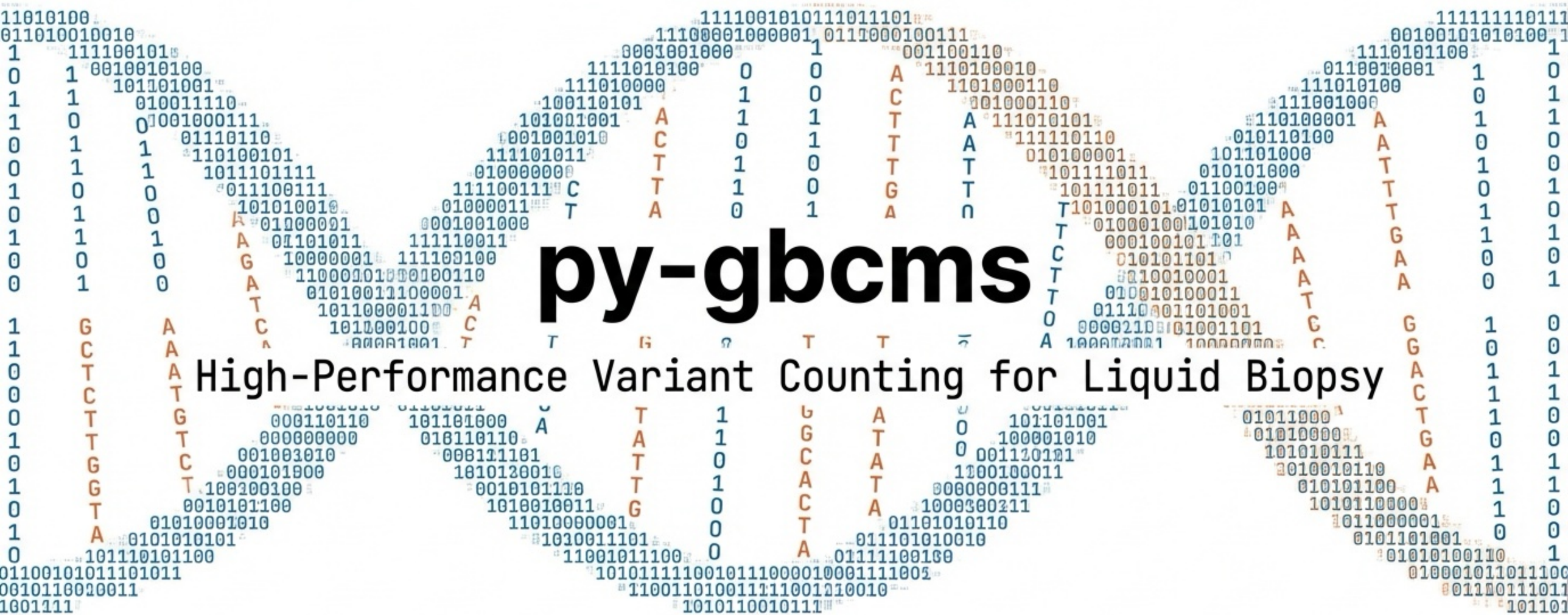


High-Performance Variant Counting for Liquid Biopsy

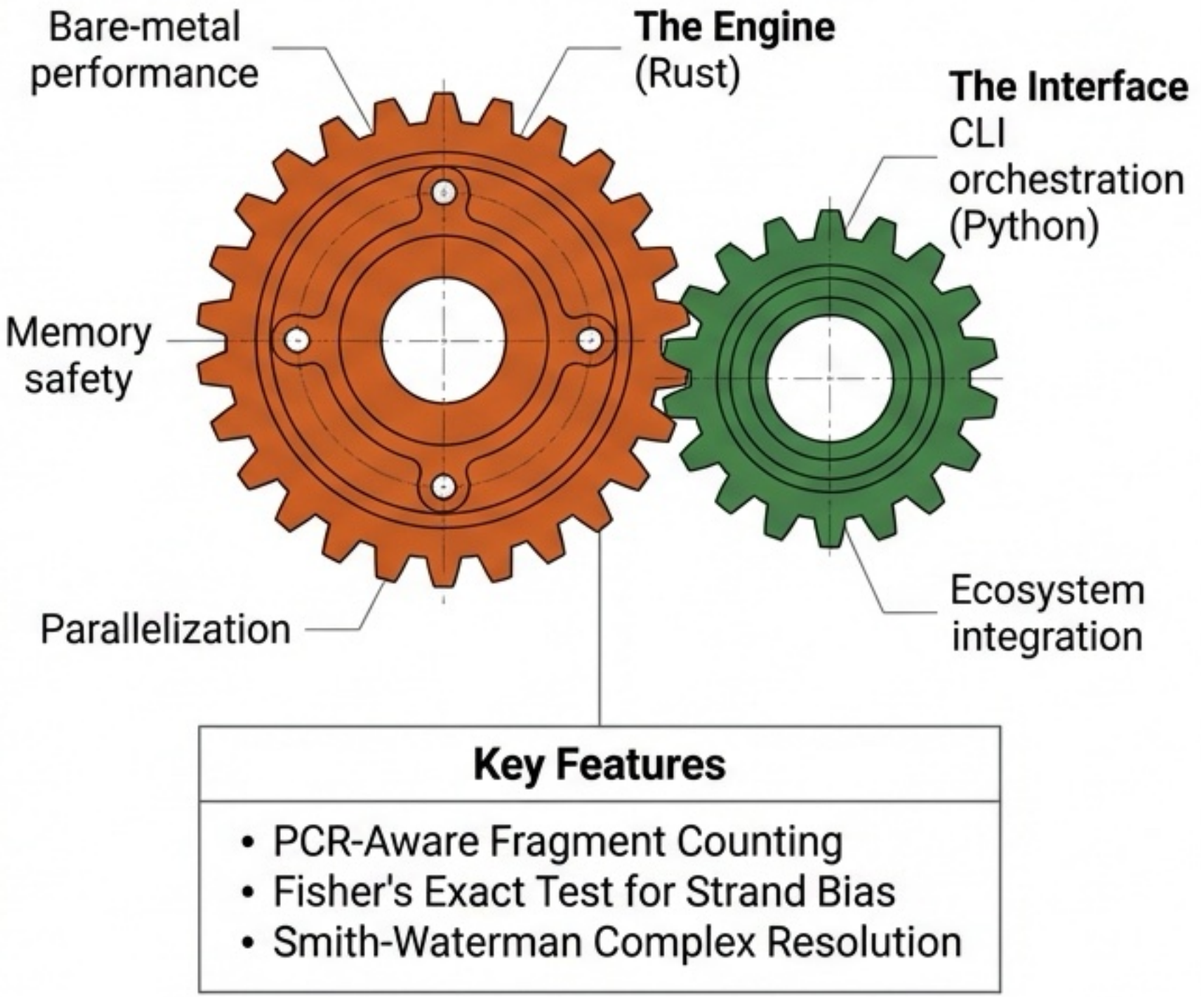
A Rust-Accelerated Architecture for Accurate cfDNA Analysis

Author: MSK-ACCESS |

Precision Counting at at Production Scale

- **Core Function:** Extracts allele counts, VAF, and fragment metrics from BAM files.
- **The Focus:** Designed for Liquid Biopsy (cfDNA).
- **The Output:** Standardized VCF and extended MAF formats.

