

Background

Comprehensive genomic profiling (CGP) panels are widely used in translational studies and clinical trials to support patient stratification, exploratory biomarker analysis, and retrospective molecular characterization. However, secondary analysis and reporting workflows are often delivered as tightly coupled, proprietary components of commercial assay ecosystems. This coupling limits transparency, reproducibility across sites, and the flexibility to adapt analytical methods as study designs and sequencing technologies evolve.

We present **twistcgp**, an open-source bioinformatics workflow for analysis of data generated with the Twist Oncology DNA CGP Panel that targets 562 cancer-associated genes across 2.4 Mb with the ability to customize content. This collaboration pairs a focused hybrid-capture CGP assay with a transparent, configurable, and portable analysis framework that operates independently of licensed software, proprietary execution environments, or fixed reporting layers.

To evaluate pipeline performance, Horizon Discovery, Seracare, and Twist reference standards (with technical replicated) were evaluated with the Twist CGP Panel.

Across all samples, the workflow generated expected classes of somatic variation with standardized quality metrics suitable for cross-sample and cross-site comparison. Known variants in reference materials were detected at anticipated VAFs, and TMB and MSI estimates were consistent with expectations, derived directly from panel data without reliance on proprietary analysis software.



Check out our Github repository for code, instructions, and a full list of citations!
<https://github.com/fulcrumgenomics/twistcgp>

CGP Panel Details

Example sequencing QC metrics averaged across data downsampled to 2000x raw coverage (32M 2x150 reads). 4.5 Gb of data is enough for VAF detections ranging ~2%. Data available upon request.

Example Sequencing Metrics	
Panel target size	2.4 Mb
Example input	50 ng
Mean Target Coverage	515x
On-Target Rate	77%
Fold-80 Base Penalty	1.32
Duplication Rate	20%
Target bases covered >100x	99.5%

Panel: 562 Genes - SNV + Hotspots

MSI (~50 sites)

Selected Fusions

CNV (57 Genes)

