hDNApipe Manual

2024/12/17 Version 1.0

This document is designed to assist users in operating hDNApipe variant calling pipeline, with an complete example and finer details provided.
Should any issues or bugs arise during the use, feel free to submit detailed descriptions at the GitHub page https://github.com/TJ-lab-ustc/hDNApipe/issues.

1. Set up example

1.1 Download hDNApipe

```
git clone https://github.com/TJ-lab-ustc/hDNApipe
cd hDNApipe
chmod +x ./dnapipe
```

1.2 Download annotation files

```
bash download_annoatation.sh /home/data3/ZYX/download
```

2. Preparation example

2.1 Enter the docker

```
docker run  \
        -v "/home/data3/ZYX/hDNApipe":"/hDNApipe" \
        -v "/home/data3/ZYX/hg002_wes_bam2fastq":"/input" \
        -v "/home/data3/ZYX/download":"/annotation_dir"
        --net=host -e DISPLAY=:10.0 \
        -v /$HOME/.Xauthority:/root/.Xauthority \
        -it hDNApipe /bin/bash
```

2.2 Edit config (first-time only)

```
cd /hDNApipe
vim ./dnapipe.config
```

Carry out the completion of the `dnapipe_dir` and `annotation_dir` items.

```
dnapipe_dir="/hDNApipe"
annotation_dir="/annotation_dir"
```

2.3 Sample information table preparation

The sample information table is a textual document in which users are obligated to supply pertinent details regarding the samples.

The requisite components encompass the sample name and the location path of the sequencing files.

Moreover, supplementary details such as the sample condition (tumor or control) and the sample sex can optionally be incorporated.

Each of these elements is demarcated by a comma, thereby generating a format exemplified by "sample name, location path of sequencing file 1, location path of sequencing file 2, sample condition, sample sex". An illustration of the full-fledged

format is presented as follows:

```
sample1,/path/sample1_1.fastq.gz,/path/sample1_2.fastq.gz,tumor,male
sample2,/path/sample2_1.fastq.gz,/path/sample2_2.fastq.gz,control,male
```

When the sequencing data is in BAM file format, only one path of the sequencing file needs to be provided. The omitted part in the middle must retain the "," symbol for the program to disassemble according to the format (the symbol can be deleted if there is no content behind). The text format is as follows:

```
sample1,/path/sample1.bam,,tumor,male
sample2,/path/sample2.bam,,control,male
```

The sample status is divided into two types: "tumor" and "control". It must be provided when somatic mutation analysis is selected, while it can be ignored for germline mutation detection.

The gender is divided into "male" and "female", which affects the detection of copy number mutations. It can be omitted.

To sum up, the most concise form of the sample information table is as follows:

```
sample1,/path/sample1.bam
sample2,/path/sample2.bam
```

In our following running example, what is used is:

```
hg002,/input/read_1.fq.gz,/input/read_2.fq.gz
```

3. Command-line interface usage example

3.1 Reference setting (first-time only)

* If the recommended genome is used and the built-in files within Docker have not been modified, this step can be skipped.

```
./dnapipe ref \
        --set /opt/download/reference/GRCh38_no_alt_analysis_set.fna
```

If no relevant index files are detected within its directory, an indexing process will be initiated.

The ending prompt message is as follows:

```
**********Reference setting Done**********
```

3.2 Variant calling pipeline example

```
./dnapipe var \
        --mode germline --file-type fastq --seq-method wes \
        --region /input/idt_capture_novogene_no_alt.bed \
        -i /input/sample.txt \
        --variant short --short deepvariant \
        -o /hg002_deepvar/ \
        -t 20
```

`--mode germline`: set the detection mode as germline

`--file-type`: declare the sequencing data files are in FASTQ format

`--seq-method wes`: declare the sequencing strategy adopted is WES (whole exome sequencing)

`--region`: specify the target region

`-i /input/sample.txt`: designate the sample information table

`-o /hg002_deepvar`: specify the output directory for storage

`--variant short` and `--short deepvariant`: define the type of variants to focus on (short, represents SNV and INDEL), and the specific caller (DeepVariant)