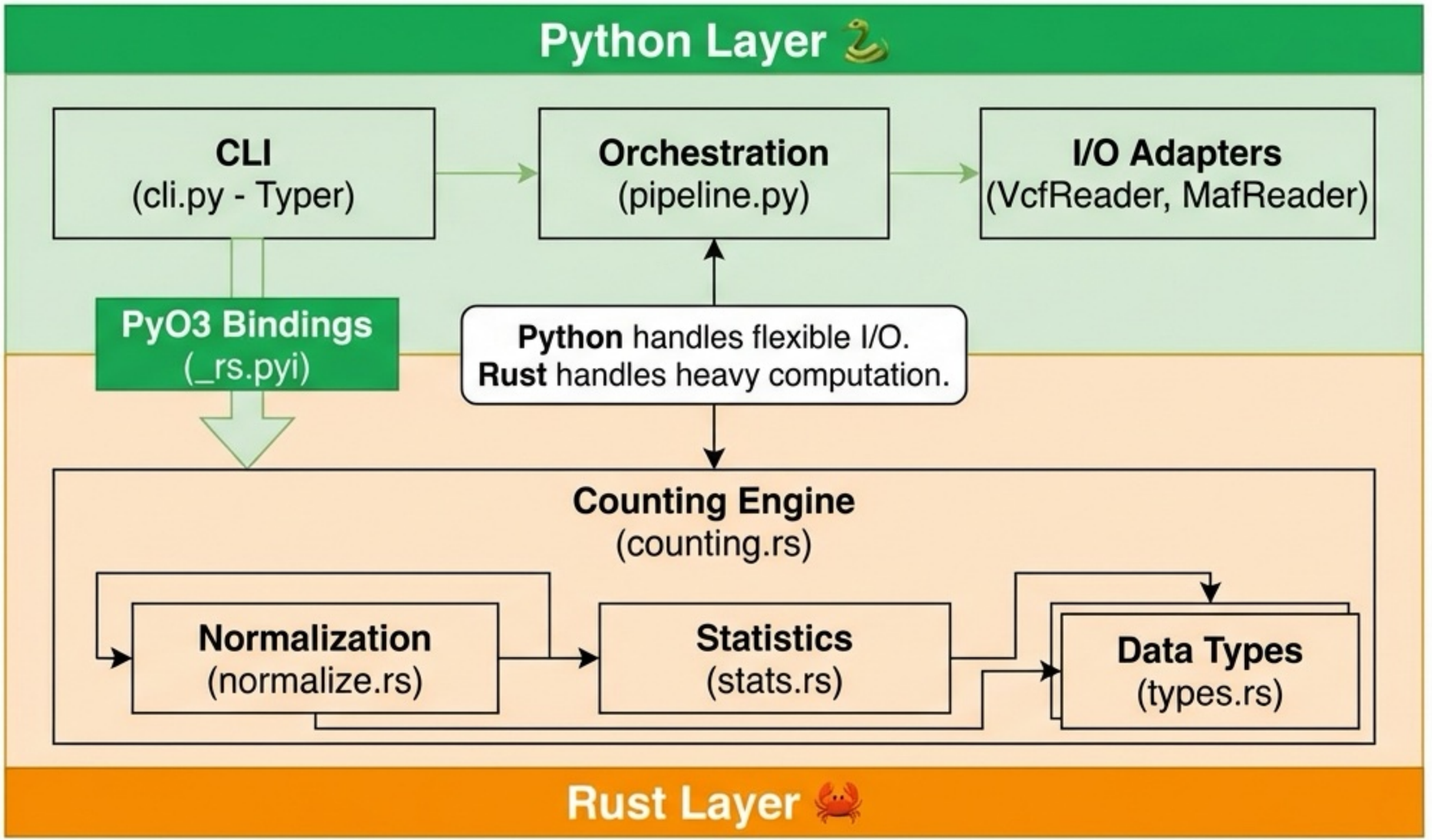
The Hybrid Advantage



Key Features

- PCR-Aware Fragment Counting
- Fisher's Exact Test for Strand Bias
- Smith-Waterman Complex Resolution

The Hybrid Architecture



Deployment Strategies

Standard (PyPI)

`pip install py-gbcms`

- Requires: Python 3.10+, glibc 2.34+
- OS: Ubuntu 22.04+, RHEL 9+

Containerized (Production)