TMB

Panel + Software Customization

Differentiate biomarkers by adding custom content blended into panel

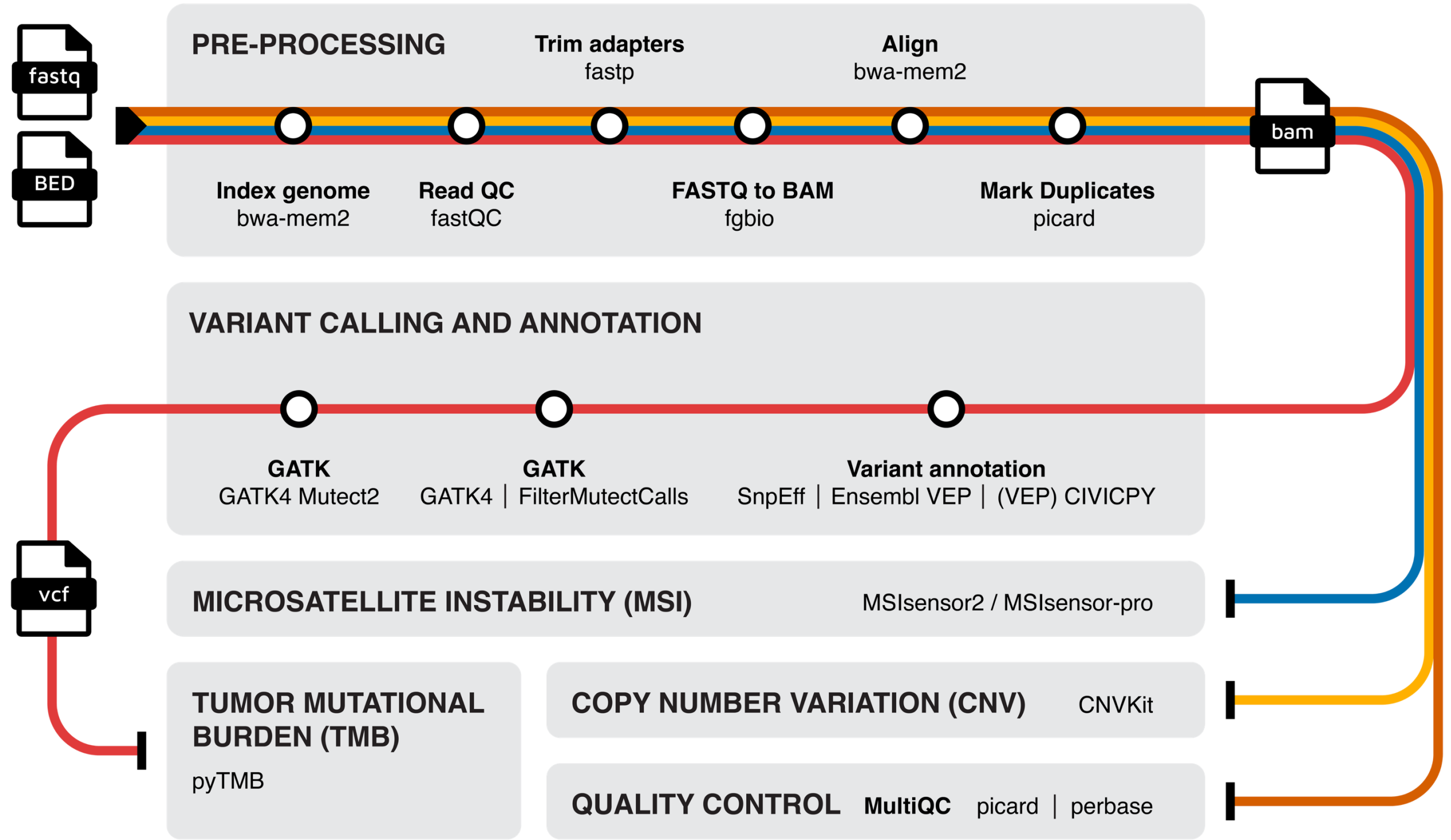
End-to-end NGS workflows include:

- Twist Library Preparation
 - Mechanical Fragmentation, cfDNA
 - Enzymatic Fragmentation (TrueAmp Kit)
- Hybridization Solutions



