Krewlyzer: The Comprehensive Fragmentomics Toolkit

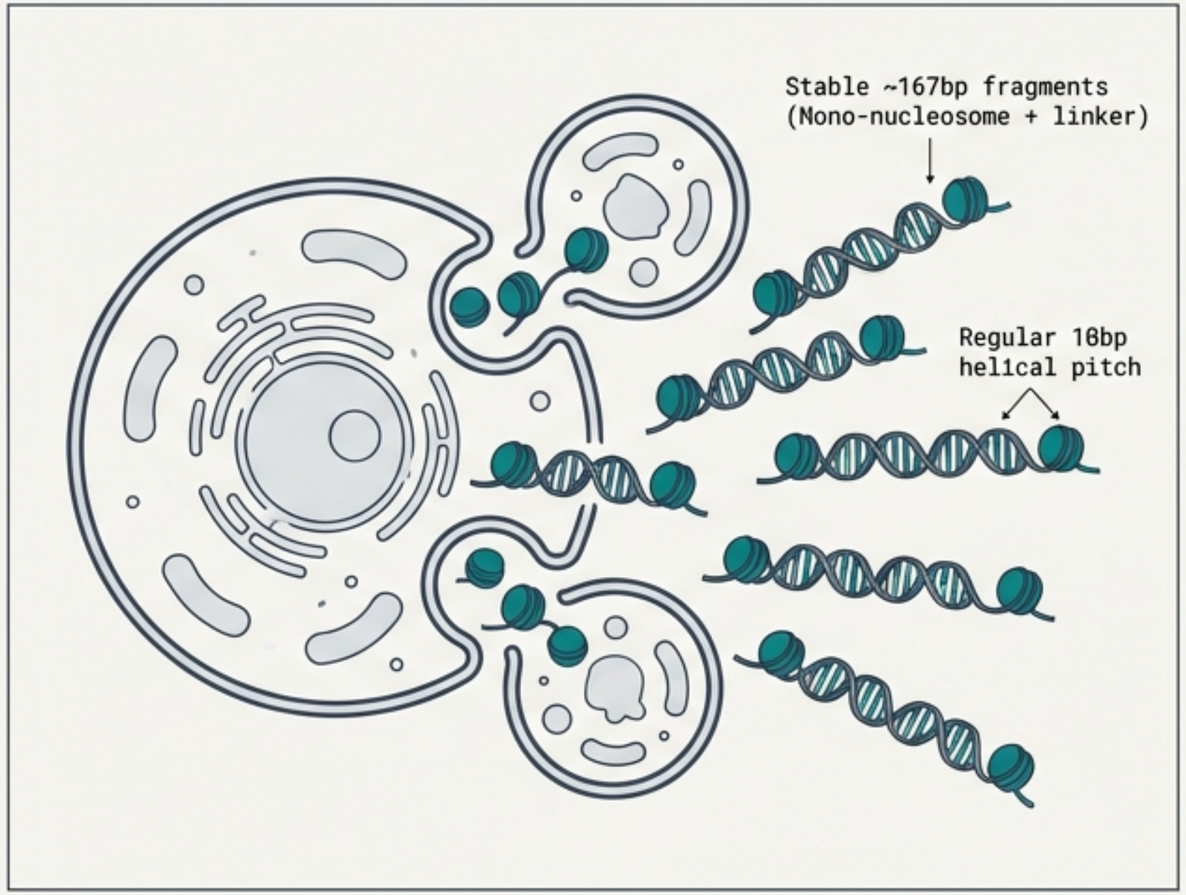
Unlocking the physical biology of cell-free DNA for cancer detection and monitoring

Architecture: Python/Rust Hybrid | **Application:** WGS & Targeted Panels | **Version:** 2.0

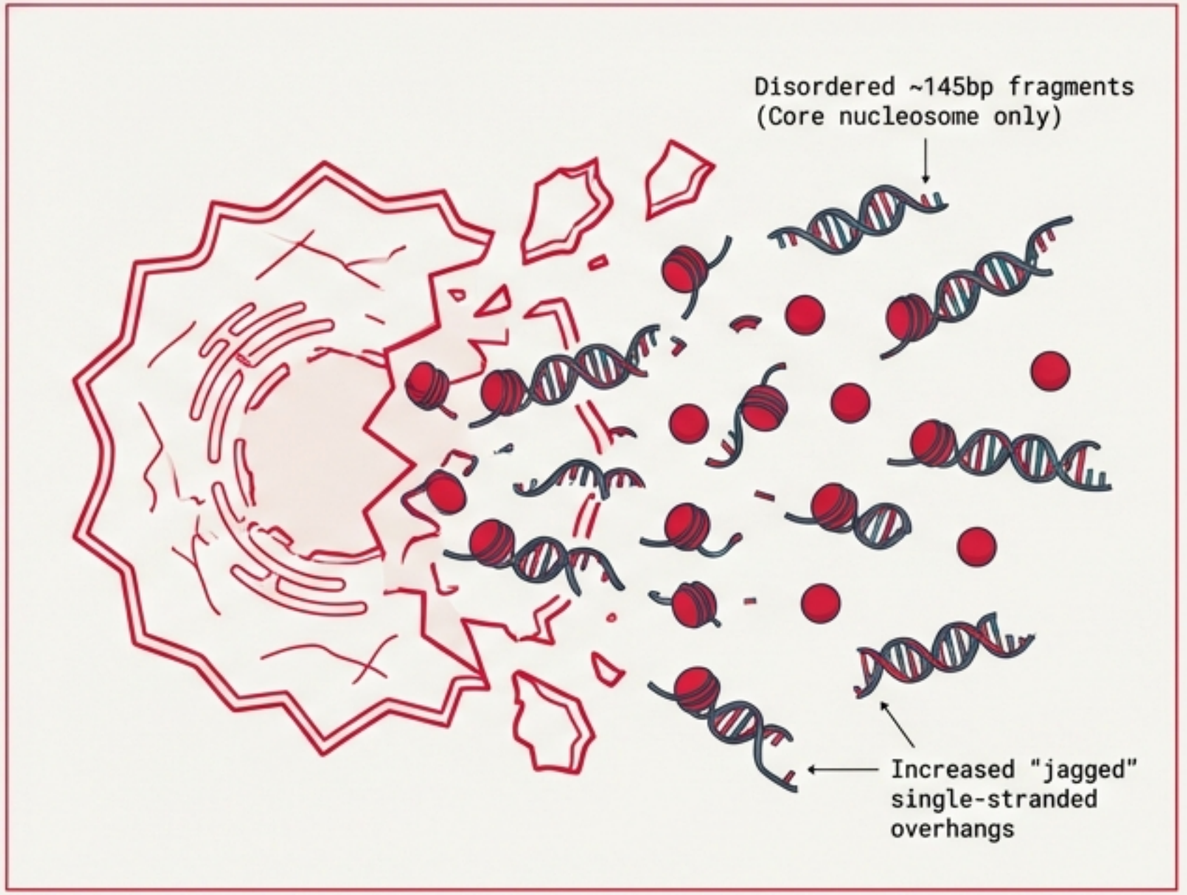
DNA is more than a sequence of letters; it has a physical shape.