Docker: `ghcr.io/msk-access/py-gbcms`

- Singularity / Apptainer
- Essential for clinical reproducibility

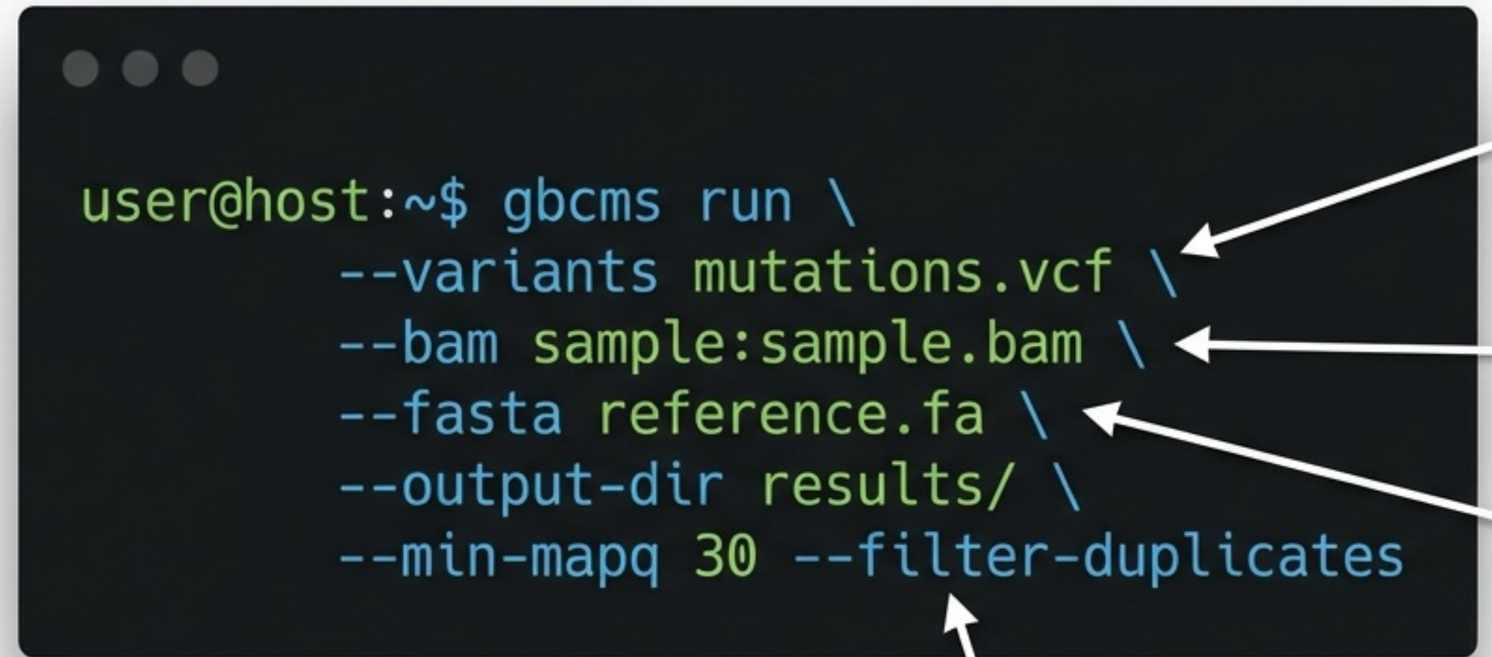
Legacy / HPC Source

Build from source via `maturin`

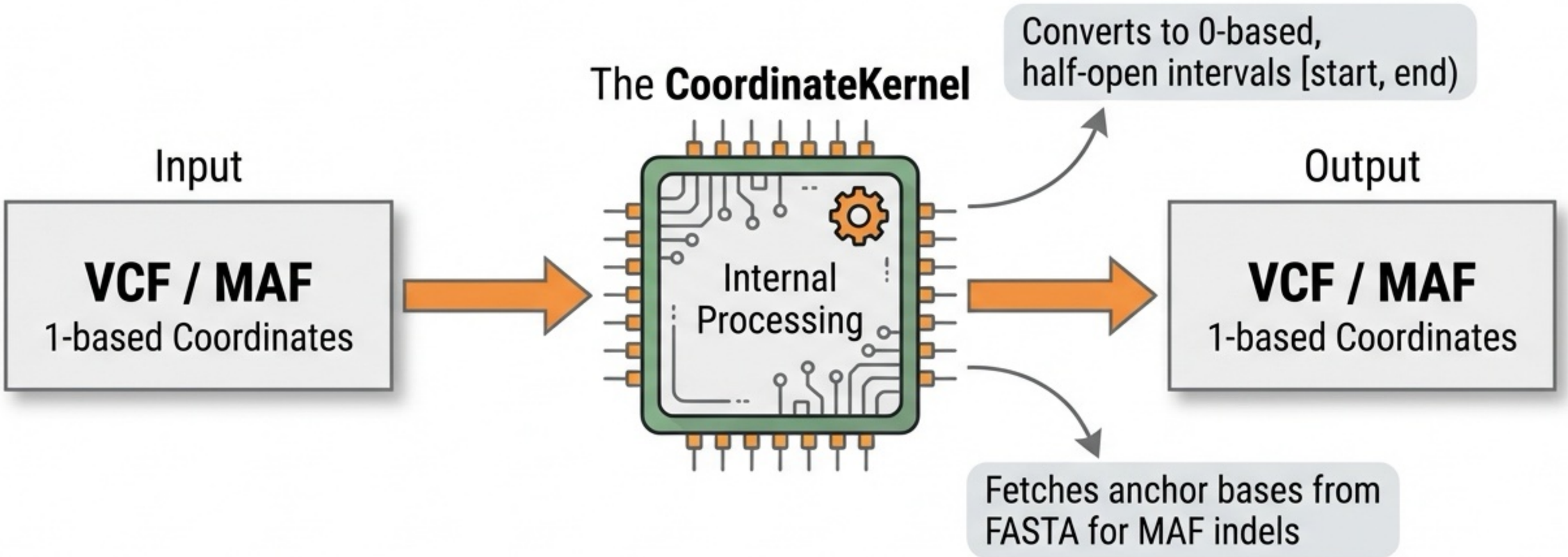
- For RHEL 8 / CentOS 8
- Requires: clangdev + rust toolchain

Pre-built wheels available for Linux (x86_64, aarch64), macOS (Intel/Silicon), and Windows.

The CLI Workflow



Coordinate Normalization & Integrity



Solves the **off-by-one errors** common in bioinformatics by centralizing logic.

The Metrics That Matter

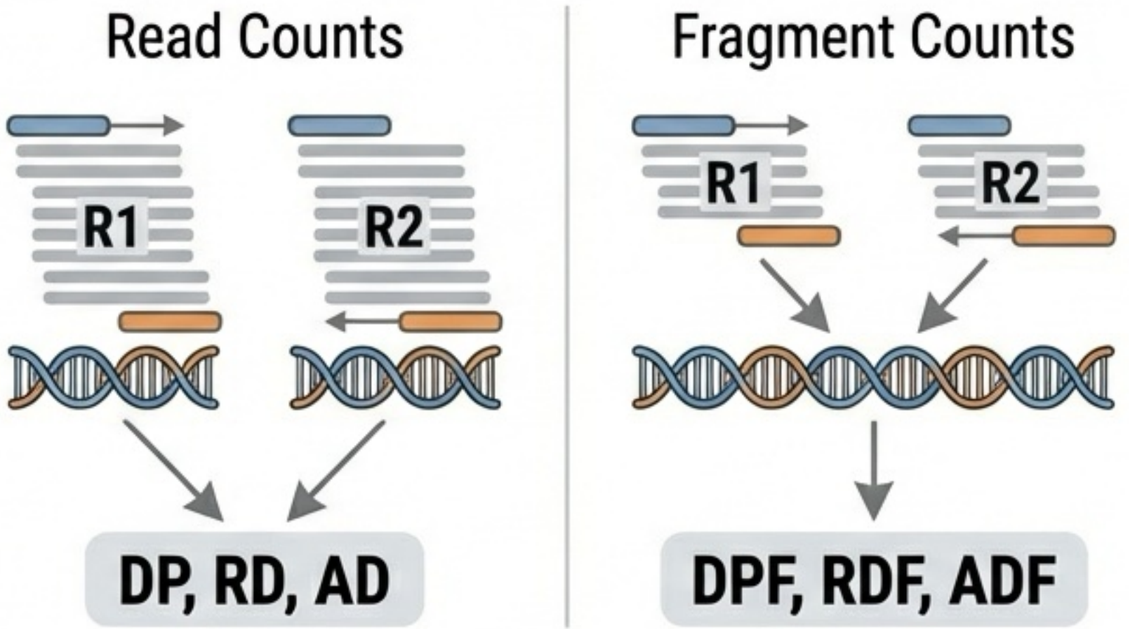
Primary Metrics

$$\mathrm{VAF} = \frac{AD}{RD + AD}$$

Variant Allele Frequency. The core measure of mutation abundance.

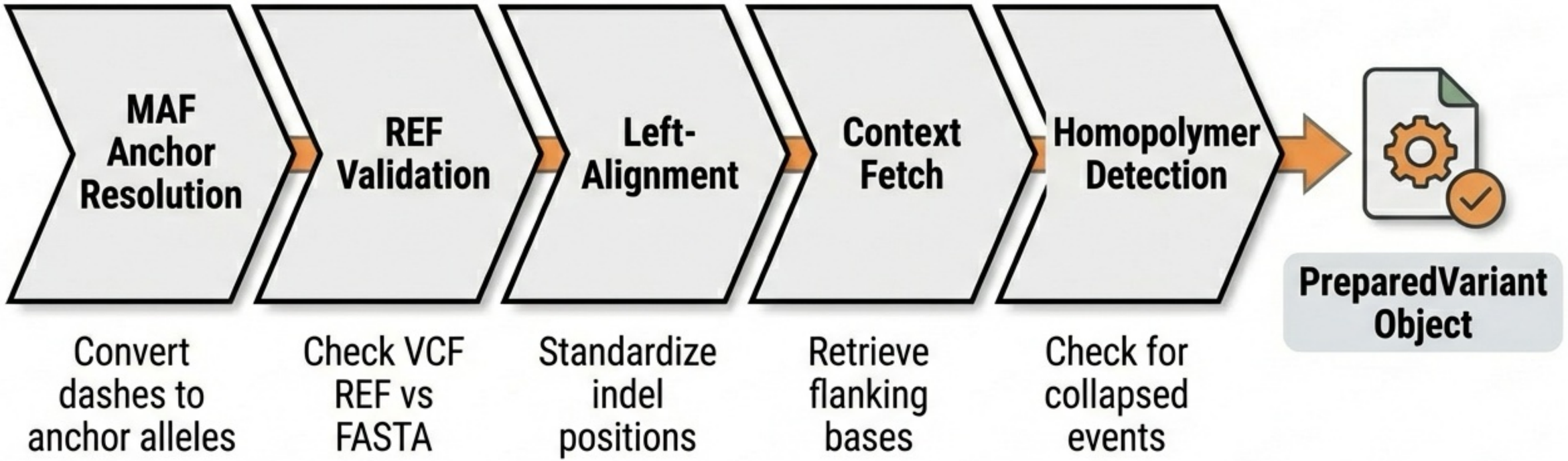
Strand Bias: Fisher's Exact Test (p-value & Odds Ratio). Detects sequencing artifacts.