Twistcgp Pipeline

Fig.1 Pipeline Overview



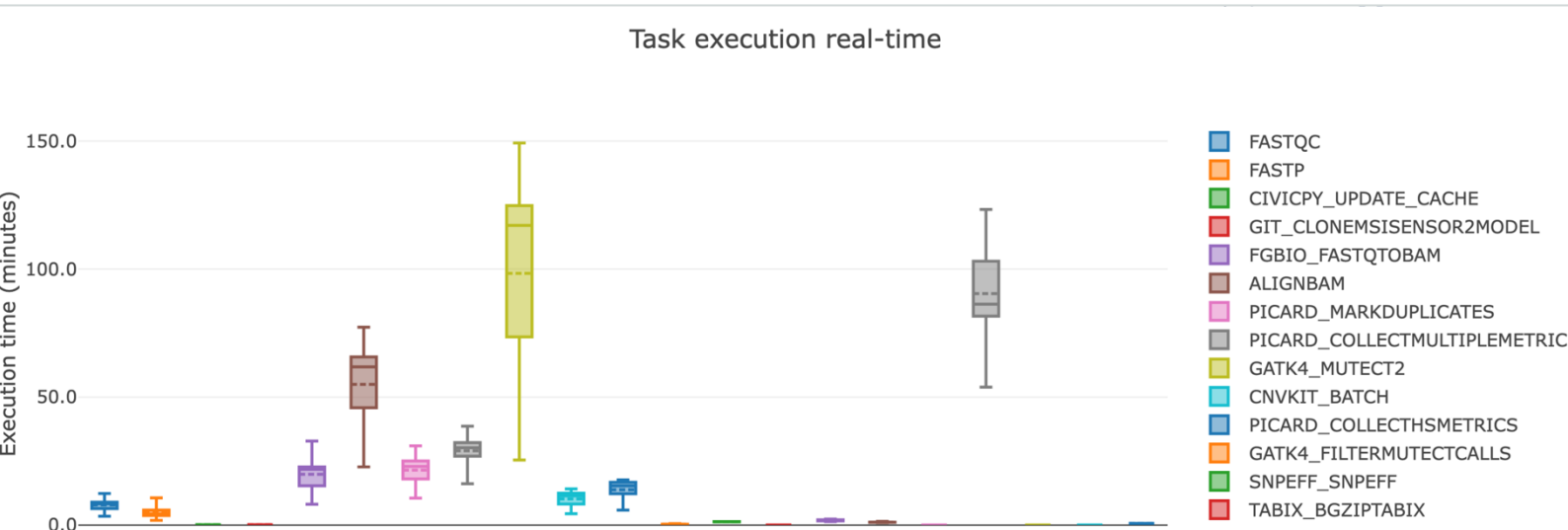
Deployment Modes

twistcgp runs anywhere Nextflow runs. Select a container or package manager with `-profile`.

Execute locally on a single node, on an HPC cluster with Slurm, at scale on AWS Batch, or via Seqera Platform for managed orchestration and monitoring.