Standard sequencing reads mutations (typos).
Fragmentomics reads structure—nucleosome positioning, chromatin accessibility, and enzymatic cutting patterns.

Healthy Cell Turnover (Apoptosis)



Tumor Cell Turnover (Necrosis/Apoptosis)



Krewlyzer transforms raw sequencing data into biological features.

A unified engine that distills terabytes of BAM alignments into compact, ML-ready metrics.

Versatility

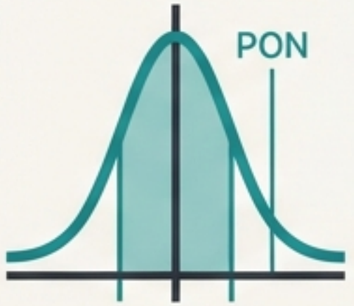
Works on Whole Genome Sequencing (WGS) and Targeted Panels (e.g., MSK-ACCESS).

Performance

Rust backend for single-pass processing. Reads massive BAM files once, extracts all features simultaneously.

Correction

Built-in **GC bias removal** using LOESS regression applied per fragment length bin.



Normalization

Z-score standardization against a **Panel of Normals** (PON) to define biological baselines.

BAM **Input:** BAM File BAM → `krewlyzer run-all` → **Output:** Unified JSON & Parquet JSON Parquet

A multi-dimensional view of the cancer genome.

Feature Landscape

Fragmentation (Size & Ratios)

FSC: Fragment Size Coverage (6 channels)

FSR: Fragment Size Ratios (Short vs. Long)

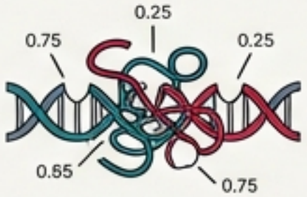
FSD: Size Distributions per chromosome arm

mFSD: Mutant-specific fragment sizes

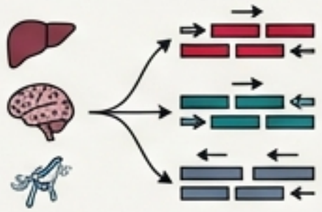
Chromatin Structure

WPS: Windowed Protection Scores (Nucleosomes)

Region Entropy: Chromatin chaos at TFBS/ATAC sites

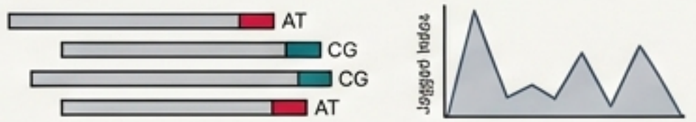


OCF: Orientation-aware fragmentation (Tissue of origin)