The **cfDNA Distinction**



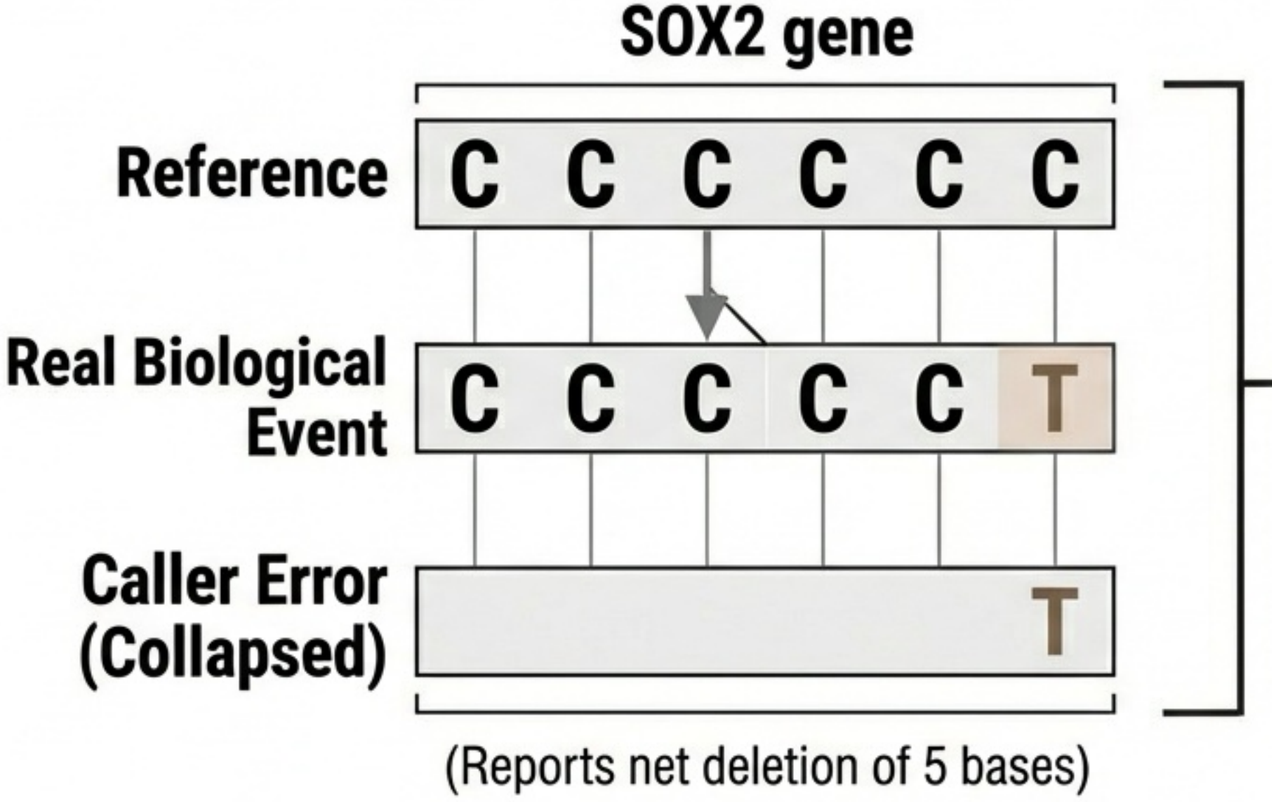
Fragment counting deduplicates read pairs to prevent artificial inflation of evidence.

The Variant Preparation Pipeline



Ensures biological consistency before counting begins.

Edge Case: Homopolymer Decomposition



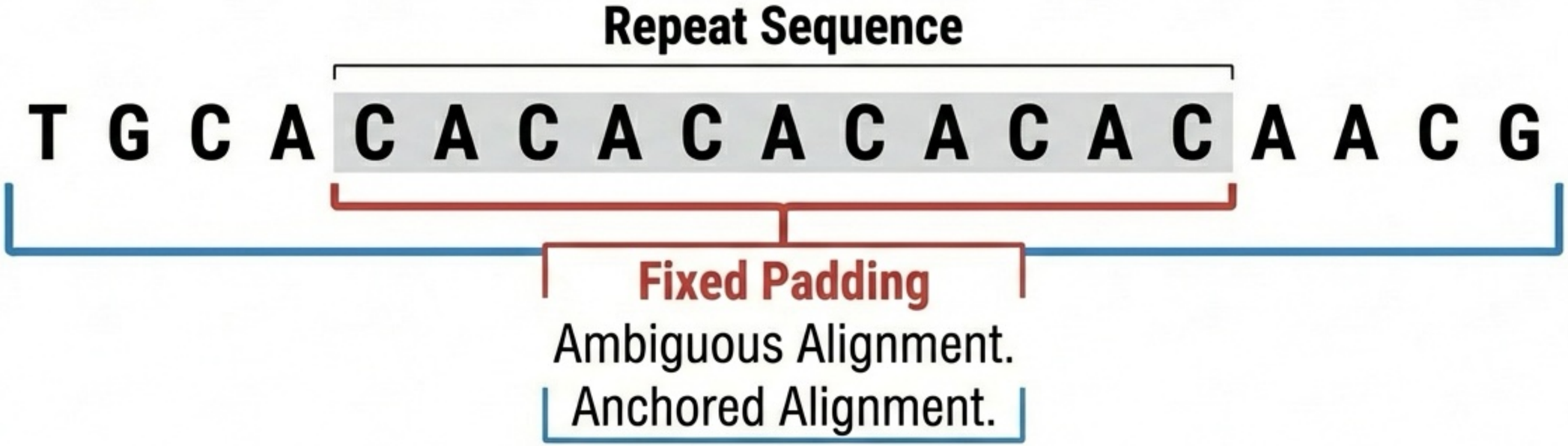
Mechanism: Dual-Counting

1. Construct Corrected Allele (**CCCCT**)
2. Count support for Original (**T**) vs Corrected (**CCCCT**)
3. Winner takes all.