`-t 20`: specify that the number of threads to be used during the operation is 20

The operation details will be presented:

```
root@tjlab:/hDNApipe# ./dnapipe var --mode germline --file-type fastq --seq-method wes --region /input/idt_capture_novogene
_no_alt.bed -i /input/sample.txt --variant short --short deepvariant -o /hg002_deepvar/ -t 20
============ hg002 align and sort ============
[M::bwa_idx_load_from_disk] read 0 ALT contigs
[M::process] read 1324504 sequences (200000104 bp)...
[M::process] read 1324504 sequences (200000104 bp)...
[M::mem_pestat] # candidate unique pairs for (FF, FR, RF, RR): (4, 550791, 9, 3)
[M::mem_pestat] skip orientation FF as there are not enough pairs
```

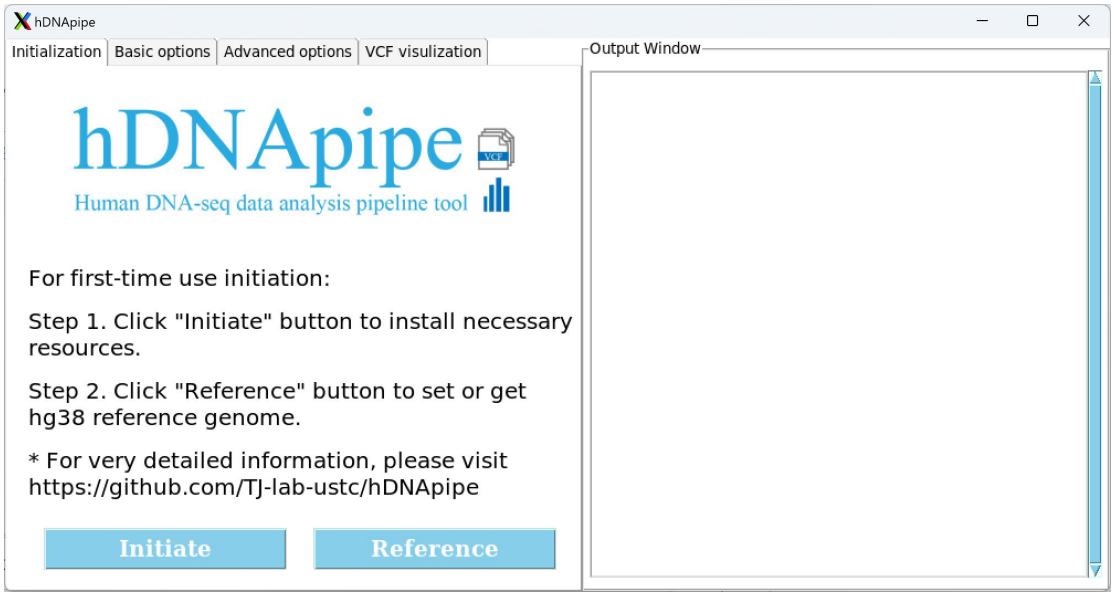
Upon completion of the operation, all files will be stored in the designated output folder `/hg002_deepvar`.

4. Graphical user interface usage steps and example

4.1 Starts up

```
cd /hDNApipe
python dnapipe.py
```

Upon entering the command, a GUI pop-up window appears:

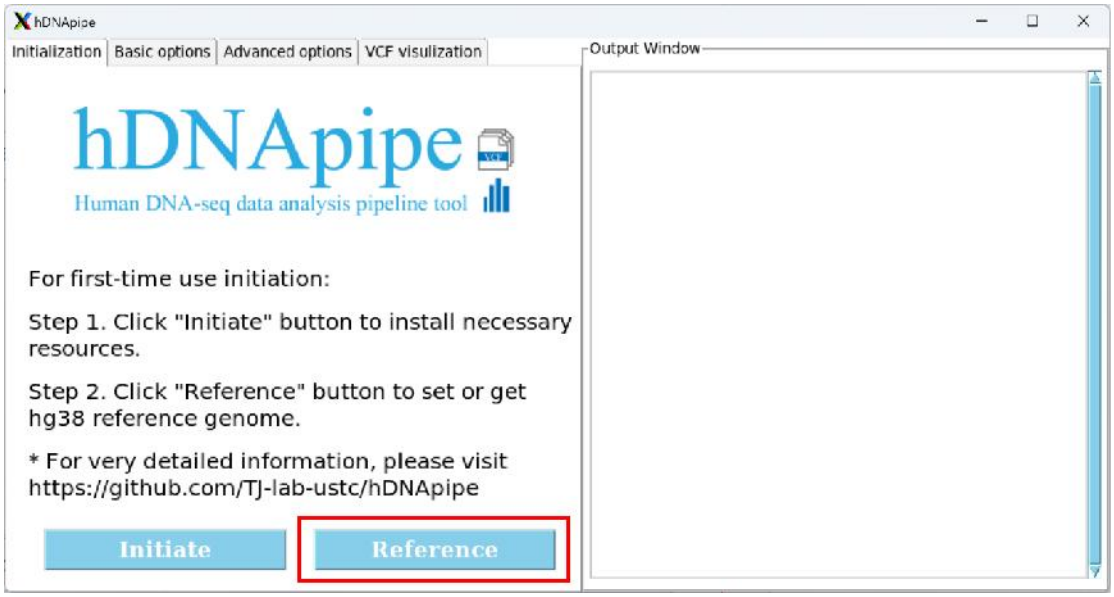


* The operations on the interface are for non-Docker users. Docker users just need to follow the manual instructions.

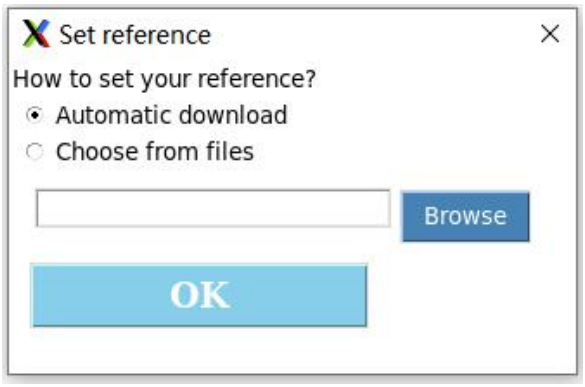
4.2 Reference setting (first-time only)

* If the recommended genome is used and the built-in files within Docker have not been modified, this step can be skipped.

Click 'Reference' button:

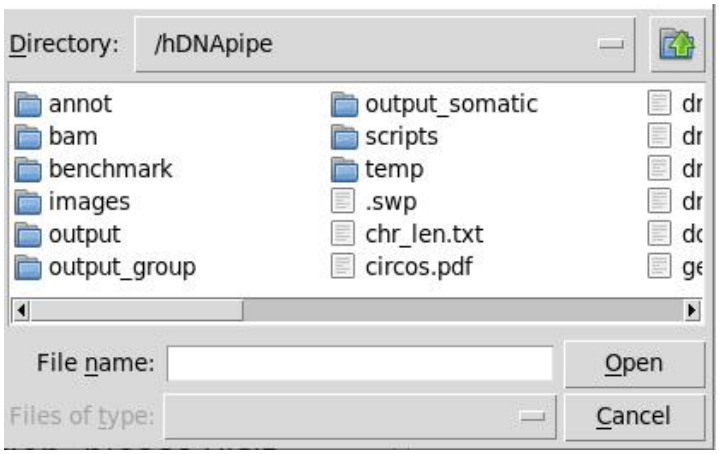


A pop-up window materializes：

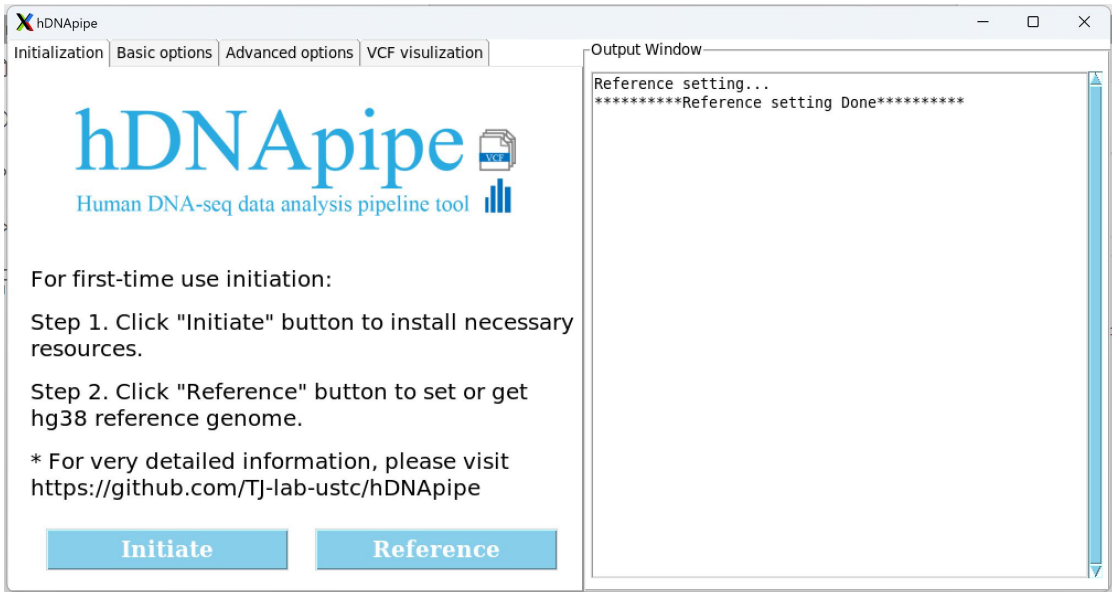


If the 'Automatic download' option is selected, the files will be downloaded to the current folder automatically.

Or select "Choose from files" to pick from existing files. Clicking the "Browse" tab allows for directly browsing the files.



Click 'OK' button and then the running information can be seen.



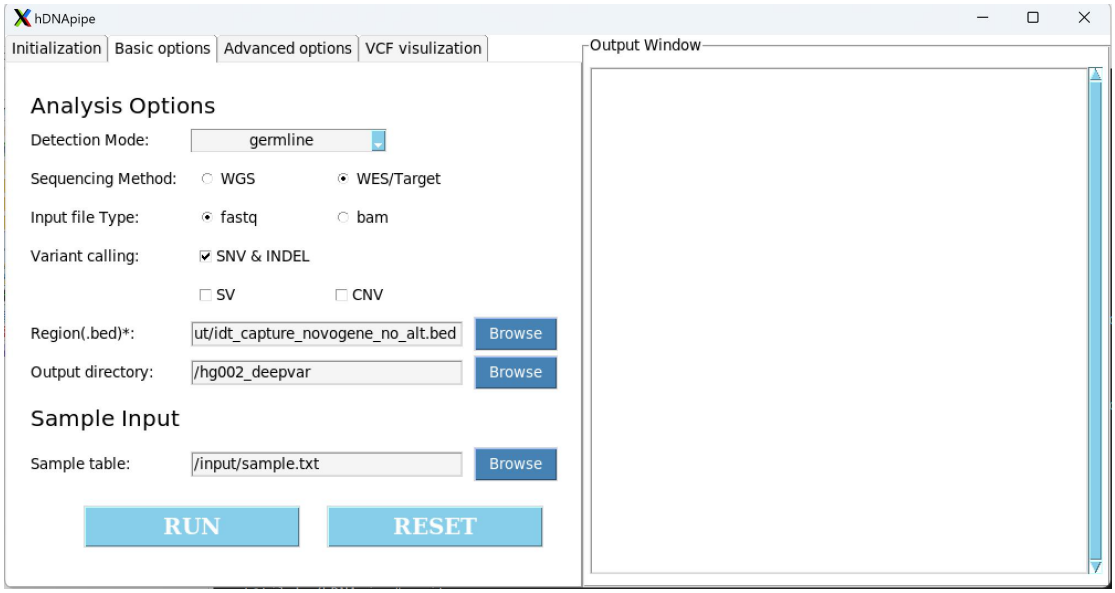
4.3 Variant calling pipeline

In the GUI, to provide more convenient usage for users with different needs, the parameters related to this function are distributed in two tabs.