Sequence & Modification

Motif: End motif diversity (MDS) & Jagged Index

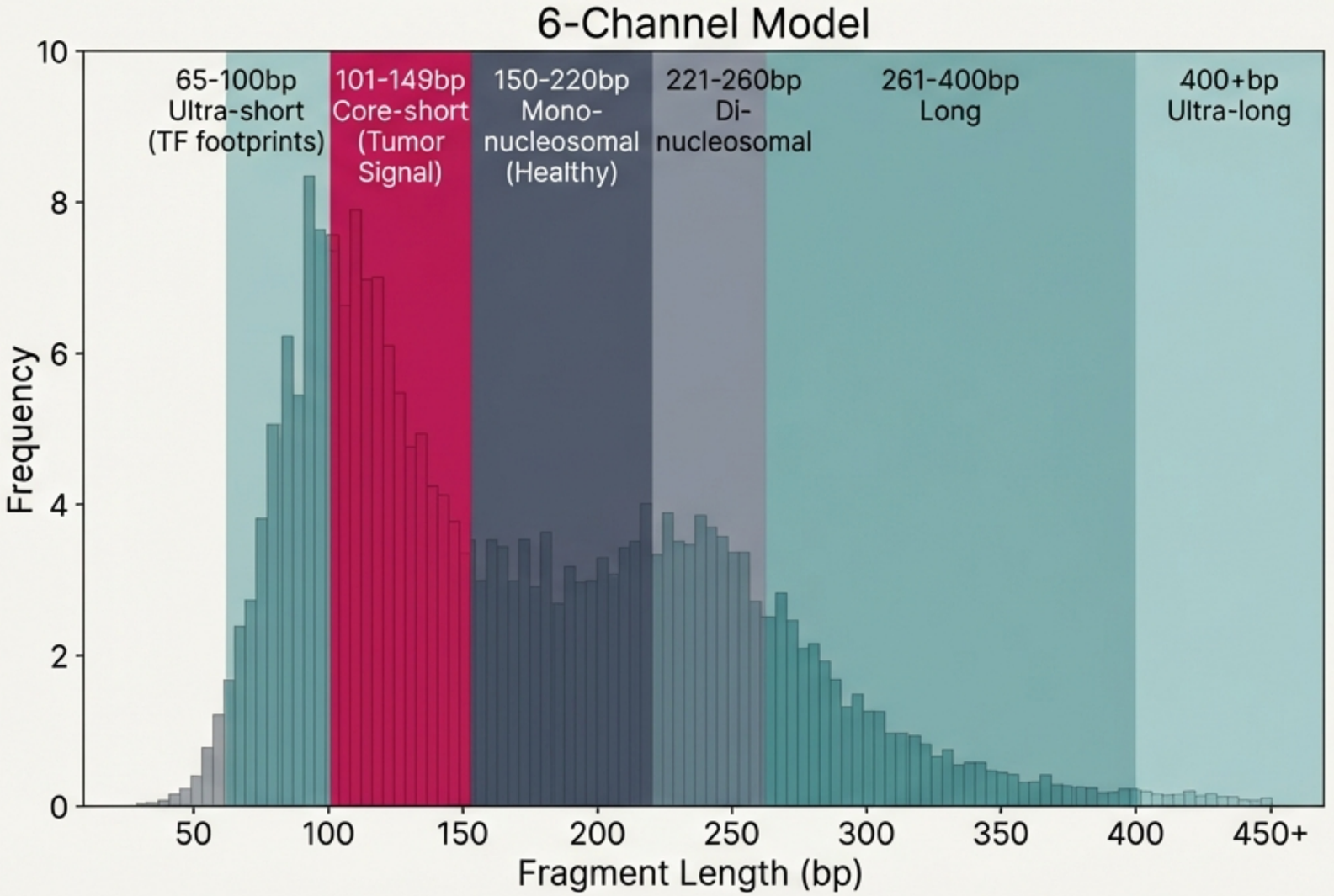


Region MDS: Per-gene motif diversity

UXM: Methylation deconvolution

Quantifying the ‘Left-Shift’: Fragment Size Analysis

FSC (Coverage) and FSR (Ratios) detect tumor burden via size anomalies



Key Metric:

`core_short_long_ratio`

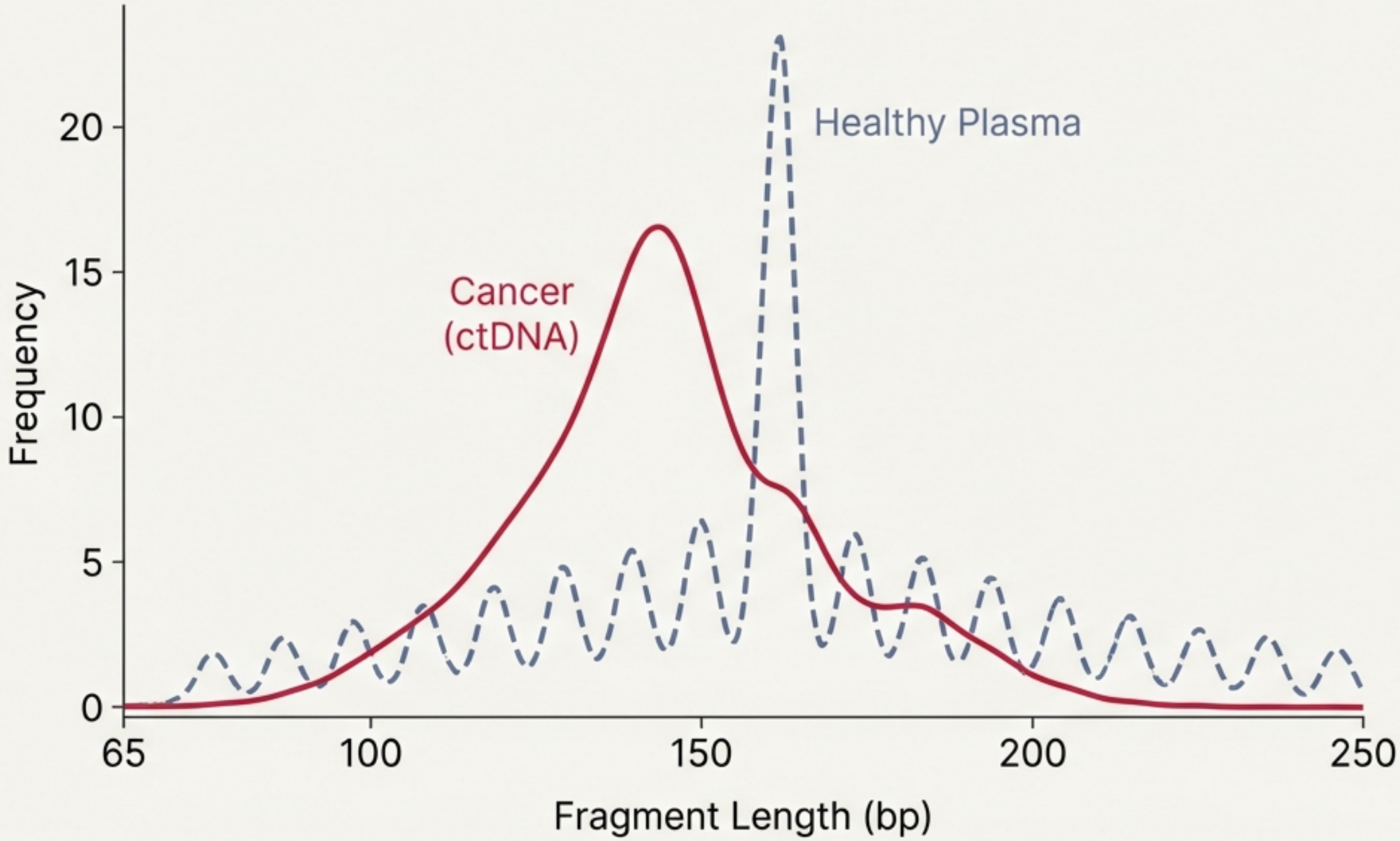
Healthy Ratio:
~0.8 – 1.0

Cancer Ratio:
> 1.2 (Elevated)

Counts are GC-corrected using LOESS regression before ratio calculation.

High-resolution distributions detect aneuploidy.

FSD creates a ‘fingerprint’ histogram for every chromosome arm.

