results

- Open `results/tumor_sample1_clinical_significance.xlsx` in Excel or R to view the combined score and clinical classification for each SV.
- The sheet `results/tumor_sample1_annotated.tsv` includes raw annotations (CIViC/OncoKB/COSMIC flags, functional impact, etc.).
- Any PDF or PNG plots appear under `results/plots/`.

---

## Folder Structure

```
PacBio-SV-Pipeline/
├── README.md
├── docs/
│   └── README.pdf
├── data/
│   ├── tumor_sample1.tsv
│   ├── normal_sample1.tsv
│   └── cosmic.tsv
└── scripts/
    └── structural_variant_finder.R
```

- **`README.md`**: A concise overview and quick-start guide (GitHub landing page).
- **`docs/README.pdf`**: This full PDF documentation (detailed methods, advanced notes).
- **`data/`**: User-supplied inputs (SV call files, COSMIC reference).
- **`scripts/`**: Main R script (`structural_variant_finder.R`).
- **`results/`**: Created automatically after running the pipeline.

---

## Usage Notes & Customization

### Customizing API Queries

- By default, the script queries CIViC and OncoKB via REST endpoints using `httr` and `jsonlite`.
- To skip API lookups (e.g., no internet), comment out lines 85–120 (the "Clinical API" section) in `structural_variant_finder.R`. The script will still run, but clinical scores rely only on local COSMIC.

### Filtering Thresholds

- Inside `structural_variant_finder.R`, locate the block (lines 45–60) that defines functional-impact penalties and scoring weights.
- To adjust, modify the `case_when` logic for `Functional_Impact` or change the points added for each category.

### Output Directory

- By default, the script creates a `results/` folder in the working directory.
- To change the output path, set the environment variable `OUTDIR` before running:

  ```
  export OUTDIR="/my/custom/output/path"
  Rscript scripts/structural_variant_finder.R data/tumor.tsv data/normal.tsv
  ```

- The code will detect `OUTDIR` and write all outputs there instead of `./results/`.

### SLURM/HPC Integration

- For HPC environments, copy the first ~10 lines of `structural_variant_finder.R` (the "Argument Parsing" block) into a Slurm job script (`.sbatch`).
- **Example `job_svs.sbatch`:**

  ```
  #!/bin/bash
  #SBATCH --job-name=SV_Sample1
  ```

```
#SBATCH --cpus-per-task=4
#SBATCH --mem=16G
#SBATCH --time=02:00:00

module load R/4.0
Rscript /path/to/PacBio-SV-Pipeline/scripts/structural_variant_finder.R \
    /path/to/data/tumor_sample1.tsv \
    /path/to/data/normal_sample1.tsv
```

- Submit with:

```
sbatch job_svs.sbatch
```

---

## Full Methods & Scoring Criteria

### Germline Filtering (Tumor-Normal Comparison)

1. Load tumor and normal SV call tables (tab-delimited, columns: `Chr`, `Start`, `End`, `SV_Type`, `Gene_name`, etc.).
2. Merge tables on matching coordinates and SV type to identify shared ("germline") SVs.
3. Retain tumor-only SVs for downstream annotation.

### Functional Annotation

1. **Gene Overlap**: Map each SV to overlapping genes using `GenomicRanges` (R). Annotate gene symbols and gene biotypes.
2. **Cancer Gene List**: Tag SVs affecting genes in a curated list of known oncogenes and tumor suppressors (from COSMIC's Cancer Gene Census).
3. **Gene Ontology Enrichment** (optional): For large SV sets, run `clusterProfiler::enrichGO()` on affected gene lists.

### Clinical Database Queries

- **CIViC**: Query using REST API to retrieve evidence items for gene and variant pairs. Score based on evidence level and significance.
- **OncoKB**: Query OncoKB REST endpoints for variant-level annotations (e.g., known actionable fusions or amplifications).
- **COSMIC**: Use a local `cosmic.tsv` to flag SVs previously observed in cancer samples. Assign a baseline "COSMIC_score" based on recurrence frequency.

### SV-Specific Scoring

1. **Size-Based Weighting**:
    - Deletions > 1 kb: +2 points
    - Duplications > 1 kb: +1 point
    - Translocations/fusions: +3 points
    - Inversions > 5 kb: +1 point
2. **Functional Impact**:
    - SV overlaps coding region (+2)
    - SV disrupts known tumor suppressor (+3)
    - SV creates potential fusion involving oncogene (+4)
    - SV in intergenic region (0)
3. **Clinical Evidence Points**:
    - CIViC Level A: +5
    - CIViC Level B: +4
    - OncoKB Level 1: +5
    - OncoKB Level 2: +4
    - COSMIC recurrence > 10 samples: +3
    - COSMIC recurrence 1–10 samples: +1
4. **Total Score Calculation**:

```
total_score <- size_points + functional_points + clinical_points
```

5. **ACMG-Style Classification**:
   - `>= 8`: Pathogenic
   - `>= 5 & < 8`: Likely Pathogenic
   - `>= 2 & < 5`: VUS (Variant of Uncertain Significance)
   - `< 2`: Benign

### Excel Report Generation

- Use `openxlsx::write.xlsx()` to create `sample01_clinical_significance.xlsx` with two sheets:
  1. **`Annotated_SVs`**: All SVs with columns for `Chr`, `Start`, `End`, `SV_Type`, `Gene_name`, `size_points`, `functional_points`, `clinical_points`, `total_score`, `classification`.
  2. **`Summary_Stats`**: Counts of SVs in each classification and distribution plots (if applicable) embedded.

---

## Example Outputs

- **`results/sample01_clinical_significance.xlsx`**: Shows 45 SVs for `sample01` with scores and classifications.
- **`results/sample01_annotated.tsv`**:

```
Chr     Start   End     SV_Type Gene_name       size_points     functional_points       clinical_points tota
chr3    123456  223456  Deletion        TP53    2       3       5       10      Pathogenic
chr7    98765   198765  Fusion  BCR-ABL1        3       4       5       12      Pathogenic
...
```

- **`results/plots/size_distribution.png`**: Histogram of deletion/duplication lengths.
- **`results/plots/classification_pie.png`**: Pie chart of SV classification proportions.

---

## Credits & Contact

- **Author**: Sally L. Yepes Torres
- **Email**: sallyepes233@gmail.com
- **Last Updated**: June 2025
- **License**: MIT

---

## Change Log

- **v1.0 (June 2025)**
  - Initial public release with core SV annotation and clinical scoring functionality.
  - Integrated CIViC, OncoKB, COSMIC local annotations, and simple ACMG/AMP scoring.
  - Output: annotated TSV + clinical_significance.xlsx + summary plots.
- **v1.1 (forthcoming)**
  - Add SV size–distribution plots (e.g., histogram of deletion/duplication lengths).
  - Incorporate normal-adjacent filtering pipeline (to remove germline SVs).
  - Expand annotation with FusionCatcher for gene fusions.