

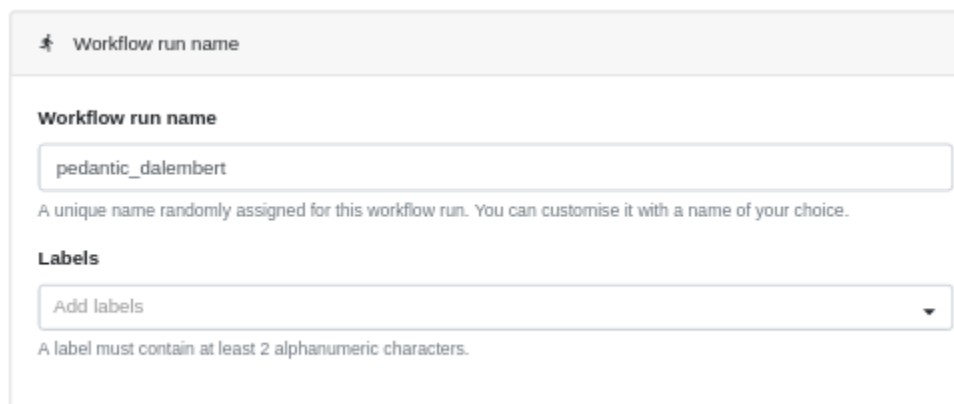
# Nextflow tower pipelines - HGI\_sarek

This pipeline is a fork from nf-core sarek <https://github.com/nf-core/sarek> modified to run on the WTSI LSF farm5 cluster.

## Inputs

The inputs to the pipeline can be defined on the launch page via a form generated by the [nextflow\\_schema.json](#) file in the repo.

- Name your workflow using a name which indicates which data is being processed



The screenshot shows a web form for launching a workflow. At the top, there is a header bar with a star icon and the text "Workflow run name". Below this, the form is divided into two sections. The first section, titled "Workflow run name", contains a text input field with the value "pedantic\_dalembert". Below the input field, there is a note: "A unique name randomly assigned for this workflow run. You can customise it with a name of your choice." The second section, titled "Labels", contains a dropdown menu with the text "Add labels" and a downward arrow. Below the dropdown, there is a note: "A label must contain at least 2 alphanumeric characters."

- There are several entry points (step) to the pipeline, but as of June 2023, only the variant calling entry has been tested
- Specify the CSV file containing details of the input CRAM files to be using in the variant calling process
  - CSV should have four fields "patient,sample,cram,crai" ( patient and sample can be the same )

Input/output options

Define where the pipeline should find input data and save output data.

▶ **step**

variant\_calling

Starting step

**input**

https://raw.githubusercontent.com/nf-core/test-datasets/sarek/testdata/csv/NA12878\_WGS\_3...

Path to comma-separated file containing information about the samples in the experiment.

A design file with information about the samples in your experiment. Use this parameter to specify the location of the input files. It has to be a comma-separated file with a header row. See [usage docs](#).

If no input file is specified, sarek will attempt to locate one in the {outdir} directory.

**outdir**

results

The output directory where the results will be saved. You have to use absolute paths to storage on Cloud infrastructure.

- Most of the main options are self explanatory or not important
  - split\_fastq is not used if entry is at the variant calling step
  - wes should be selected if CRAMs are from exome sequencing
  - intervals is a bed format file indicating your variant calling regions
  - nucleotides\_per\_second is used to determine the size of the regions called so that the jobs complete in an appropriate time ( 40000 is appropriate for WGS and results in approx. 139 fragments for WGS calling ) This number should be reduced depending on the number of input samples 40000 results in a genomicDB import time of around 6hr for 200 WGS samples and a slightly shorter time for the genotypeGVCF step. The job time is not quite linear by sample number but close.
  - the tools tested so far are "haplotypcaller" and "deepvariant" ( GATK gVCFs are only produced if joint calling is selected in the next list of parameters. )

Main options

Most common options used for the pipeline

**split\_fastq**

Specify how many reads each split of a FastQ file contains. Set 0 to turn off splitting at all.

☒ **wes**

Enable when exome or panel data is provided.

**intervals**

Path to target bed file in case of whole exome or targeted sequencing or intervals file.

**nucleotides\_per\_second**

Estimate interval size.

☒ **no\_intervals**

Disable usage of intervals.

**tools**

Tools to use for variant calling and/or for annotation.

**skip\_tools**

Disable specified tools.

- Ignore this section of parameters as they refer to steps in the pipeline before "variant calling"

Preprocessing

Configure preprocessing tools

☒ **save\_bam\_mapped**

Save mapped BAMs.


☒ **save\_output\_as\_bam**

Saves output from Markduplicates & Baserecalibration as BAM file instead of CRAM

**use\_gatk\_spark**

Enable usage of GATK Spark implementation for duplicate marking and/or base quality score recalibration

- Turn on the `joint_germline` option to enable the creation of GATK gVCF files and run GATK joint calling

 Variant Calling

Configure variant calling tools

☒ only\_paired\_variant\_calling

?

If true, skips germline variant calling for matched normal to tumor sample. Normal samples without matched tumor will still be processed through germline variant calling tools.

☒ joint\_germline

?

Turn on the joint germline variant calling for GATK haplotypcaller

- Reference parameters, most of these can be left blank and the pipeline will use appropriate public reference data but enter local files for the genome fasta and fasta index e.g. /lustre/scratch125/humgen/resources/ref/Homo\_sapiens/HS38DH/hs38DH.fa and /lustre/scratch125/humgen/resources/ref/Homo\_sapiens/HS38DH/hs38DH.fa.fai