Status:

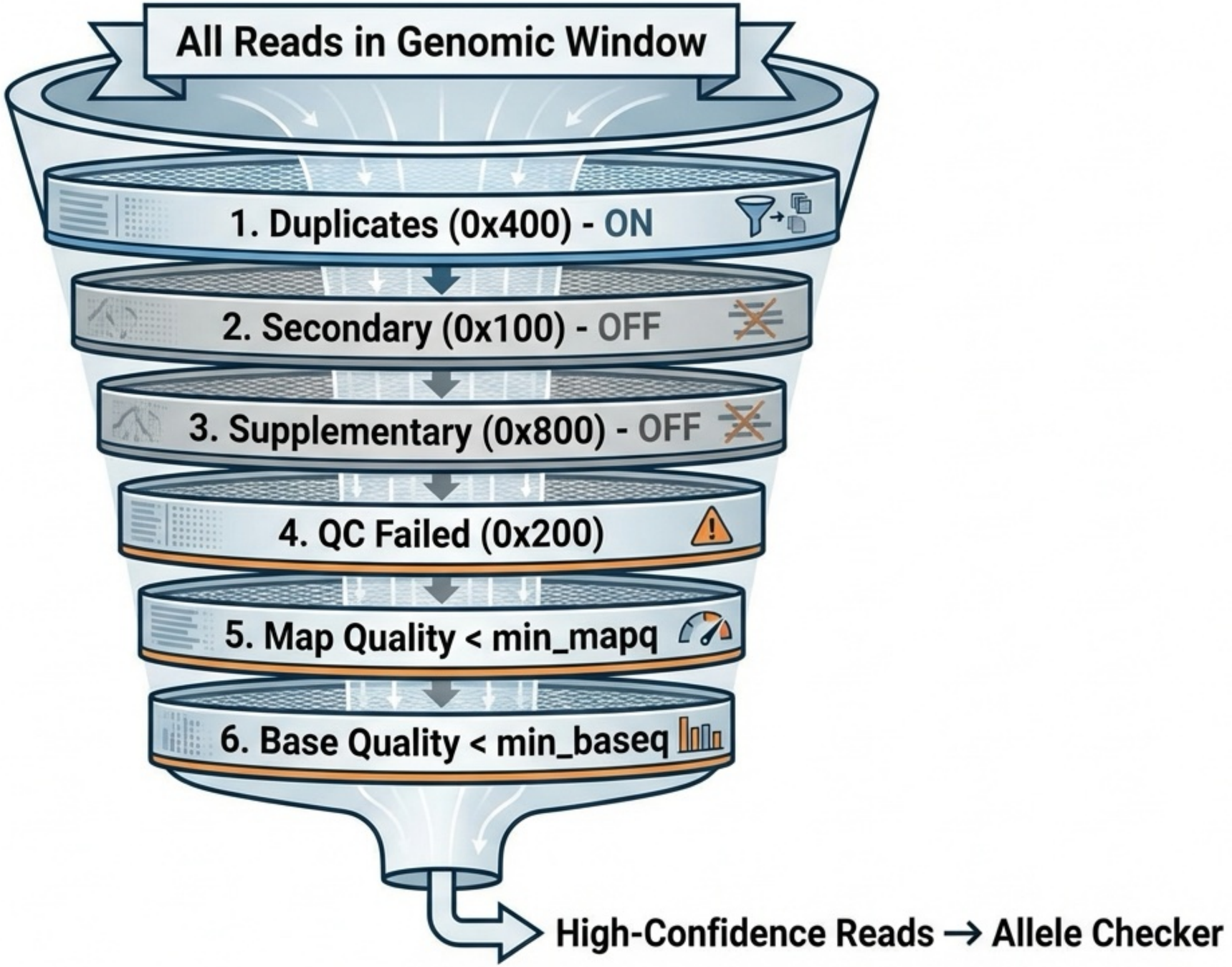
PASS_WARN_HOMOPOLYMER_DECOMP

Adaptive Context & Windowed Scanning

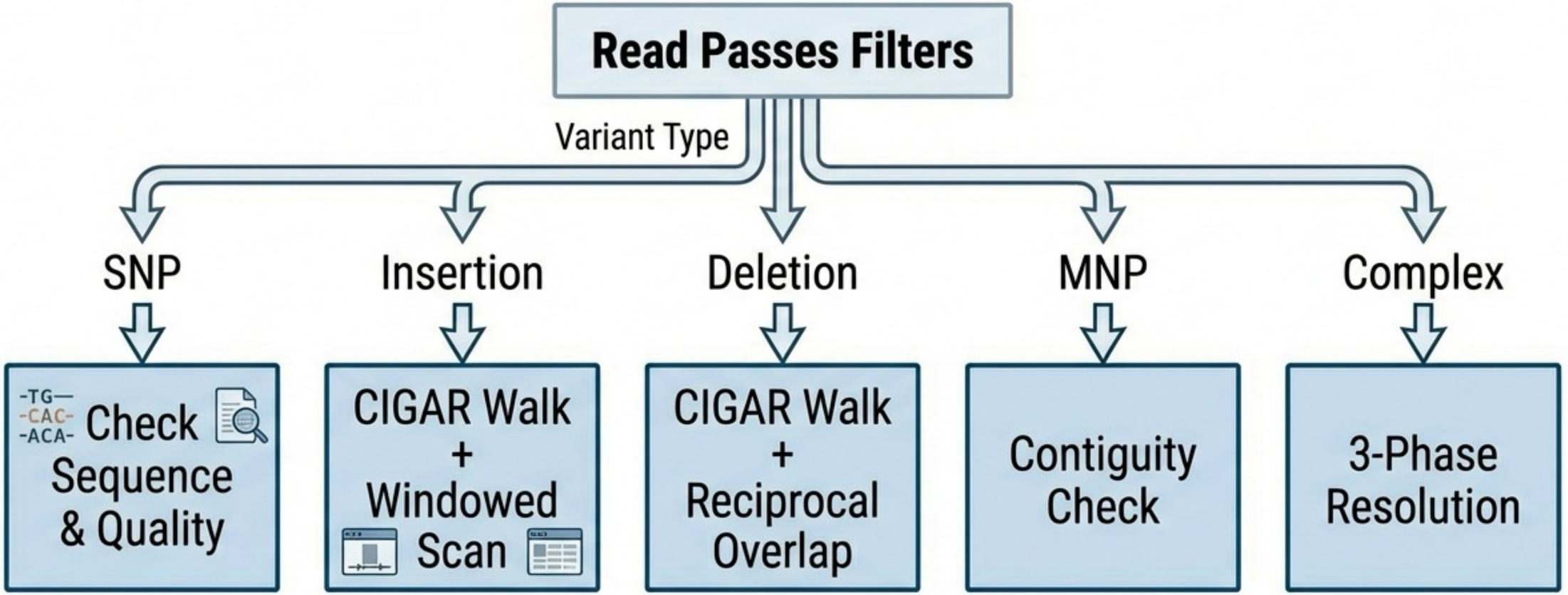



```
padding = max(default, repeat_span / 2 + 3)
```

Result: Expands field of view in repetitive regions.


