Calling **germline** and **somatic mutations** from whole genome sequencing (WGS) data in glioblastoma (GBM) patients requires a matched tumor–normal analysis pipeline. Below is a step-by-step workflow using best-practice tools and methods:

**🧬 1. Data Preparation**

**Input:**

* **Tumor WGS** FASTQ (e.g., GBM tumor tissue)
* **Matched normal WGS** FASTQ (e.g., blood or adjacent non-tumor brain tissue)

**Preprocessing:**

* Quality control (e.g., with FastQC)
* Adapter trimming (e.g., Trimmomatic or fastp)
* Alignment with **BWA-MEM** to **hg38** (or hg19) reference genome
* Post-alignment processing:
  + Sort BAM (samtools sort)
  + Mark duplicates (GATK MarkDuplicates)
  + Base quality score recalibration (GATK BaseRecalibrator)

**🔍 2. Germline Mutation Calling**

You only need the **normal sample** for germline calls.

**Tool:**

* **GATK HaplotypeCaller**
* gatk HaplotypeCaller -R reference.fasta -I normal.bam -O germline.vcf

**Output:**

* germline.vcf: Contains SNPs and indels present in the germline.

Optional:

* Filter by population frequency using gnomAD or 1000 Genomes (to annotate rare vs common variants).

**🔬 3. Somatic Mutation Calling**

You need **tumor-normal paired BAMs**.

**Recommended Tools:**

**A. GATK Mutect2**

gatk Mutect2 -R reference.fasta \

-I tumor.bam -tumor TUMOR\_SAMPLE \

-I normal.bam -normal NORMAL\_SAMPLE \

--germline-resource af-only-gnomad.vcf.gz \

--panel-of-normals pon.vcf.gz \

-O somatic.vcf

**B. Other Tools (for consensus calling):**

* **Strelka2**
* **VarScan2**
* **Lancet**
* **MuSE**
* **SomaticSniper**

You can use multiple callers and intersect/filter them using tools like **bcbio**, **SomaticSeq**, or **GATK Funcotator**.

**🧪 4. Post-processing & Filtering**

* **Filtering:** gatk FilterMutectCalls or caller-specific filters
* **Annotation:** Use ANNOVAR, VEP, or Oncotator to annotate variant effects (e.g., coding, splice site, etc.)
* **Filtering Criteria:**
  + Variant allele frequency (VAF)
  + Read depth and base quality
  + Known artifacts (e.g., strand bias)
  + Population databases (gnomAD, dbSNP)
  + Cancer databases (COSMIC)

**🧠 5. Special Considerations for GBM**

* Focus on known GBM driver genes: **TP53, PTEN, EGFR, IDH1, ATRX, TERT promoter**
* Consider calling **TERT promoter mutations** with custom scripts due to GC-rich region
* For **RNA-seq validation**, check allele-specific expression of somatic variants

**🧬 Optional: Structural Variants, CNVs, and MSI**

* **CNVs**: Use GATK CNV, CNVkit, or FACETS
* **SVs**: Use Manta, LUMPY, Delly
* **Microsatellite instability (MSI)**: MSIsensor or MANTIS

**🧰 6. Visualization & Interpretation**

* **IGV**: Inspect individual variants manually
* **cBioPortal format**: For large-cohort sharing
* **Pathway-level impact**: Map variants to GBM-related pathways (e.g., PI3K-Akt, p53, RTK/RAS)

