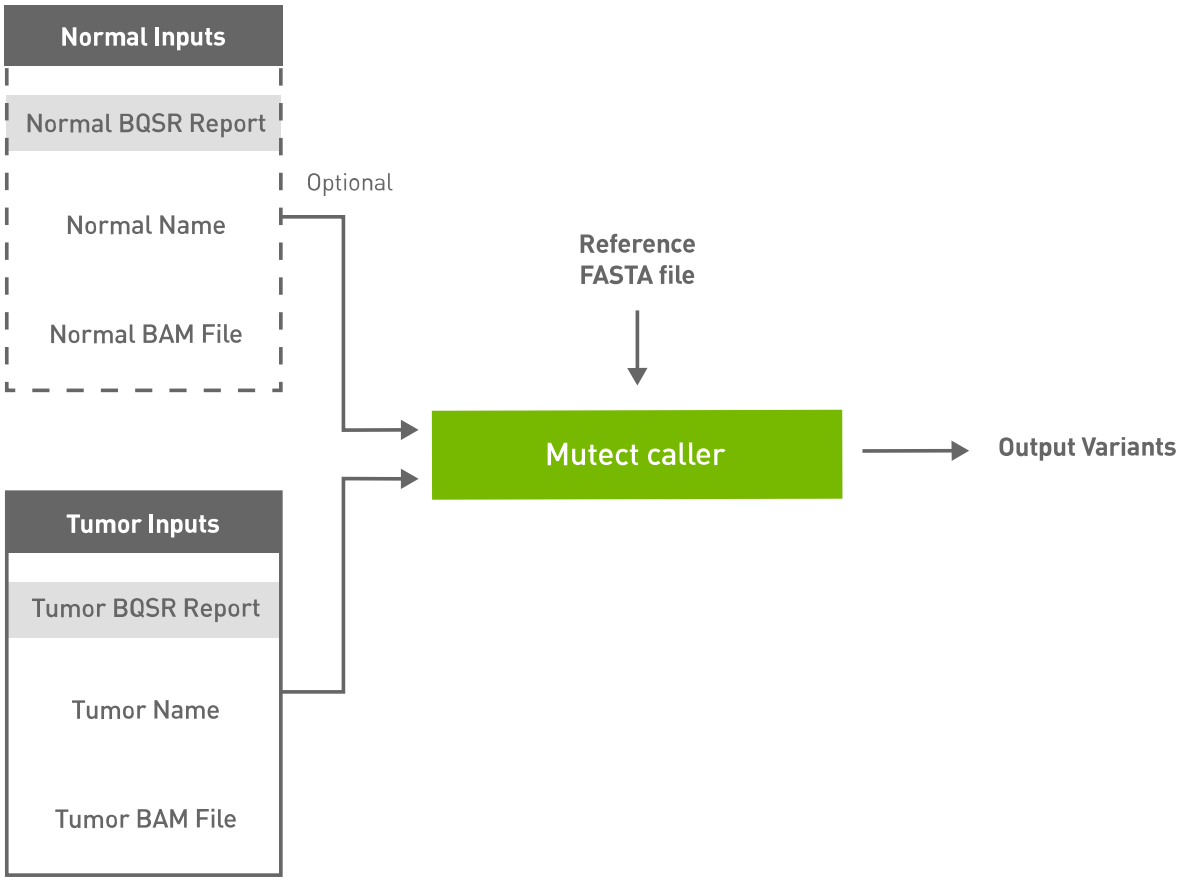
mutectcaller

Table of contents

This tool is an accelerated version of the GATK somatic variant caller, Mutect2, which takes aligned BAMs from the FQ2BAM tool, and outputs a VCF file. This can take as input either a single ("tumor-only") BAM, or a pair of BAMs ("tumor-normal") to provide a baseline to call somatic variants against.

The figure below shows the high-level functionality of mutectcaller. All dotted boxes indicate optional data, with some constraints.



The names of the tumor sample (for the `--tumor-name` option) and the normal sample (for the `--normal-name` option) can be extracted from the headers of their respective BAM files with samtools, which can be installed through apt-get:

```
$ sudo apt-get install samtools
```

Or you can build it from source codes by following the instructions in samtools repo.

Once you have samtools installed on your system you can run this command to get the sample name (SM) field:

```
$ samtools view NA12878.bam -H | grep '@RG' @RG ID:HJYFJ.4 SM:NA12878 LB:Pond-492093 PL:illumina PU:HJYFJCCXX160204.4.GCCGCAAC CN:BI DT:2016-02-04T00:00:00-0500
```

The sample name is the value after "*SM:*" (NA12878, in this example)

If there are multiple read group (@RG) lines and all of them have the same sample name you may safely use the common sample name. If there are multiple read group lines with multiple sample names, choose one sample name as the input. All reads with that sample name will be processed by `mutectcaller` and all other reads will be ignored. Currently only one sample name per BAM file is supported.

