



Genotoul
Bioinfo

Nextflow nf-core/SAREK

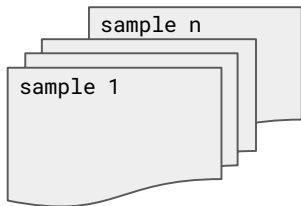


December 2024



- ❖ Context
- ❖ Nextflow / workflow repository (nf-core)
- ❖ Run a workflow
- ❖ Outputs

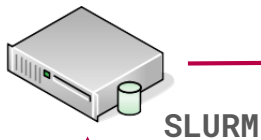




QC



SRUN

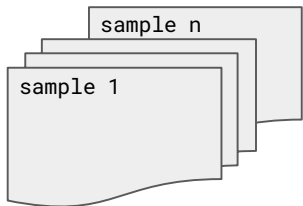


Workflow example





Workflow example



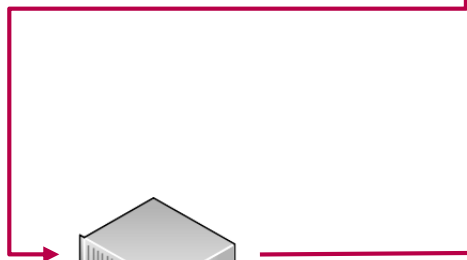
QC

- \$ fastqc sample 1
- \$ fastqc sample 2
- \$ fastqc sample 3
- ...
- \$ fastqc sample n

ALN

- \$ bwa mem sample 1
- \$ bwa mem sample 2
- \$ bwa mem sample 3
- ...
- \$ bwa mem sample n

SRUN



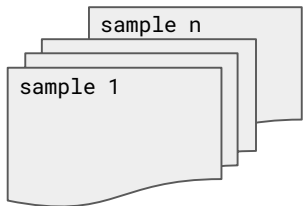
SLURM



HPC



Workflow example



QC

- `$ fastqc sample 1`
- `$ fastqc sample 2`
- `$ fastqc sample 3`
- ...
- `$ fastqc sample n`

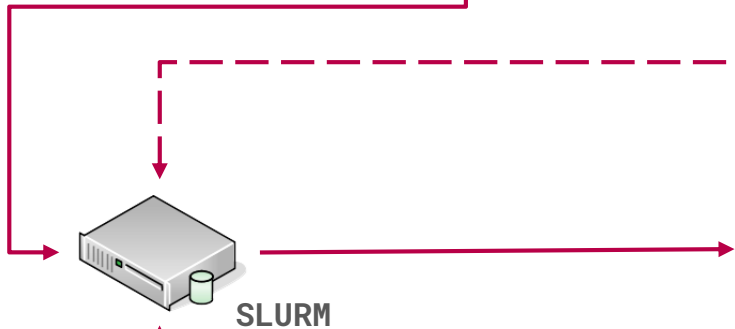
ALN

- `$ bwa mem sample 1`
- `$ bwa mem sample 2`
- `$ bwa mem sample 3`
- ...
- `$ bwa mem sample n`

GATK

- `$ gatk MarkDuplicates sample 1`
- `$ gatk MarkDuplicates sample 2`
- `$ gatk MarkDuplicates sample 3`
- ...
- `$ gatk MarkDuplicates sample n`