All Reads in Genomic Window
1. Duplicates (0x400) - ON
2. Secondary (0x100) - OFF
3. Supplementary (0x800) - OFF
4. QC Failed (0x200)
5. Map Quality < min_mapq
6. Base Quality < min_baseq
High-Confidence Reads → Allele Checker

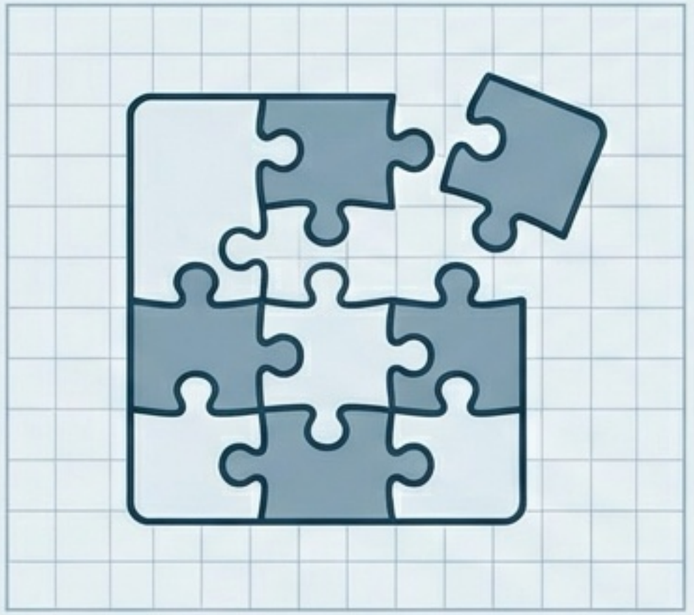
Allele Classification Logic



Auto-detection falls back to Complex check for unknown types.

Deep Dive: Complex Variant Resolution

Phase 1: Reconstruction

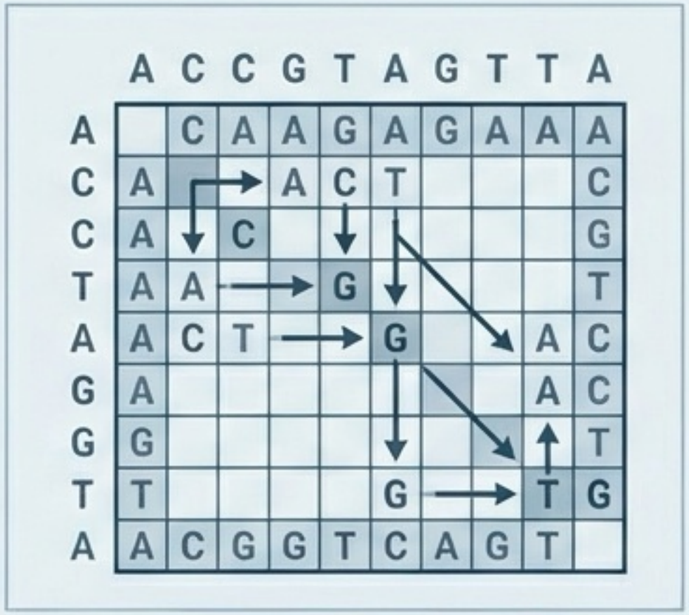


Haplotype Reconstruction
via CIGAR parsing.

Phase 2: Masked Comparison

Quality-aware matching.
Low-quality bases are masked.

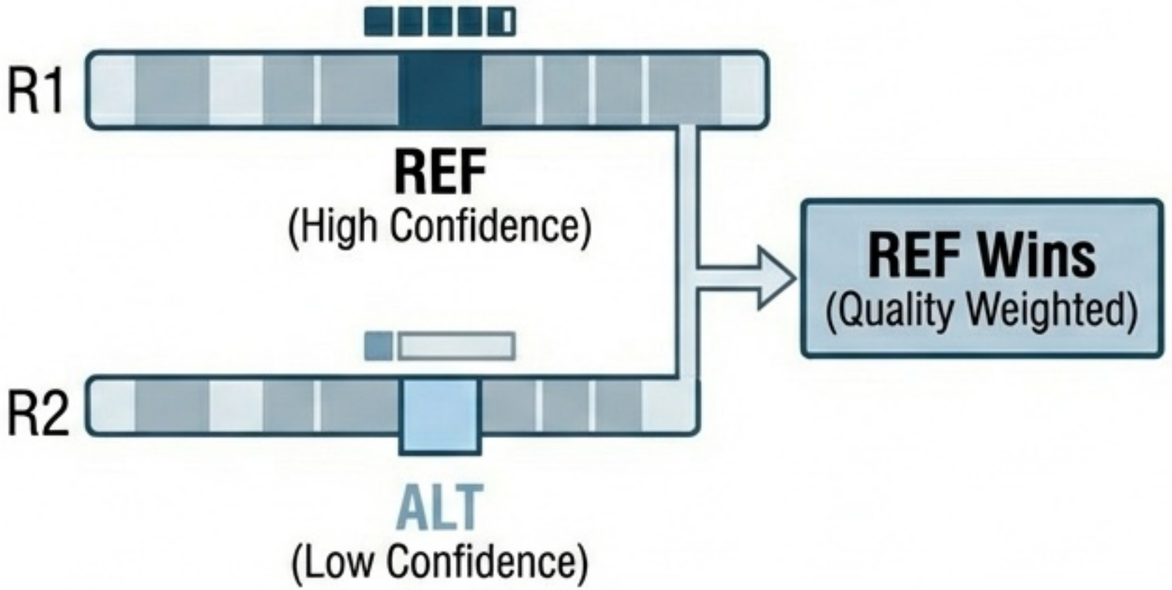
Phase 3: Smith-Waterman



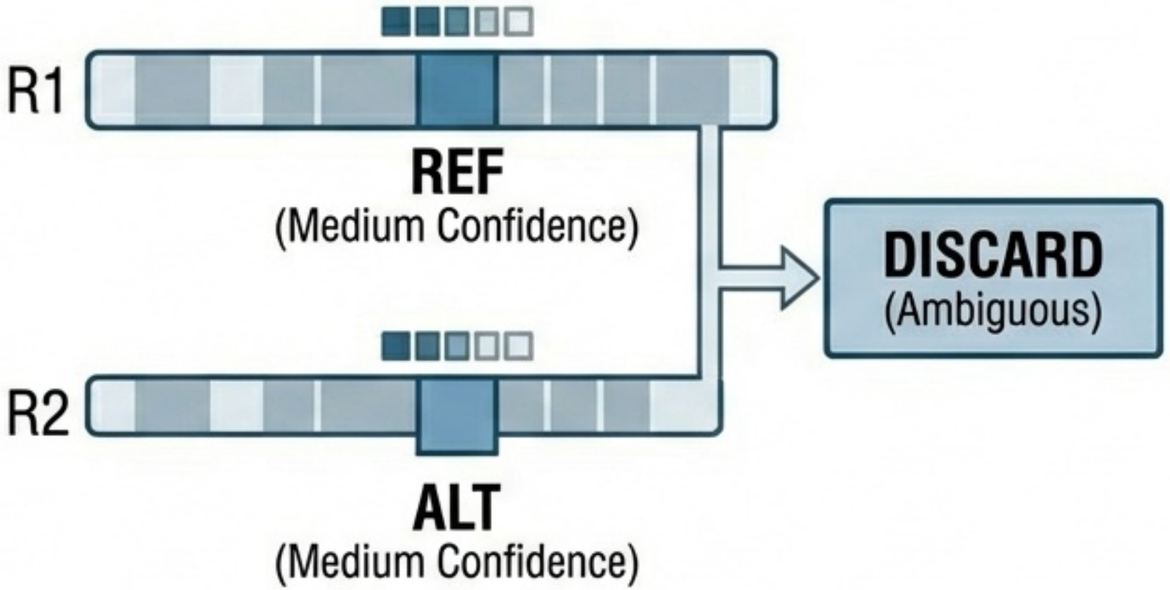
Fallback for difficult alignments.
Semiglobal alignment.

Fragment Counting & Consensus

Case 1: Conflict Resolution



Case 2: Ambiguity



Does NOT default to REF. Prevents VAF deflation.