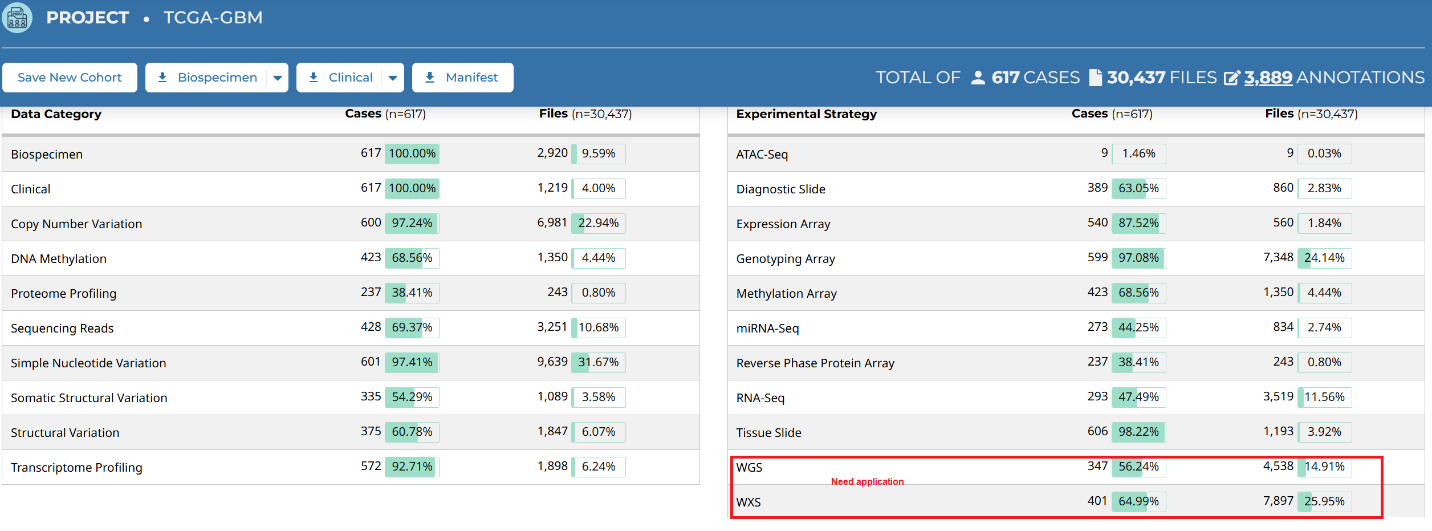
**Summary Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variant Type** | **Tool** | **Input** | **Output** |
| Germline | GATK HaplotypeCaller | Normal BAM | germline.vcf |
| Somatic SNVs/Indels | GATK Mutect2 / Strelka2 | Tumor & Normal BAMs | somatic.vcf |
| Annotation | VEP / ANNOVAR / Oncotator | VCF | Annotated VCF |
| CNVs | CNVkit / GATK CNV | Tumor & Normal BAMs | CNV files |
| SVs | Manta / Delly | Tumor & Normal BAMs | SV calls |

**Data availability**

DNA sequence data: WGS or WXS

1. [**https://portal.gdc.cancer.gov/projects/TCGA-GBM**](https://portal.gdc.cancer.gov/projects/TCGA-GBM)**, need apply for the data, which is controlled.**

****

1. Nature paper: <https://www.nature.com/articles/s41588-025-02167-5.pdf>

Processed snRNA-seq data generated for this study are available at the Gene Expression Omnibus under accession no. [GSE274546](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE274546) (10x Genomics) and [GSE274548](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE274548) (Smart-seq2). All de-identified somatic mutation and copy number alteration calls, and genomic analysis tables, are available via Synapse ([https://www.synapse.org/care\_glioblastoma](https://urldefense.com/v3/__https:/www.synapse.org/care_glioblastoma__;!!KVu0SnhVq1hAFvslES2Y!Jj11HYsvR3DXWkKRTTAfMOWYjmgNIZoDpaIuv9srphfrjcK7z0zgxw3DZKeqoiM38istqfBEtsxwHgQ3Sl5MQR_NnV4Ye8r4Y8I$)). Raw sequencing data are available with limitations in accordance with the consent forms from the **Data Use Oversight System (DUOS) at**[**https://duos.boardinstitute.org**](https://duos.boardinstitute.org/)**under the following IDs: DUOS-000475; DUOS-000476; DUOS-000477; DUOS-000478; DUOS-000479; and DUOS-000480. need apply for the data, which is controlled.**

1. source Chinese CGGA [https://www.cgga.org.cn/](https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.cgga.org.cn%2F&data=05%7C02%7CLijun.Cheng%40osumc.edu%7C43bb087387174c6d93e308dddab4b65b%7C0b95a125791c4f0a9f9e99e363117506%7C0%7C0%7C638907191884296997%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=11xeZO4KCVS%2BvQm%2Fa%2BAOeks994%2BOSNh4BFOcEKUHBqA%3D&reserved=0) , **need apply for the data, which is controlled.**

TCGA GBM whole genome somatic mutation calling and data analysis results see excel file.

TCGA\_GBM\_mutation.xls