SRUN



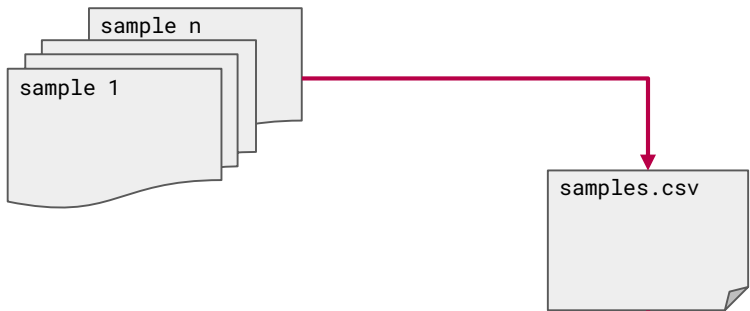
SLURM



HPC

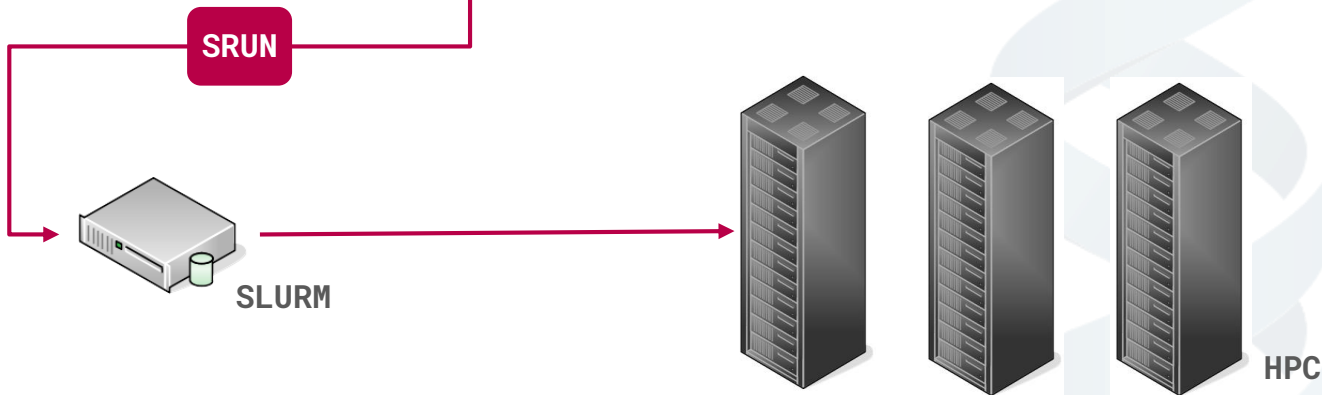


Workflow example



Nextflow

```
$ nextflow run nf-core/sarek --input samples.csv --genome genome.fa --tools haplotypcaller
```





- ❖ Execution and parallelization
- ❖ Reproducibility
- ❖ Workflows **versioning** in a repository
- ❖ **Containers** with dependencies (software)
- ❖ Few manual configuration
- ❖ Same usage on Gentoul, IFB, Amazon...

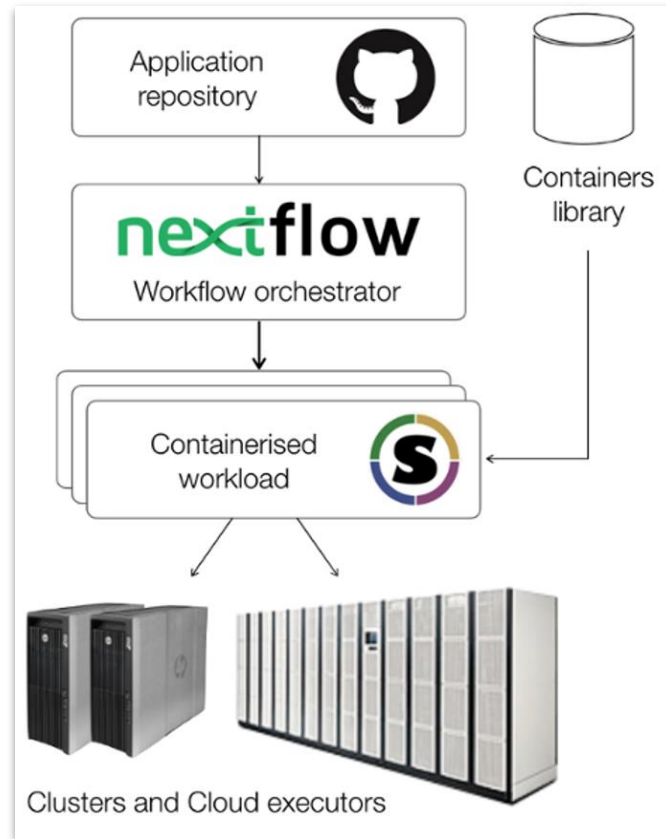




General concepts: nextflow



- ❖ Developed at CRG
- ❖ Java
- ❖ Large user community





General concepts: nextflow



<https://www.nextflow.io/docs/latest>

Nextflow can run a workflow from:

