

# VARIANT++ step by step guide

---

## Table of Contents

1. [Prerequisites](#)
2. [Other considerations](#)
3. [Step 1 – QC trimming and merge reads](#)
4. [Step 2 – Deduplicate merged reads](#)
5. [Step 3 – Remove host DNA](#)
6. [Step 4 – Filter reads with Kraken](#)
7. [Step 5 – Classification with Themisto and mSWEEP](#)
8. [Explore the results](#)

Make sure that you have:

1. An updated VARIANT++\_env and github repository \*\* Install it with conda from the VARIANT++ directory with:

```
git clone https://github.com/Microbial-Ecology-Group/VARIANTplusplus.git

cd VARIANTplusplus/

conda env create -f envs/VARIANT++_env.yaml

conda activate VARIANT++_env
# if that doesn't work, use "source" instead of "conda" for the command above.
```

2. The coreNT kraken database. It's about 240 GB and can be downloaded like this:

```
# Download database
wget https://genome-idx.s3.amazonaws.com/kraken/k2_core_nt_20241228.tar.gz

# Make directory for db contents
mkdir -p k2_core_nt_20241228

# unzip it
tar -xzf k2_core_nt_20241228.tar.gz -C k2_core_nt_20241228/
```

## Other considerations

- I recommend using the flag "--profile local\_slurm".
  - This will submit individual processes for each job based on what they typically require.
  - You would need to submit an sbatch script with a header that looks like this:

```
#!/bin/bash
#SBATCH -J AMR++ -o GSV_1_log.out -t 48:00:00 --mem=5G --nodes=1 --ntasks=1 --
cpus-per-task=1

nextflow run main_VARIANT++.nf -profile local_slurm --pipeline GSV_1 -with-report
report_GSV_1_slurm.html --output BRDnoBRD_GSV_result -resume
```

- The various parts of this pipeline can require a lot of temporary storage, so I recommend adding `-w /path/to/work_dir` so that you can place the working directory somewhere other than your working directory.
  - For example, we can move the working directory to the shared space for our group.

## Step 1: QC trimming and merge reads

Parameters that have to change:

- `--pipeline ==> --pipeline GSV_1`
- `--reads ==> --reads "/path/to/your/reads/*R{1,2}.fastq.gz"`

Defaults for Trimmomatic

- `--leading = 3`
- `--trailing = 3`
- `--slidingwindow = "4:15"`
- `--minlen = 36`

Optional

- `--threads = 4`

Example command:

```
nextflow run main_VARIANT++.nf --pipeline GSV_1 --output GSV_analysis --reads
"data/raw/*_R{1,2}.fastq.gz"
```

## Step 2: Deduplicate merged reads

Parameters that have to change:

- `--pipeline ==> --pipeline GSV_2`
- `--merged_reads ==> --merged_reads 'GSV_analysis/Flash_reads/*.{extendedFragments,notCombined}.fastq.gz'`
  - If you named you used `--output GSV_analysis`, then the command below should work with your data, otherwise just change it to match your output directory name.
  - Also note, that this parameter requires the use of single quotes `'`, anything else will not work.

Example command:

```
nextflow run main_VARIANT++.nf --pipeline GSV_2 --output GSV_analysis --merged_reads 'GSV_analysis/Flash_reads/*.{extendedFragments,notCombined}.fastq.gz' -profile local_slurm
```

## Step 3: Remove host DNA

Parameters that have to change:

- `--pipeline ==> --pipeline GSV_3`
- `--merged_reads ==> --merged_reads 'GSV_analysis/Deduped_reads/*_{merged,unmerged}.dedup.fastq.gz'`
- `host ==> --host "/path/to/your/host/chr21.fasta.gz"`
  - remember, you can change this in `params.config` file or add it to your nextflow command.

Example command:

```
nextflow run main_VARIANT++.nf --pipeline GSV_3 --output GSV_analysis --merged_reads 'GSV_analysis/Deduped_reads/*_{merged,unmerged}.dedup.fastq.gz' -profile local_slurm
```

## Step 4: Filter reads with kraken

Parameters that have to change:

- `--pipeline ==> --pipeline GSV_4`
- `--merged_reads ==> --merged_reads 'GSV_analysis/Deduped_reads/*_{merged,unmerged}.dedup.fastq.gz'`
- `--kraken_db ==> --kraken_db /path/to/k2_core_nt_20241228`

Example command:

```
nextflow run main_VARIANT++.nf --pipeline GSV_4 --output GSV_analysis --merged_reads 'GSV_analysis/HostRemoval/NonHostFastq/*_{merged,unmerged}.non.host.fastq.gz' -profile local_slurm
```

## Step 5: Perform classification with themisto and mSweep

Parameters that have to change:

- `--pipeline ==> --pipeline GSV_5`
- `--merged_reads ==> --merged_reads 'GSV_analysis/MicrobiomeAnalysis/Kraken/extracted_reads/*_{merged,unmerged}.dedup.fastq.gz'`

Example command:

```
nextflow run main_VARIANT++.nf --pipeline GSV_5 --output GSV_analysis --merged_reads 'GSV_analysis/MicrobiomeAnalysis/Kraken/extracted_reads/*_Mh_extracted_{merged,unmerged}.fastq.gz' -profile local_slurm
```

## Explore the results

Check the "Results" folder for the kraken analytic matrix and the mSweep results. The "mSweep\_results\_summary.tsv" file contains all results for each the merged and unmerged reads, but the "mSweep\_results\_count\_matrix.tsv" file has a count matrix with the combined results.