If there are no read group lines in the BAM header, or there is no sample name in the read group line, you will need to add read group information to the BAM file. This may be done by running this command:

```
$ samtools addreplacerg \ -r "@RG\tID:sample_rg1\tLB:lib1\tPL:bar\tSM:sample_sm\tPU:sample_rg1" \ original_file.bam \ -o updated_file.bam \ -O BAM
```

This will update the sample name of all reads in this BAM file to "sample_sm", and you can pass "sample_sm" as the sample name of this BAM file. Make sure you use the *updated_file.bam* as input to `mutectcaller`.

Quick Start

You can download the mutect sample dataset from here. Extract all files by running:

```
$ tar -xvzf mutect_sample.tar.gz mutect_sample/
mutect_sample/germline_resource.vcf.gz.tbi mutect_sample/force_call.vcf.gz.tbi
mutect_sample/germline_resource.vcf.gz mutect_sample/tumor.bam.bai
mutect_sample/GCA_000001405.15_GRCh38_no_alt_analysis_set.fa
mutect_sample/force_call.vcf.gz mutect_sample/tumor.bam
mutect_sample/normal.bam.bai mutect_sample/normal.bam
```

Inside the `mutect_sample` folder you will find the necessary input files including:

- one reference fasta (GCA_000001405.15_GRCh38_no_alt_analysis_set.fa),
- one tumor bam (tumor.bam),
- one normal bam (normal.bam),
- one force_calling.vcf.gz VCF file and
- one germline_resource.vcf.gz VCF file

with all necessary indexes.

```
# This command assumes all the inputs are in INPUT_DIR and all the outputs go to
OUTPUT_DIR. docker run --rm --gpus all --volume INPUT_DIR:/workdir --volume
OUTPUT_DIR:/outputdir \ --workdir /workdir \ nvcr.io/nvidia/clara/clara-
parabricks:4.3.1-1 \ pbrun mutectcaller \ --ref /workdir/${REFERENCE_FILE} \ --
tumor-name tumor_name_inside_bam_file \ --in-tumor-bam
/workdir/${INPUT_TUMOR_BAM} \ --in-normal-bam
/workdir/${INPUT_NORMAL_BAM} \ --normal-name normal_name_inside_bam_file \
--out-vcf /outputdir/${OUTPUT_VCF}
```

Compatible GATK4 Command

The command below is the GATK4 counterpart of the Parabricks command above. The output from this command will be identical to the output from the above command. See the Output Comparison page for comparing the results.

```
$ gatk Mutect2 \ -R <INPUT_DIR>/${REFERENCE_FILE} \ --input <INPUT_DIR>/${INPUT_TUMOR_BAM} \ --tumor-sample tumor_name_inside_bam_file \ --input <INPUT_DIR>/${INPUT_NORMAL_BAM} \ --normal-sample normal_name_inside_bam_file \ --output <OUTPUT_DIR>/${OUTPUT_VCF}
```

Mutect2 with Panel of Normals

Parabricks Mutect2 from version 3.7.0-1 has started supporting Panel of Normals to process variants. There are three steps involved:

- prepon
- running mutectcaller with the index generated by prepon
- postpon, updating the vcf with pon information

```
# The first command will generate input.pon that should be done once for the input.vcf.gz # This command assumes all the inputs are in INPUT_DIR and all the outputs go to OUTPUT_DIR. docker run --rm --gpus all --volume INPUT_DIR:/workdir --volume OUTPUT_DIR:/outputdir \ --workdir /workdir \ nvcr.io/nvidia/clara/clara-parabricks:4.3.1-1 \ pbrun prepon --in-pon-file /workdir/${INPUT_PON_VCF} # Run mutectcaller with the pon index # This command assumes all the inputs are in INPUT_DIR and all the outputs go to OUTPUT_DIR. docker run --rm --gpus all --volume INPUT_DIR:/workdir --volume OUTPUT_DIR:/outputdir \ --workdir /workdir \ nvcr.io/nvidia/clara/clara-parabricks:4.3.1-1 \ pbrun mutectcaller \ --ref /workdir/${REFERENCE_FILE} \ --tumor-name tumor \ --in-tumor-bam /workdir/${INPUT_TUMOR_BAM} \ --in-normal-bam /workdir/${INPUT_NORMAL_BAM} \ --pon /workdir/${INPUT_PON_VCF} \ --normal-name normal \ --out-vcf /outputdir/${OUTPUT_VCF} # Add the annotation to the output.vcf generated above # This command assumes all the inputs are in INPUT_DIR and all the outputs go to OUTPUT_DIR. docker run --rm --gpus all --volume INPUT_DIR:/workdir --volume OUTPUT_DIR:/outputdir \ --workdir /workdir \ nvcr.io/nvidia/clara/clara-parabricks:4.3.1-1 \ pbrun postpon \ --in-vcf
```

```
/workdir/${OUTPUT_VCF} \ --in-pon-file /workdir/${INPUT_PON_FILE} \ --out-vcf /outputdir/${OUTPUT_ANNOTATED_VCF}
```

mutectcaller Reference

Run GPU mutect2 to convert BAM/CRAM to vcf

Input/Output file options

--ref REF

Path to the reference file. (default: None)

Option is required.