- ❖ a file (.nf)
- ❖ a repository:
 - Github
 - Gitlab
 - BitBucket



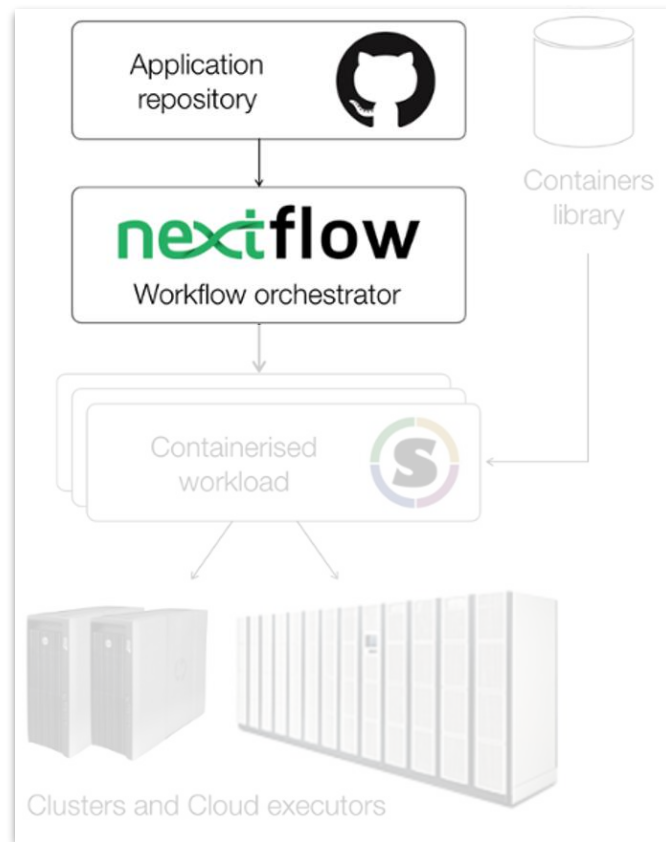


General concepts: repository



<https://nf-co.re/>

nf-core

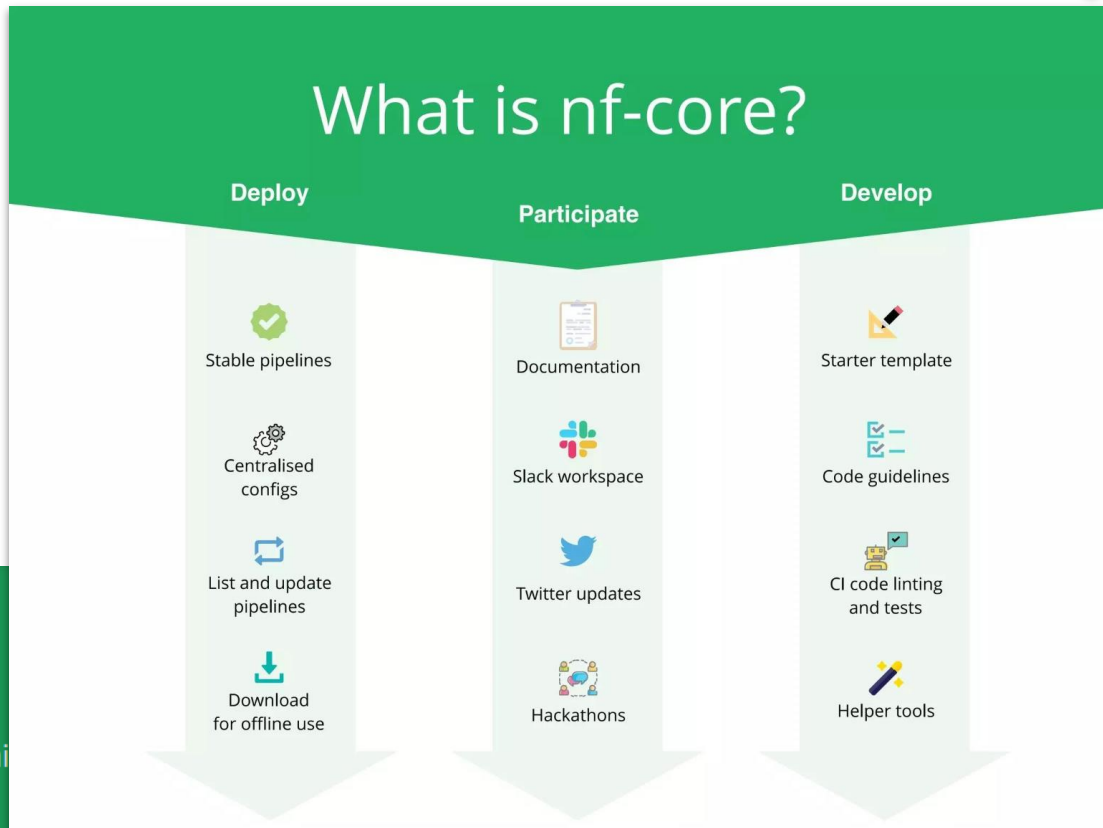




- ❖ Start of 2018
NGI Stockholm
- ❖ A community effort to collect a curated set of analysis pipelines built using Nextflow

Pipelines


Browse the 118 pipelines that are currently available





nf-core