The "Basic options" contains common parameters that users must select or provide.

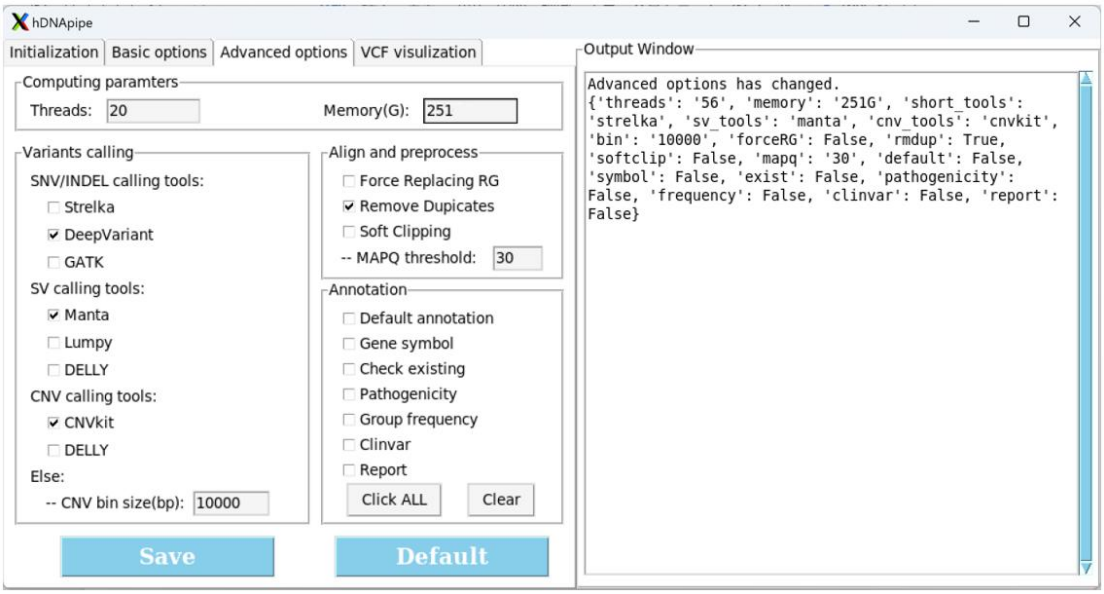
It includes: analysis mode (two types: germline detection and somatic mutation detection), sequencing type which is either whole genome sequencing (WGS) or exome/targeted sequencing (WES/Target), the type of sequencing data provided by the user (FASTQ file or BAM file), the types of mutations to be detected, the region BED file, the output folder, and the sample information table.

To conduct a germline short variant detection analysis on the HG002 WES data (consistent with the command-line analysis content mentioned previously), we have selected the following options:

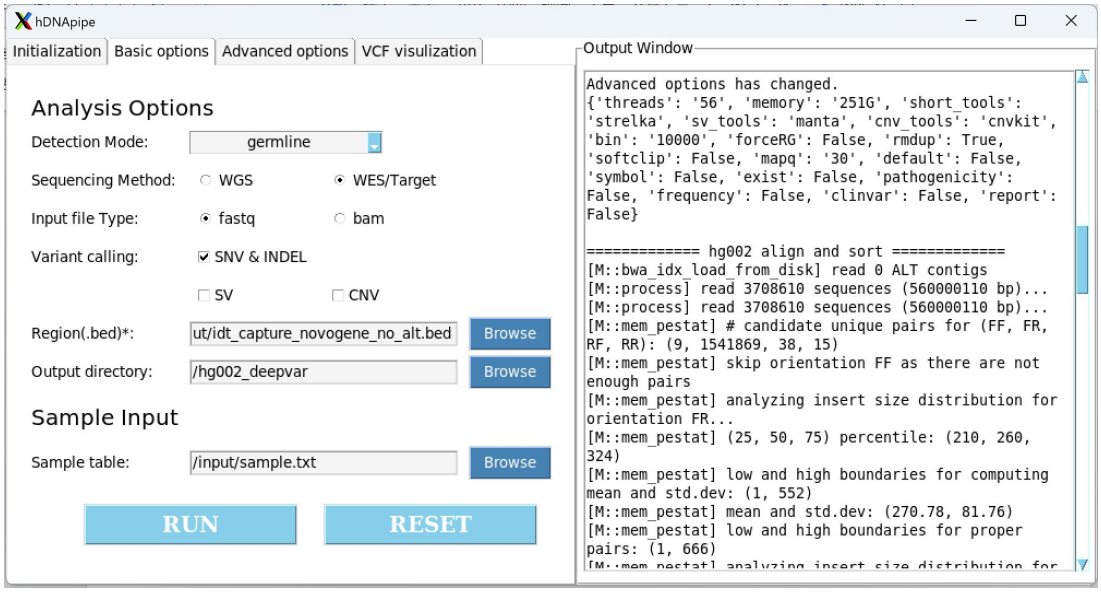


To ensure full consistency with the previous analysis, we entered the "Advanced options" page and modified the number of threads to 20 and the short variant caller to DeepVariant. The other settings remained at their default values.

Don't forget to click the `**Save**` button to save the adjustments.
If the saving is successful, the information in the output window as shown in the following figure will be displayed.



Click 'RUN' button in "Basic options" tab and the running information can be seen.



5. Advanced options introduction

"Advanced options" provides more detailed parameters for advanced users. It is divided into several sections.

Firstly, there is the "Computing parameters" section, where the number of threads and memory used during program execution can be set. By default, resources are automatically detected and set to half of the total available resources.For example, with 112 threads and 502G of memory, it is automatically set to use 56 threads and 251G of memory. Users can make modifications if needed.

The "Variant calling" section pertains to mutation detection. It can be used to adjust the mutation detection tools employed in the process. This section is further divided into three parts.
For SNV and INDEL, Strelka, DeepVariant, and GATK can be selected.
For SV, Manta, Lumpy, and DELLY are available.
For CNV, CNVkit and DELLY can be used.
The defaults are Strelka, Manta, and CNVkit respectively.
Advanced users can make selections based on their requirements and understanding of the characteristics of these tools. Additionally, since the detection of copy number mutations often involves the size of the analysis window (in base pairs, bp), modifications can be made at "CNV bin size".

The "Align and preprocess" section contains parameters related to the alignment and preprocessing stage.
It includes some parameters related to BWA: removal of duplicates (by default, duplicate sequences are marked but not removed); whether to use soft clipping or hard clipping when dealing with low-quality read ends; the MAPQ (mapping quality) threshold for the generated BAM file, which is set to 30 by default, meaning only reads with MAPQ >= 30 are included.
Furthermore, there is the option to force replacement of the RG (Read Group). For FASTQ files, there is no need to select this option as RG information is automatically added; for BAM files, when there is no RG information or when it needs to be modified, this option should be selected.

The "Annotation" section allows users to choose whether annotation is required and what content should be annotated in the result file.
The default "Default annotation" includes ENSG gene nomenclature, region type, severity of the mutation result in that type, cDNA position, CDS coding region position, protein position, amino acid mutation details, codon mutation details, and impact.
Users can additionally select: gene name, whether the mutation has been reported, pathogenicity, mutation frequency in the population, records in the Clinvar clinical database, and whether to generate a report of the annotated content. If none of these are selected, it is considered that no annotation is needed and the annotation stage will be skipped.