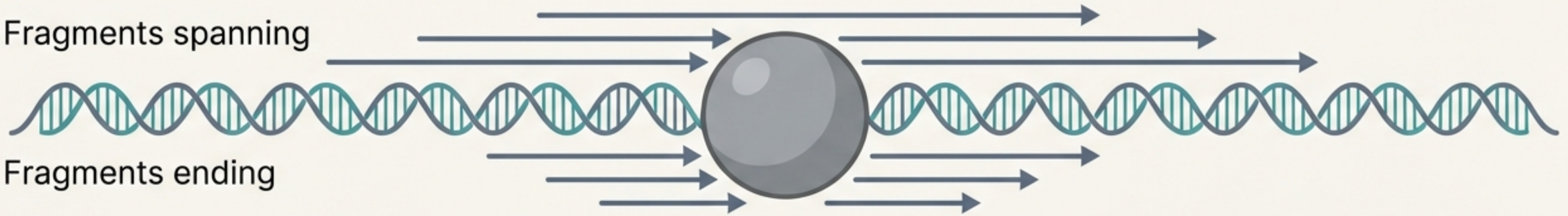


The Signal:

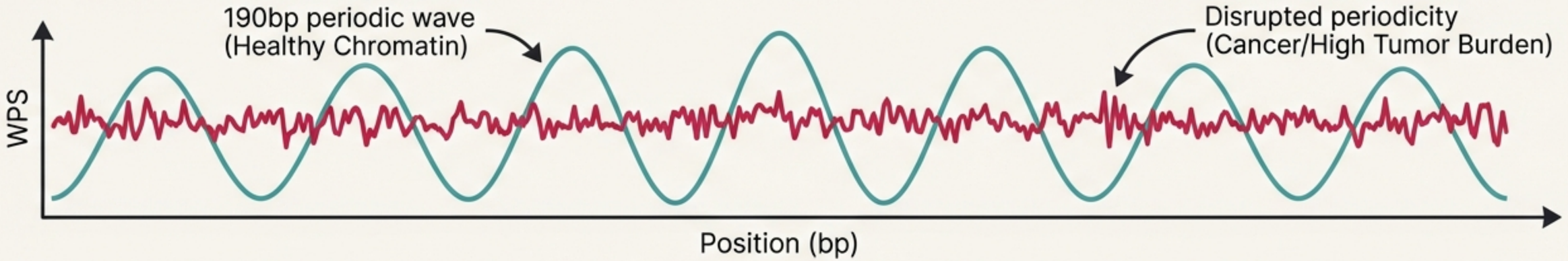
- Left-shifted peak toward 145bp
- Disrupted 10bp helical periodicity
- Arm-level deviations indicate Copy Number Alterations (CNAs)

Mapping the physical protection of the genome.

WPS (Windowed Protection Score) reveals protein footprints on DNA.




```
WPS = Spanning - Ending
```

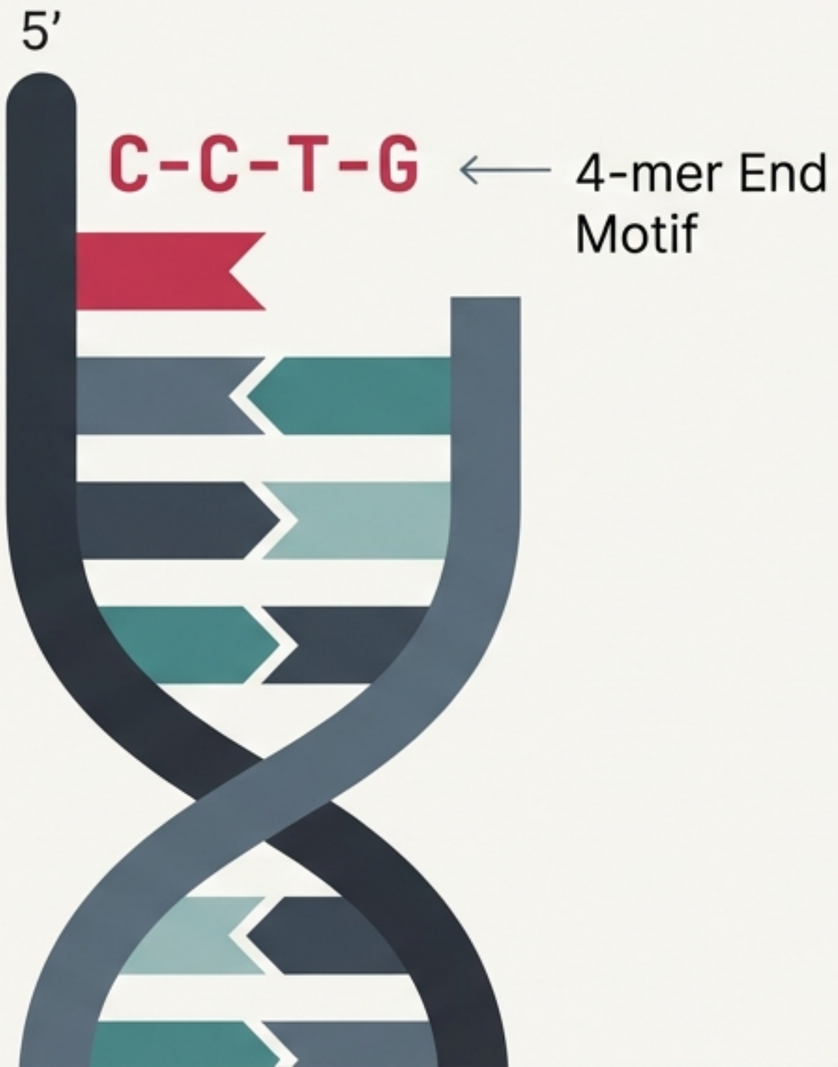


Dual Output Streams

1. **Foreground:** Gene-specific profiles (TSS/CTCF)
2. **Background:** Global Alu element stacking (Chromatin Health)

The molecular signature of enzymatic cutting

How the DNA was cut reveals its origin.



MDS (Motif Diversity Score)

High Diversity (Random Cutting).

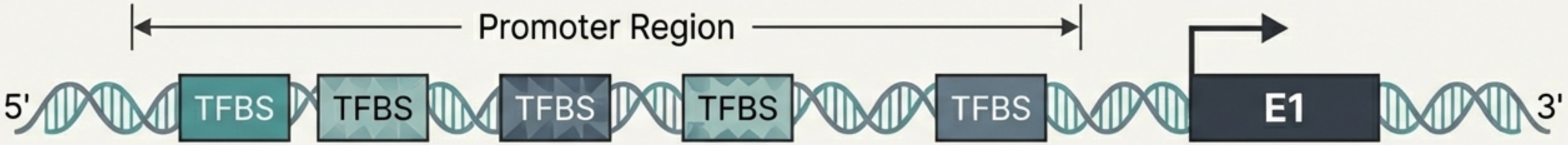
Low Diversity (Stereotyped Cutting).