 **slack**

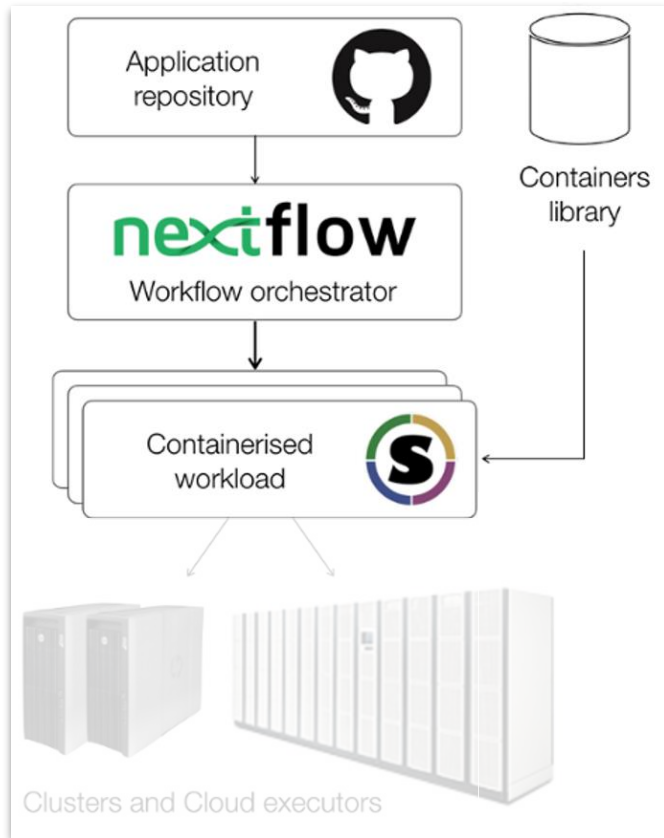
Sign in to nf-core

nfcore.slack.com

You plan to develop workflows?
You can join INRAE mattermost
<https://team.forgemia.inra.fr/tlse-nextflow>



General concepts: containers





General concepts: containers



OR



« allows you to run one or more linux applications inside an isolated and reproducible environment called a container, which shares the linux kernel of the machine you are on »

