# deepvariant

# Table of contents

Run a GPU-accelerated DeepVariant algorithm.

## What is DeepVariant?

DeepVariant is a deep learning based variant caller developed by Google for germline variant calling of high-throughput sequencing data. It works by taking aligned sequencing reads in BAM/CRAM format and utilizes a convolutional neural network (CNN) to classify the locus into true underlying genomic variation or sequencing error. DeepVariant can therefore call single nucleotide variants (SNVs) and insertions/deletions (InDels) from sequencing data at high accuracy in germline samples.

## Why DeepVariant?

DeepVariant's approach is able to detect variants that are often missed by traditional (for example Bayesian) variant callers, and is known to reduce false positives. It offers several advantages over similar tools, including its ability to detect a wide range of variants with high accuracy, its scalability for analyzing large datasets, and its open source availability. Additionally, its deep learning-based approach allows it to provide better support for different sequencing platforms, as it can be retrained to provide higher accuracy for specific protocols or research areas.

## How should I use DeepVariant?

DeepVariant is designed for use as a germline variant caller that can apply different models trained for specific sample types (such as whole genome and whole exome samples) to yield higher accuracy results. DeepVariant can be deployed within NVIDIA's Parabricks software suite, which is designed for accelerated secondary analysis in genomics, bringing industry standard tools and workflows from CPU to GPU, and delivering the same results at up to 60x faster runtimes. A 30x whole genome can be run through DeepVariant in as little as 8 minutes on an NVIDIA DGX station, compared to 5 hours on a CPU instance (m5.24xlarge, 96 x vCPU). DeepVariant in Parabricks is used in the same way as other command line tools that users are familiar with: It takes a BAM/CRAM and the reference genome as inputs and produces the variants (a VCF file) as outputs. **Currently, DeepVariant is supported for V100 and newer GPUs out of the box.**

**Note**

In version 3.8 the *--run-partition* option was added, which can lead to a significant speed increase. However, using the *--run-partition*, *--proposed-variants*, and *--gvcf* options at the same time will lead to a substantial slowdown. A warning will be issued and the *--run-partition* option will be ignored.

## Available Operating Modes

Parabricks DeepVariant can run in one of three operating modes:

1. shortread
2. PacBio
3. ONT

See the **--mode** option below.

## Quick Start

```
# 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 deepvariant \ --ref /workdir/${REFERENCE_FILE} \ --in-bam /workdir/${INPUT_BAM} \ --out-variants /outputdir/${OUTPUT_VCF}
```

## Compatible Google DeepVariant Commands

The commands below are the Google counterpart of the Parabricks command above. The output from these commands will be identical to the output from the above command. See the Output Comparison page for comparing the results.

```
sudo docker run \ --volume <INPUT_DIR>:/input \ --volume <OUTPUT_DIR>:/output \ google/deepvariant:1.6.1 \ /opt/deepvariant/bin/run_deepvariant \ --model_type WGS \ --ref /input/${REFERENCE_FILE} \ --reads /input/${INPUT_BAM} \ --output_vcf /output/${OUTPUT_VCF} \ --num_shards $(nproc) \ --make_examples_extra_args "ws_use_window_selector_model=true"
```

# Models for additional GPUs

Parabricks DeepVariant supports the following models:

1. Short-read WGS
2. Short-read WES
3. PacBio
4. ONT

DeepVariant models for T4, V100 and all other GPUs which are Ampere and above architecture ship with the software.

# deepvariant Reference

Run DeepVariant to convert BAM/CRAM to VCF.

### Input/Output file options

--ref REF

Path to the reference file. (default: None)

Option is required.

--in-bam IN_BAM

Path to the input BAM/CRAM file for variant calling. (default: None)

Option is required.
--interval-file INTERVAL_FILE

Path to a BED file (.bed) for selective access. This option can be used multiple times. (default: None)

--out-variants OUT_VARIANTS

Path of the vcf/g.vcf/g.vcf.gz file after variant calling. (default: None)

Option is required.

--pb-model-file PB_MODEL_FILE

Path to a non-default parabricks model file for deepvariant. (default: None)

--pb-model-dir PB_MODEL_DIR

Path to a non-default parabricks model dir that contains multiple engine files for one model (default: None)

--proposed-variants PROPOSED_VARIANTS

Path of the vcf.gz file, which has proposed variants for the make examples stage. (default: None)

## Tool Options:

--disable-use-window-selector-model

Change the window selector model from Allele Count Linear to Variant Reads. This option will increase the accuracy and runtime. (default: None)

--gvcf

Generate variant calls in .gvcf Format. (default: None)

--norealign-reads

Do not locally realign reads before calling variants. Reads longer than 500 bp are never realigned. (default: None)

--sort-by-haplotypes

Reads are sorted by haplotypes (using HP tag). (default: None)

--keep-duplicates

Keep reads that are duplicate. (default: None)

--vsc-min-count-snps VSC_MIN_COUNT_SNPS

SNP alleles occurring at least this many times in the AlleleCount will be advanced as candidates. (default: 2)

--vsc-min-count-indels VSC_MIN_COUNT_INDELS

Indel alleles occurring at least this many times in the AlleleCount will be advanced as candidates. (default: 2)

--vsc-min-fraction-snps VSC_MIN_FRACTION_SNPS

SNP alleles occurring at least this fraction of all counts in the AlleleCount will be advanced as candidates. (default: 0.12)

--vsc-min-fraction-indels VSC_MIN_FRACTION_INDELS

Indel alleles occurring at least this fraction of all counts in the AlleleCount will be advanced as candidates. (default: None)

--min-mapping-quality MIN_MAPPING_QUALITY

By default, reads with any mapping quality are kept. Setting this field to a positive integer i will only keep reads that have a MAPQ >= i. Note this only applies to aligned reads. (default: 5)