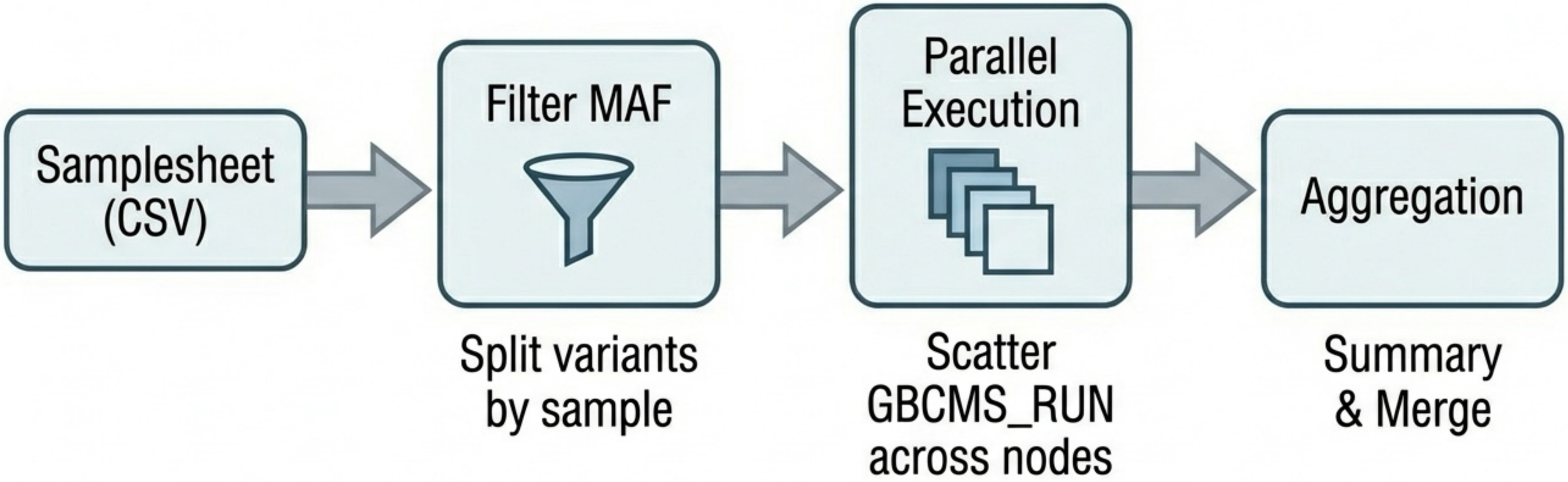
Statistical Rigor: Strand Bias

	Forward Strand	Reverse Strand
Reference Support	a	b
Alternate Support	c	d

Fisher's Exact Test

- Goal: Detect artifacts where variants appear on only one strand.
- Calculated for: Reads (SB_PVAL) and Fragments (FSB_PVAL).