--out-vcf OUT_VCF

Path of the VCF file after Variant Calling. (default: None)

Option is required.

--in-tumor-bam IN_TUMOR_BAM

Path of the BAM/CRAM file for tumor reads. (default: None)

Option is required.

--in-normal-bam IN_NORMAL_BAM

Path of the BAM/CRAM file for normal reads. (default: None)

--in-tumor-recal-file IN_TUMOR_RECAL_FILE

Path of the report file after Base Quality Score Recalibration for tumor sample. (default: None)

--in-normal-recal-file IN_NORMAL_RECAL_FILE

Path of the report file after Base Quality Score Recalibration for normal sample. (default: None)

--interval-file INTERVAL_FILE

Path to an interval file in one of these formats: Picard-style (.interval_list or .picard), GATK-style (.list or .intervals), or BED file (.bed). This option can be used multiple times. (default: None)

--mutect-bam-output MUTECT_BAM_OUTPUT

File to which assembled haplotypes should be written. (default: None)

--pon PON

Path of the vcf.gz PON file. Make sure you run prepon first and there is a '.pon' file already. (default: None)

--mutect-germline-resource MUTECT_GERMLINE_RESOURCE

Path of the vcf.gz germline resource file. Population vcf of germline sequencing containing allele fractions. (default: None)

--mutect-alleles MUTECT_ALLELES

Path of the vcf.gz force-call file. The set of alleles to force-call regardless of evidence. (default: None)

Tool Options:

--max-mnp-distance MAX_MNP_DISTANCE

Two or more phased substitutions separated by this distance or less are merged into MNPs. (default: 1)

--mutectcaller-options MUTECTCALLER_OPTIONS

Pass supported mutectcaller options as one string. The following are currently supported original mutectcaller options: -pcr-indel-model <NONE, HOSTILE, AGGRESSIVE,

CONSERVATIVE>, -max-reads-per-alignment-start <int>, (e.g. --mutectcaller-options="-pcr-indel-model HOSTILE -max-reads-per-alignment-start 30"). (default: None)
--initial-tumor-lod INITIAL_TUMOR_LOD

Log 10 odds threshold to consider pileup active. (default: None)

--tumor-lod-to-emit TUMOR_LOD_TO_EMIT

Log 10 odds threshold to emit variant to VCF. (default: None)

--pruning-lod-threshold PRUNING_LOD_THRESHOLD

Ln likelihood ratio threshold for adaptive pruning algorithm. (default: None)

--active-probability-threshold ACTIVE_PROBABILITY_THRESHOLD

Minimum probability for a locus to be considered active. (default: None)

--no-alt-contigs

Ignore commonly known alternate contigs. (default: None)

--genotype-germline-sites

Call all apparent germline site even though they will ultimately be filtered. (default: None)

--genotype-pon-sites

Call sites in the PoN even though they will ultimately be filtered. (default: None)

--force-call-filtered-alleles

Force-call filtered alleles included in the resource specified by --alleles. (default: None)

--tumor-name TUMOR_NAME

Name of the sample for tumor reads. This MUST match the SM tag in the tumor BAM file. (default: None)

Option is required.

--normal-name NORMAL_NAME

Name of the sample for normal reads. If specified, this MUST match the SM tag in the normal BAM file. (default: None)

-L INTERVAL, --interval INTERVAL

Interval within which to call the variants from the BAM/CRAM file. All intervals will have a padding of 100 to get read records, and overlapping intervals will be combined. Interval files should be passed using the --interval-file option. This option can be used multiple times (e.g. "-L chr1 -L chr2:10000 -L chr3:20000+ -L chr4:10000-20000"). (default: None)

-ip INTERVAL_PADDING, --interval-padding INTERVAL_PADDING

Amount of padding (in base pairs) to add to each interval you are including. (default: None)

Performance Options:

--mutect-low-memory

Use low memory mode in mutect caller. (default: None)

--run-partition

Turn on partition mode; divides genome into multiple partitions and runs 1 process per partition. (default: None)

--gpu-num-per-partition GPU_NUM_PER_PARTITION

Number of GPUs to use per partition. (default: None)

--num-htvc-threads NUM_HTVC_THREADS

Number of CPU threads to use. (default: 5)

Common options:

--logfile LOGFILE

Path to the log file. If not specified, messages will only be written to the standard error output. (default: None)

--tmp-dir TMP_DIR

Full path to the directory where temporary files will be stored.

--with-petagene-dir WITH_PETAGENE_DIR

Full path to the PetaGene installation directory. By default, this should have been installed at /opt/petagene. Use of this option also requires that the PetaLink library has been preloaded by setting the LD_PRELOAD environment variable. Optionally set the PETASUITE_REFPATH and PGCLOUD_CREDPATH environment variables that are used for data and credentials (default: None)

--keep-tmp

Do not delete the directory storing temporary files after completion.

--no-seccomp-override

Do not override seccomp options for docker (default: None).

--version

View compatible software versions.

GPU options:

--num-gpus NUM_GPUS

Number of GPUs to use for a run. GPUs 0..(NUM_GPUS-1) will be used.