The Jagged Index

- Tumor DNA contains more single-stranded overhangs.
- **Signal: ~87.8%** jagged ends in tumor-derived fragments.

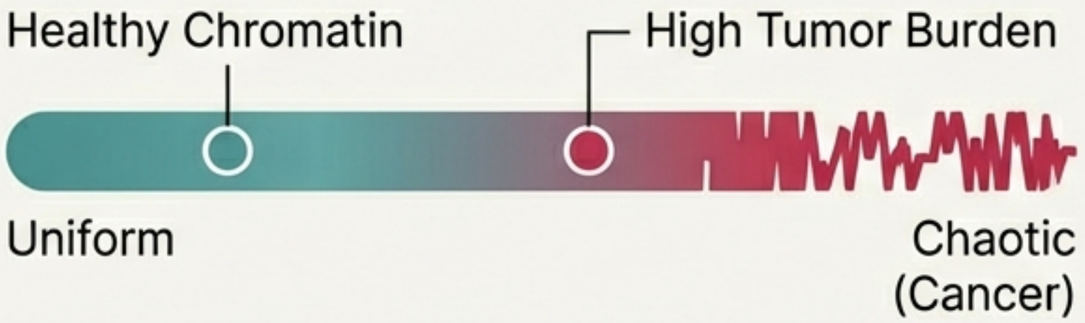
Detecting chromatin chaos at regulatory regions.

Based on Helzer et al. (2025).



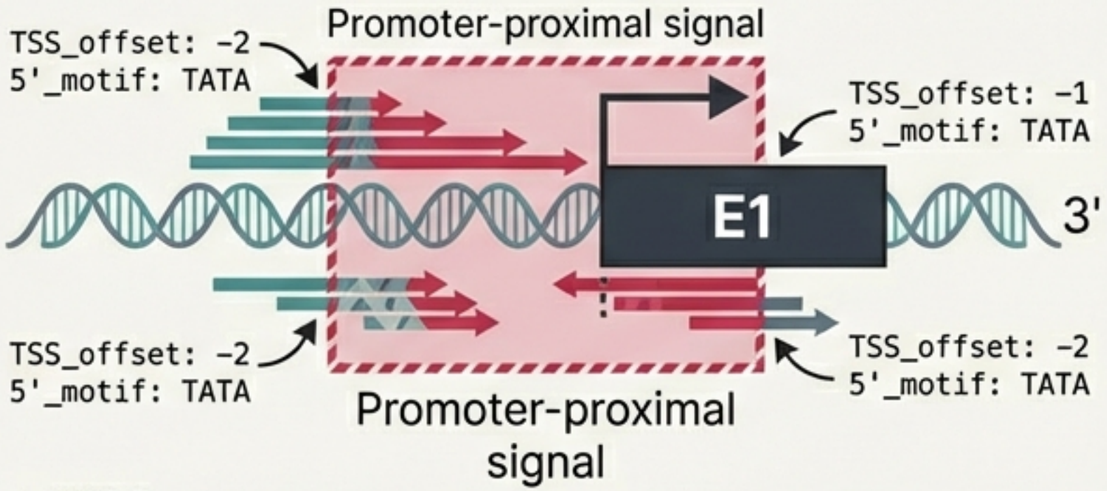
Region Entropy

Measures fragment size diversity at 808 Transcription Factor sites.



Region MDS

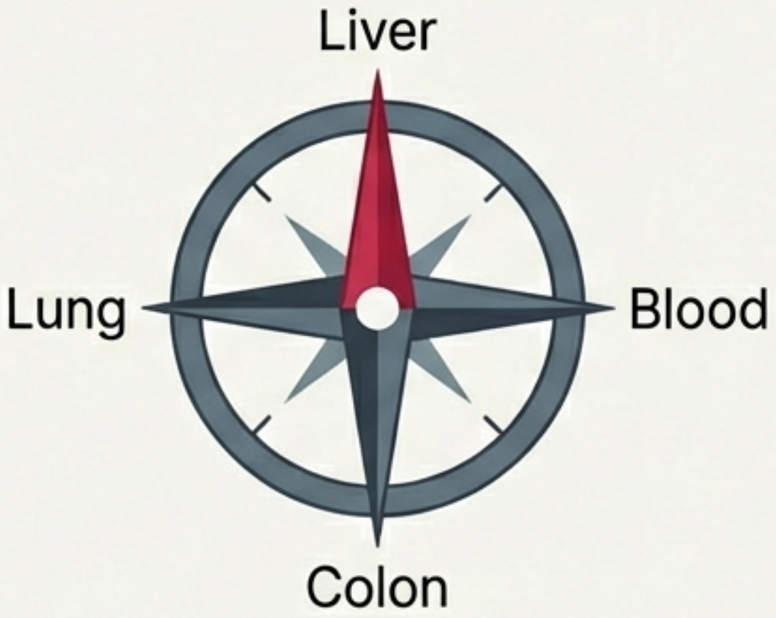
Pinpoints motif anomalies at the first exon (E1).



Panel Ready: Works effectively on commercial panels without WGS.

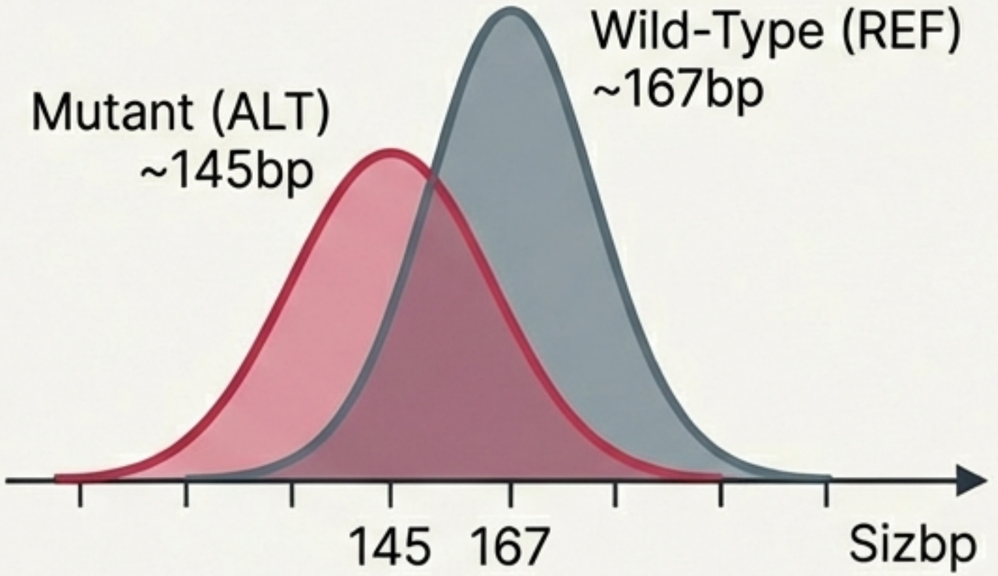
Pinpointing origin and validating variants.