Fig 2. Pipeline Artifacts

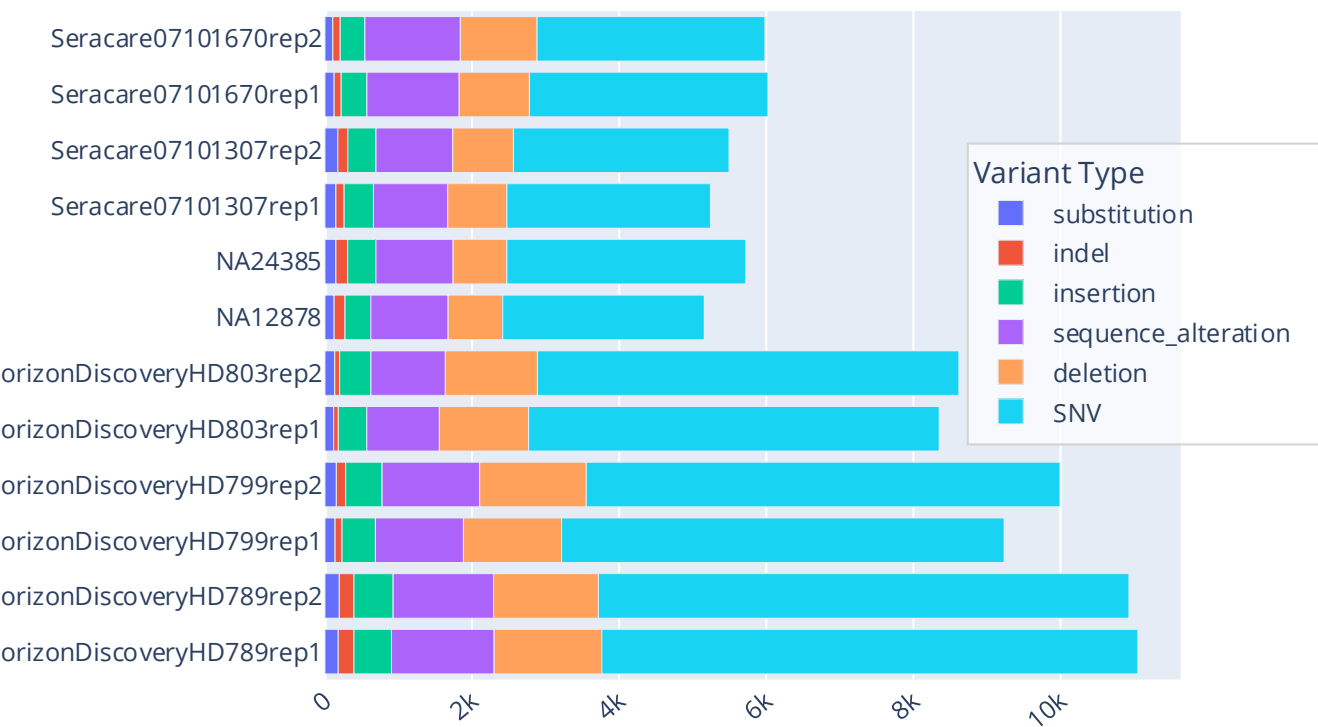


Every run produces a software manifest, an execution report with per-task CPU, memory, and wall time usage, and a timeline visualization of task parallelism across the run.

The total run time for all 12 samples on an AWS EC2 m5.16xlarge instance (~\$3/hour) was ~20 hours

Fig 3. Pipeline Results

Variant Annotation



Sample Name	Overlapped transcripts	Overlapped genes	Existing variants	Novel variants	Variants processed
HorizonDiscoveryHD789rep1	12742	1636	9286	1768	11062
HorizonDiscoveryHD789rep2	12741	1632	9194	1737	10931
HorizonDiscoveryHD799rep1	12520	1592	7508	1726	9234
HorizonDiscoveryHD799rep2	12677	1627	8152	1841	9993
HorizonDiscoveryHD803rep1	12385	1564	6856	1495	8351
HorizonDiscoveryHD803rep2	12383	1569	7002	1617	8619
NA12878	10349	1299	4292	865	5157
HorizonDiscoveryHD803rep1	10828	1337	4607	1113	5720
HorizonDiscoveryHD799rep2	10564	1281	4351	890	5241
HorizonDiscoveryHD799rep1	10625	1321	4523	971	5494
HorizonDiscoveryHD789rep2	11299	1422	4788	1235	6023
HorizonDiscoveryHD789rep1	11253	1424	4819	1163	5982

twistcgp uses multiple complementary variant annotation tools:

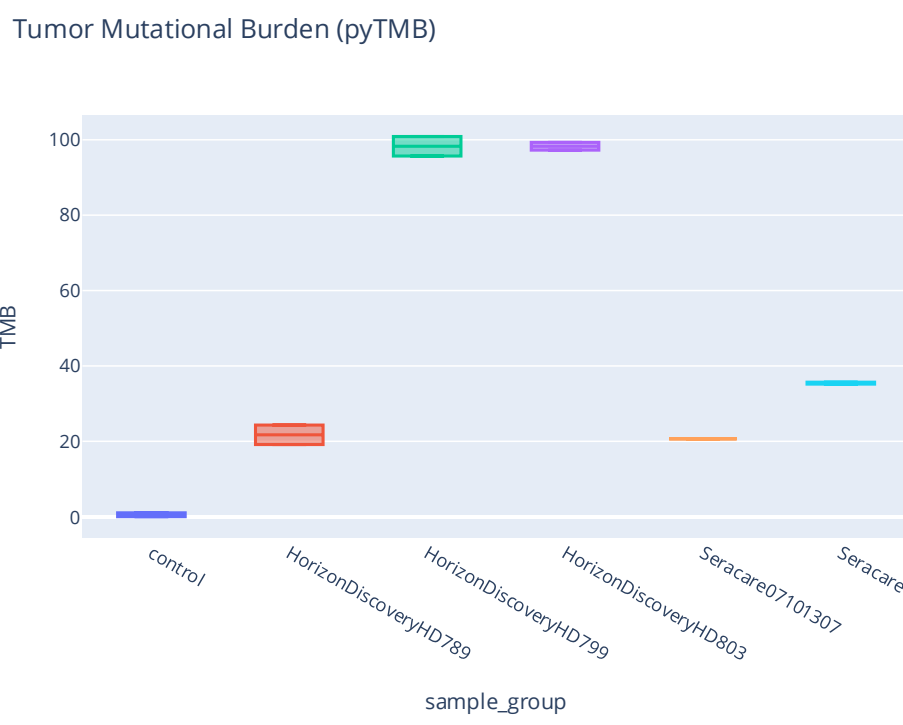
- SnpEff: functional effect prediction
- Ensembl VEP: POPAF, consequence annotation
- CIViCpy: clinical interpretation of variants in cancer