--min-base-quality MIN_BASE_QUALITY

Minimum base quality. This option enforces a minimum base quality score for alternate alleles. Alternate alleles will only be considered if all bases in the allele have a quality greater than min_base_quality. (default: 10)

--mode MODE

Value can be one of [shortread, pacbio, ont]. By default, it is shortread. If mode is set to pacbio, the following defaults are used: --norealign-reads, --alt-aligned-pileup diff_channels, --vsc-min-fraction-indels 0.12. If mode is set to ont, the following defaults are used: -norealign-reads, --variant-caller VCF_CANDIDATE_IMPORTER. (default: shortread)

--alt-aligned-pileup ALT_ALIGNED_PILEUP

Value can be one of [none, diff_channels]. Include alignments of reads against each candidate alternate allele in the pileup image. (default: None)

--variant-caller VARIANT_CALLER

Value can be one of [VERY_SENSITIVE_CALLER, VCF_CANDIDATE_IMPORTER]. The caller to use to make examples. If you use VCF_CANDIDATE_IMPORTER, it implies force calling. Default is VERY_SENSITIVE_CALLER. (default: None)

--add-hp-channel

Add another channel to represent HP tags per read. (default: None)

--parse-sam-aux-fields

Auxiliary fields of the BAM/CRAM records are parsed. If either --sort-by-haplotypes or --add-hp-channel is set, then this option must also be set. (default: None)

--use-wes-model

If passed, the WES model file will be used. Only used in shortread mode. (default: None)

--include-med-dp

If True, include MED_DP in the output gVCF records. (default: None)

--normalize-reads

If True, allele counter left align INDELs for each read. (default: None)

--pileup-image-width PILEUP_IMAGE_WIDTH

Pileup image width. Only change this if you know your model supports this width. (default: 221)

--channel-insert-size

If True, add insert_size channel into pileup image. By default, this parameter is true in WGS and WES mode. (default: None)

--no-channel-insert-size

If True, don't add insert_size channel into the pileup image. (default: None)

--max-read-size-512

Allow deepvariant to run on reads of size 512bp. The default size is 320 bp. (default: None)

--prealign-helper-thread

Use an extra thread for the pre-align step. This parameter is more useful when --max-reads-size-512 is set. (default: None)

--track-ref-reads

If True, allele counter keeps track of reads supporting ref. By default, allele counter keeps a simple count of the number of reads supporting ref. (default: None)

--phase-reads

Calculate phases and add HP tag to all reads automatically. (default: None)

--dbg-min-base-quality DBG_MIN_BASE_QUALITY

Minimum base quality in a k-mer sequence to consider. (default: 15)

--ws-min-windows-distance WS_MIN_WINDOWS_DISTANCE

Minimum distance between candidate windows for local assembly (default: 80)

--channel-gc-content

If True, add gc_content channel into pileup image (default: None)

--channel-hmer-deletion-quality

If True, add hmer deletion quality channel into pileup image (default: None)

--channel-hmer-insertion-quality

If True, add hmer insertion quality channel into pileup image (default: None)

--channel-non-hmer-insertion-quality

If True, add non-hmer insertion quality channel into pileup image (default: None)

--skip-bq-channel

If True, ignore base quality channel. (default: None)

--aux-fields-to-keep AUX_FIELDS_TO_KEEP

Comma-delimited list of auxiliary BAM fields to keep. Values can be [HP, tp, t0] (default: HP)

--vsc-min-fraction-hmer-indels VSC_MIN_FRACTION_HMER_INDELS

Hmer Indel alleles occurring at least this be advanced as candidates. Use this threshold if hmer and non-hmer indels should be treated differently (Ultima reads)Default will use the same threshold for hmer and non-hmer indels, as defined in vsc_min_fraction_indels. (default: None)

--vsc-turn-on-non-hmer-ins-proxy-support

Add read-support from soft-clipped reads and other non-hmer insertion alleles,to the most frequent non-hmer insertion allele. (default: None)

--consider-strand-bias

If True, expect SB field in calls and write it to vcf (default: None)

--p-error P_ERROR

Basecalling error for reference confidence model. (default: 0.001)

--channel-ins-size

If true, add another channel to represent size of insertions. (good for flow-based sequencing) (default: None)

--max-ins-size MAX_INS_SIZE

Max insertion size for ins_size_channel, larger insertions will look like max (have max intensity) (default: 10)

--disable-group-variants

If using vcf_candidate_importer and multi-allelic sites are split across multiple lines in VCF, set to True so that variants are not grouped when transforming CallVariantsOutput to Variants. (default: None)

--filter-reads-too-long

Ignore all input bam reads with size > 512bp (default: None)

--haploid-contigs HAPLOID_CONTIGS

Optional list of non autosomal chromosomes. For all listed chromosomes HET probabilities are not considered. (default: None)

-L INTERVAL, --interval INTERVAL

Interval within which to call the variants from the BAM/CRAM file. 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)

## Performance Options:

--num-cpu-threads-per-stream NUM_CPU_THREADS_PER_STREAM

Number of CPU threads to use per stream. (default: 6)

--num-streams-per-gpu NUM_STREAMS_PER_GPU

Number of streams to use per GPU. (default: 2)

--run-partition

Divide the whole genome into multiple partitions and run multiple processes at the same time, each on one partition. (default: None)

--gpu-num-per-partition GPU_NUM_PER_PARTITION

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

--max-reads-per-partition MAX_READS_PER_PARTITION

The maximum number of reads per partition that are considered before following processing such as sampling and realignment. (default: 1500)

--partition-size PARTITION_SIZE

The maximum number of basepairs allowed in a region before splitting it into multiple smaller subregions. (default: 1000)

## 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.