Scaling Up: The Nextflow Pipeline

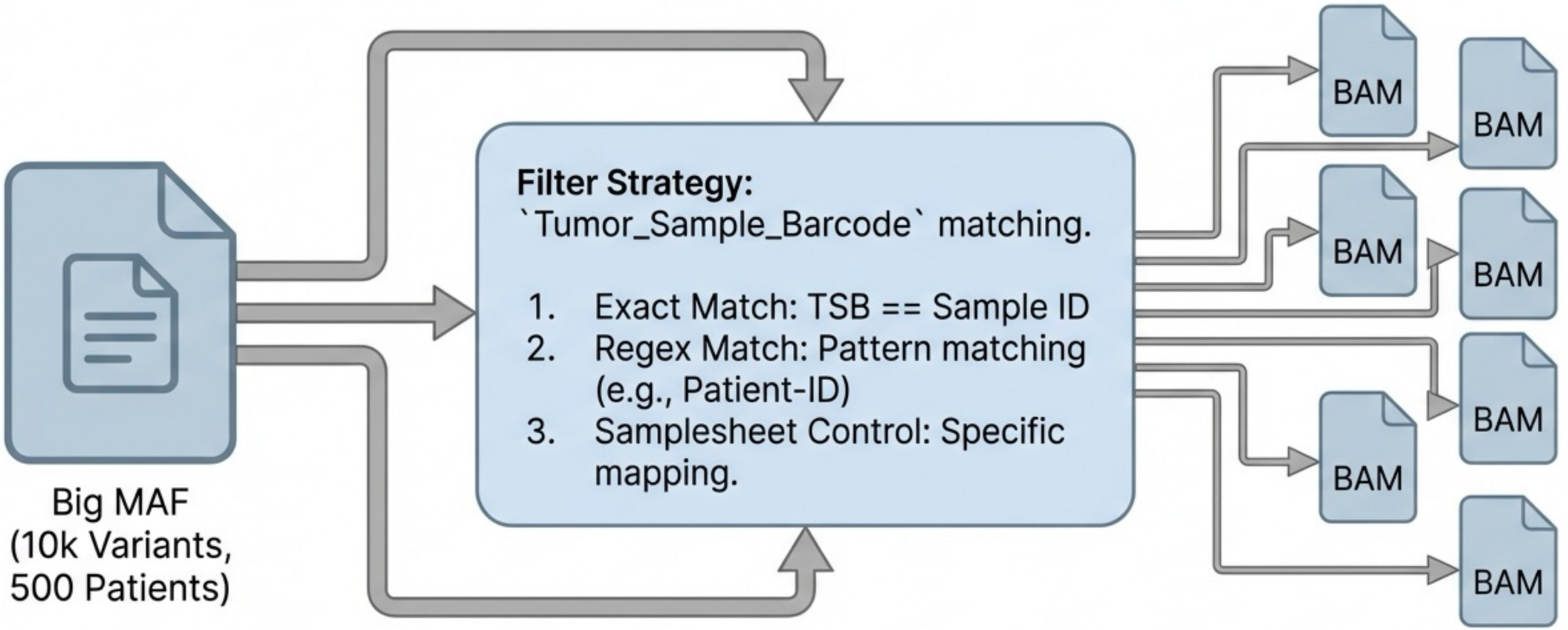


Resume Capability

Docker/Singularity

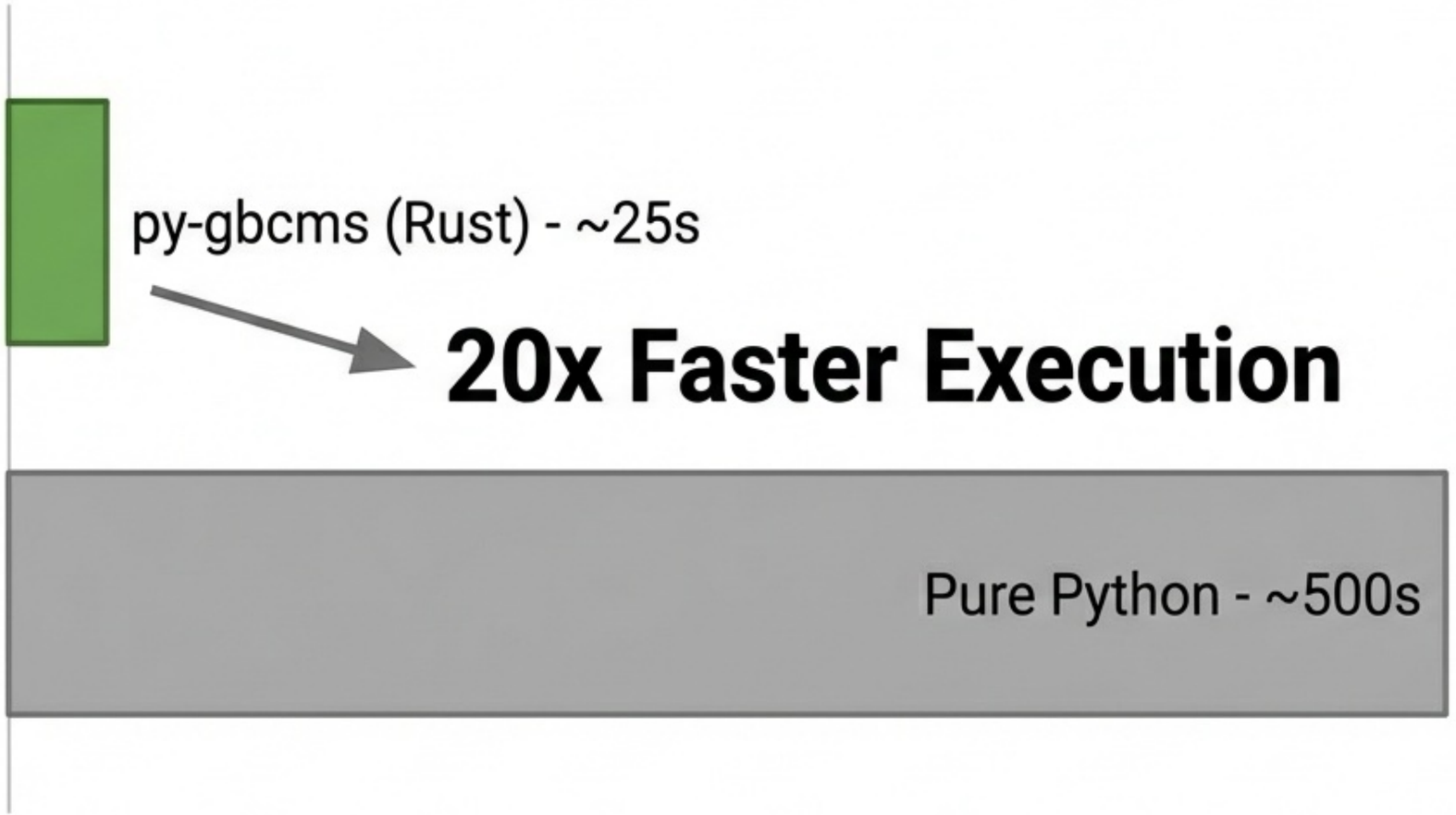
SLURM Integration

Multi-Sample Filtering



Prevents processing irrelevant variants, saving massive compute time.

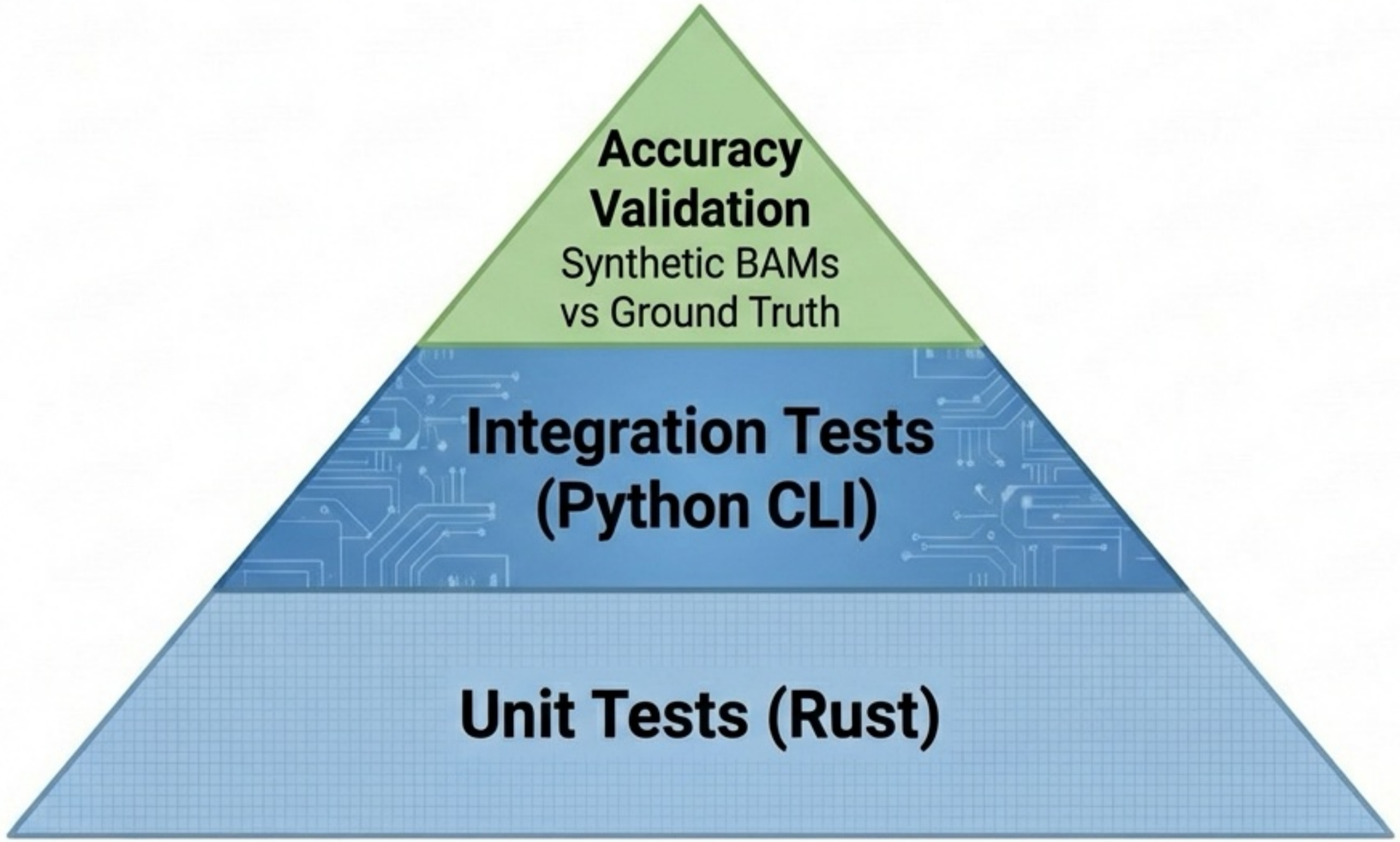
Performance & Benchmarks



Resource Details Box

- **Data:** 1.3GB cfDNA BAM (608 variants)
- **Compute:** 4 CPUs
- **Efficiency:** Rayon parallelization maximizes CPU saturation.

Validation & Quality Assurance



Specific Test Suites

- **test_shifted_indels** (Windowed detection)
- **test_fuzzy_complex** (Masked matching)
- **test_fragment_consensus** (R1/R2 resolution)

Get Started

-  github.com/msk-access/py-gbcms
-  msk-access.github.io/py-gbcms
-  ghcr.io/msk-access/py-gbcms

```
pip install py-gbcms
```

Shah, Ronak H., et al. “py-gbcms: High-performance variant counting from BAM files.” Memorial Sloan Kettering Cancer Center, 2024.