Sample Name	Fold enrichment	On-target bases	Usable bases on-target	Median target coverage	PF bases	PF bases aligned	PF reads
HorizonDiscoveryHD789rep1	708.2	3,631.5Mb	16.8%	1714.0	23,128.7Mb	23,079.7Mb	164.1M
NA12878	715.0	3,627.9Mb	21.6%	1914.0	16,760.7Mb	16,724.3Mb	115.1M
Seracare07101307rep2	695.8	3,300.8Mb	20.4%	1732.0	16,182.7Mb	16,150.6Mb	110.7M
NA24385	650.3	3,204.4Mb	19.9%	1698.0	16,079.8Mb	16,042.8Mb	109.2M
Seracare07101307rep1	760.7	2,787.8Mb	14.4%	1445.0	19,335.2Mb	19,292.7Mb	141.2M
HorizonDiscoveryHD789rep2	716.0	2,893.2Mb	17.5%	1490.0	16,559.9Mb	16,525.6Mb	115.7M
HorizonDiscoveryHD789rep1	780.7	2,737.8Mb	15.3%	1418.0	17,877.7Mb	17,839.3Mb	130.7M
Seracare07101307rep1	655.0	2,713.7Mb	21.7%	1417.0	12,514.8Mb	12,489.9Mb	85.1M
HorizonDiscoveryHD799rep2	725.8	1,641.0Mb	11.7%	1023.0	16,541.7Mb	16,510.5Mb	115.3M
HorizonDiscoveryHD799rep1	720.4	1,087.9Mb	10.9%	574.0	9,953.1Mb	9,933.7Mb	70.7M
HorizonDiscoveryHD803rep2	833.7	798.1Mb	12.8%	427.0	6,220.9Mb	6,207.9Mb	46.6M
HorizonDiscoveryHD803rep1	826.2	639.2Mb	16.6%	343.0	6,540.8Mb	6,027.5Mb	45.7M

twistcgp collects per-sample hybrid-selection quality metrics:

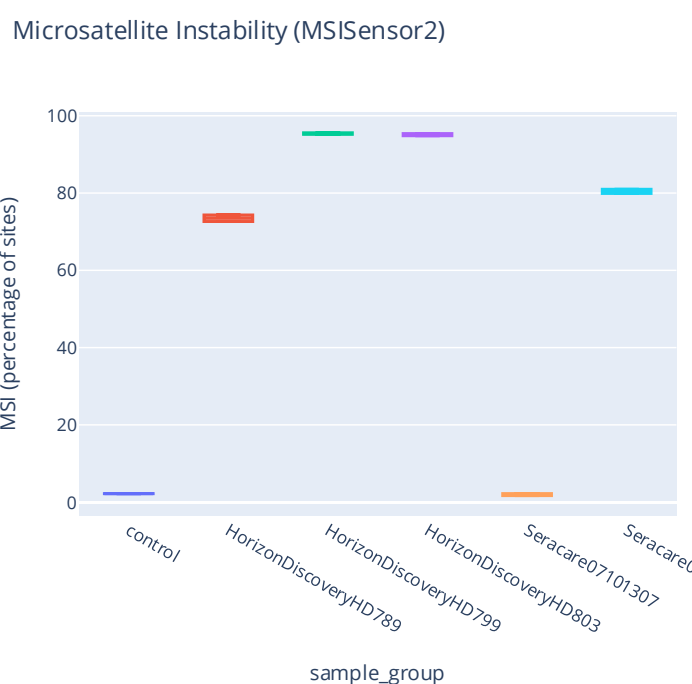
- coverage depth and uniformity
- library complexity estimates
- base-level efficiency
- sources of base loss