Tissue of Origin (OCF)



Analyzes fragment phasing at open chromatin regions to deconvolve tissue sources.

Mutant-Specific Analysis (mFSD)

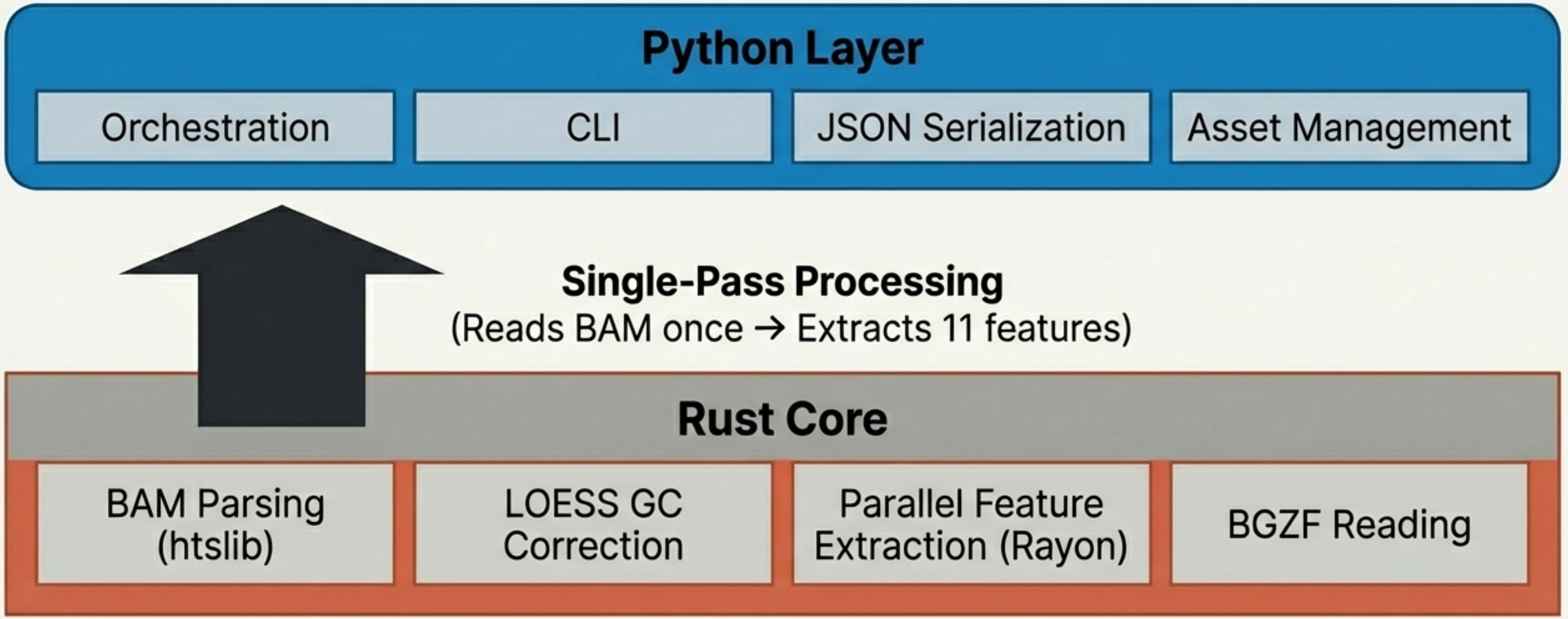


Compares size of Mutant vs. Wild-type fragments at specific variant sites.

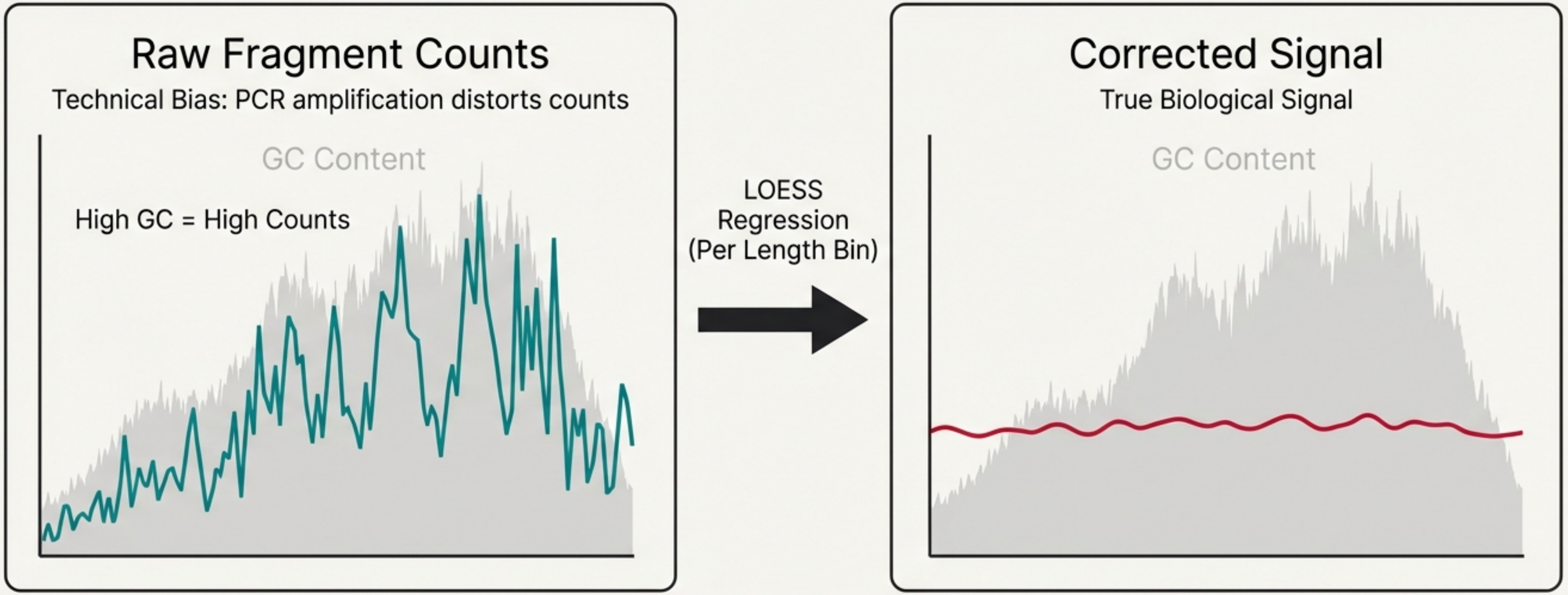
Includes Duplex Support for ultra-sensitive MRD monitoring.

Engineered for performance and scale.

The hybrid architecture combines Python flexibility with Rust performance.



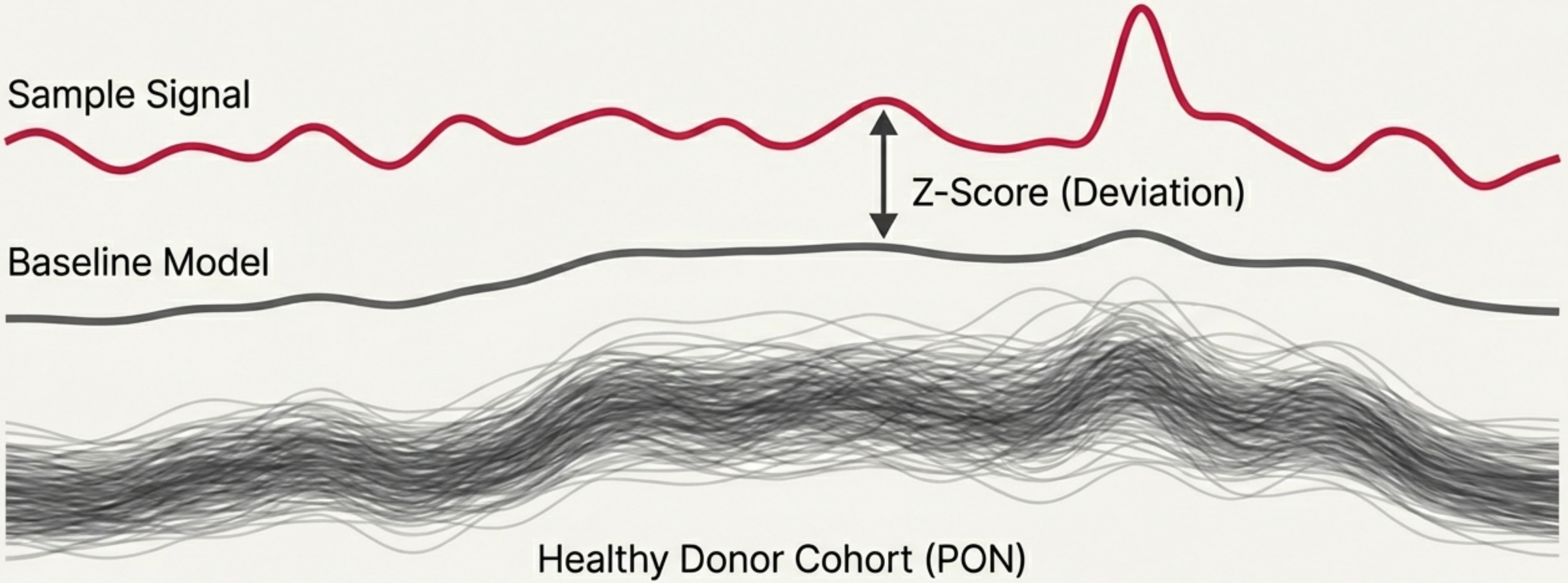
Biological signal requires rigorous bias correction.



Hybrid Correction: Removes both assay-wide systematics and sample-specific batch effects

Defining ‘Normal’ to detect the abnormal.

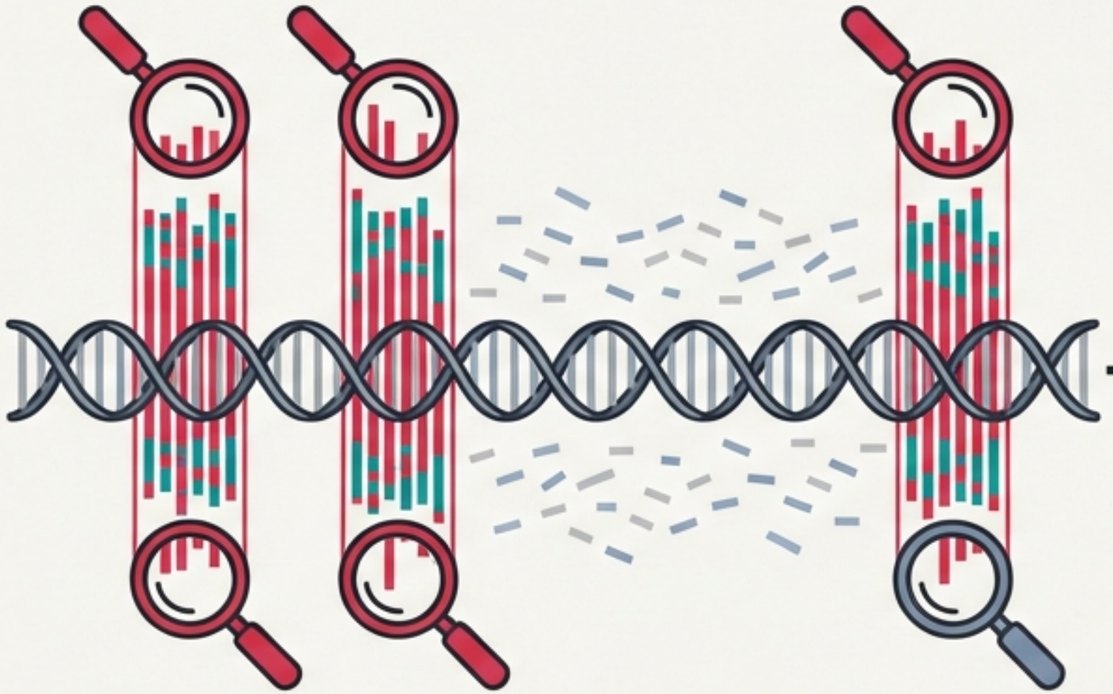
The Panel of Normals (PON) creates a statistical baseline for Z-scores.