Tumor Mutational Burden (TMB), Microsatellite Instability (MSI), and Copy Number Variant (CNV) Analysis



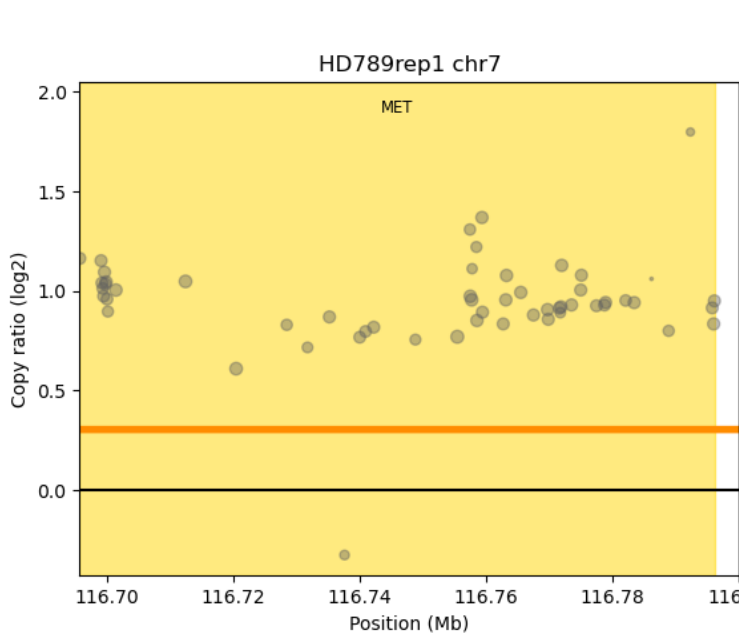
twistcgp calculates TMB and MSI:

- germline variants can be excluded using a user-supplied gnomAD VCF.
- configurable thresholds for variant allele frequency and population allele frequency
- tumor-only MSI calculation allows users to supply a pre-computed scan file



NA12878 and NA24385 are TMB-low (0 and 1 mut/Mb), as expected for germline references

NA12878 and NA24385 are MSI-stable; HD samples are expected to be MSI-high



twistcgp calls somatic copy number variants:

- log2 copy ratios estimated from read depth
- Panel of Normals can be optionally provided to increase accuracy

Sample HD789 should contain ~4.5 copies of MET and ~8.5 copies of MYC-N

Highlights

Designed for Real-World Constraints

- Tumor-only mode — no matched normal required
- Resource files (germline VCF, panel of normals, gnomAD) can be omitted — the pipeline runs with what you have

Comprehensive Biomarkers in a Single Run

- SNVs/indels, copy number variants, TMB, and MSI from one workflow
- Variant calls are soft-filtered, preserving the full call set
- CIViC annotation links variants to curated clinical evidence

Flexible & Modular

- CNV, MSI, TMB, and CIViC annotation can each be skipped independently
- Supports both paired-end and single-end reads
- Configurable thresholds for TMB filtering and target padding

Reproducible & Auditable

- Pinned tool versions across all environments
- Built-in Nextflow execution reports and software manifests on every run

Under Active Development

- Future potential improvements include:
 - UMI-aware analysis
 - Structural Variant (SV) calling
 - Homologous Recombination Deficiency (HRD)
 - Per-sample IGV reporting

Acknowledgements

We thank Tim Fennell and Tim Dunn of Fulcrum Genomics for their extensive assistance in the development of this pipeline, and Michelle Vierra and Juliann Chmielecki for abstract review and technical expertise.

We thank James Flynn, Rebecca Barnard, Tina Han, and Sarthak Sharma of Twist Bioscience for pipeline iteration and development.

We acknowledge the Nextflow nf-core community for their continued maintenance of the open-source tools used in this work and their commitment to advancing open software development in the bioinformatics community.