All features (FSC, WPS, MDS) are normalized against the PON to distinguish signal from noise.

Panel Mode: Global insights from targeted sequencing

Krewlyzer utilizes off-target data to make panels act like WGS



On-Target
(Deep Coverage)

Off-Target
(Background Waste)

Split Analysis Strategy

On-Target Data

Gene-Specific Features
(CNV, Local OCF)

Off-Target Data

GC Model Training → **Global Biomarkers**
(MDS, Genome-wide Signal)

We use the "waste" data to build unbiased GC models and extract global cancer signals.

Two distinct biological signals from one assay

Input BAM (MSK-ACCESS)

Stream A: Off-Target (Unbiased)

- Genome-wide MDS
- Global OCF
- Background WPS

Use Case:
Tumor Fraction estimation,
Global fragmentation profiles.

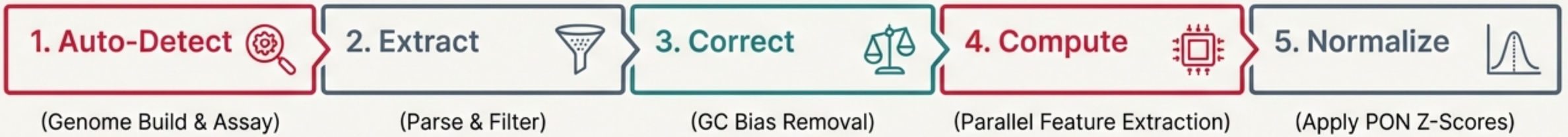
Stream B: On-Target (High-Depth)

- Gene-specific Copy Number
- Local OCF
- Gene-level FSC (146 genes)

Use Case:
Driver gene analysis (e.g.,
EGFR amplification).

From raw alignments to insights in one command.

```
> krewlyzer run-all --input sample.bam --output results/
```



Nextflow Integration: Ready for Cluster/Cloud Batch Processing.

Structured outputs built for Machine Learning.

Unified JSON

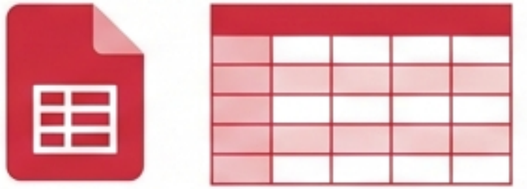
`{sample}.features.json`

Single file containing all scalar features, z-scores, and metadata for easy ML ingestion.

Parquet

WPS Profiles

Efficient storage for high-dimensional vectors (200 bins).

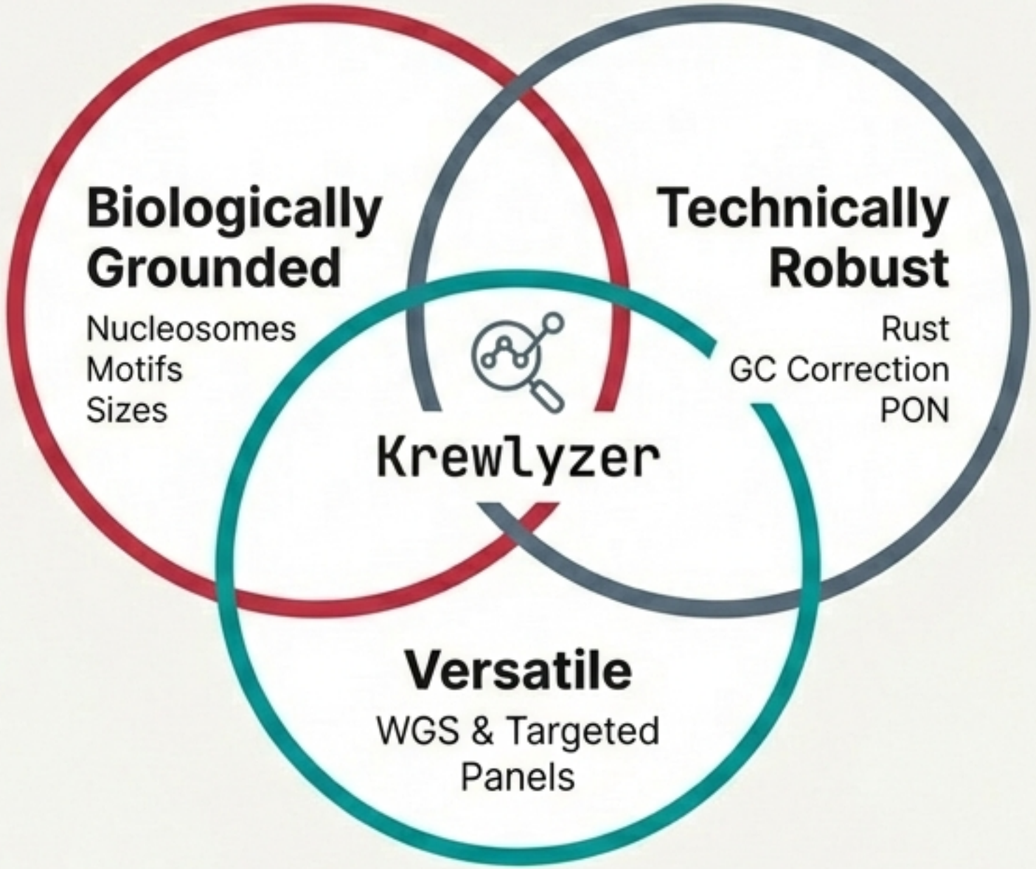


TSV

Gene-Level Data

FSC.gene.tsv with normalized depth and z-scores.

A comprehensive lens for liquid biopsy.



Moving from “Is there a mutation?” to
“What is the physical state of the cancer genome?”