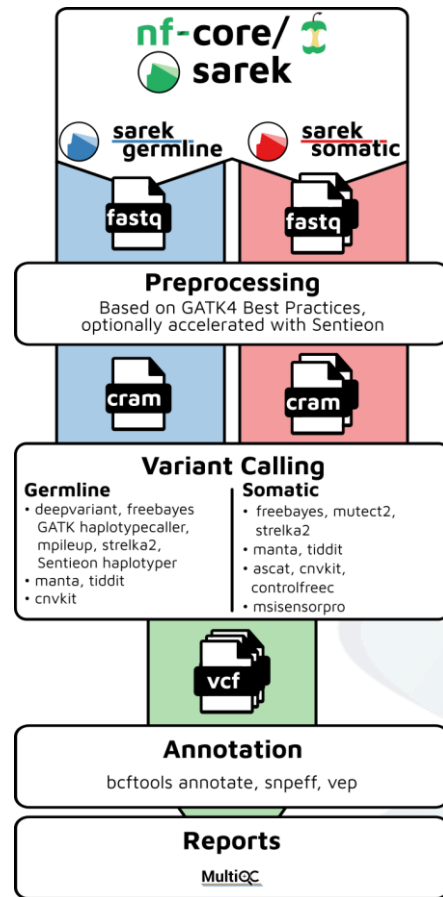
OR

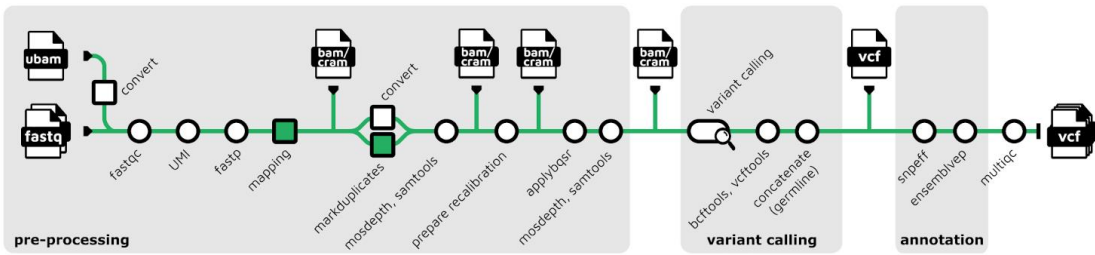


Conda is a package manager and environment management system (based on the system)

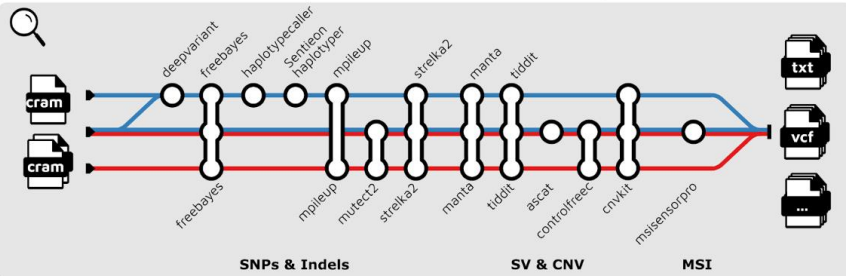


- ❖ Analysis pipeline to detect germline or somatic variants (pre-processing, variant calling and annotation) from WGS / targeted sequencing
- ❖ <https://nf-co.re/sarek>



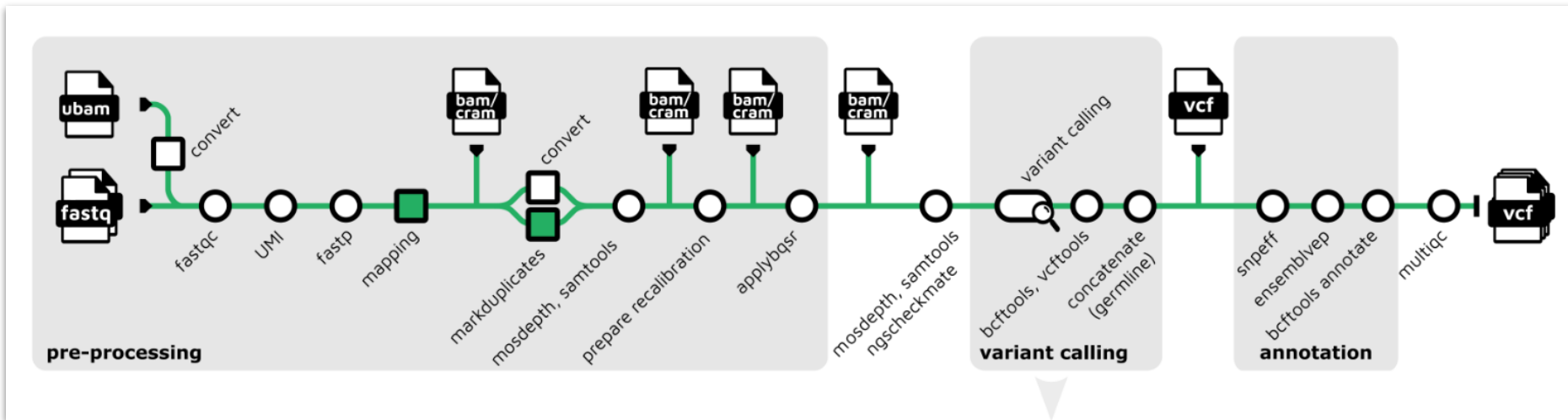


Example analysis pathways



- Mandatory
- Optional
- Optionally Sentieon accelerated
- Core workflow
- Germline variant calling
- Tumor only variant calling
- Tumor-normal pair variant calling

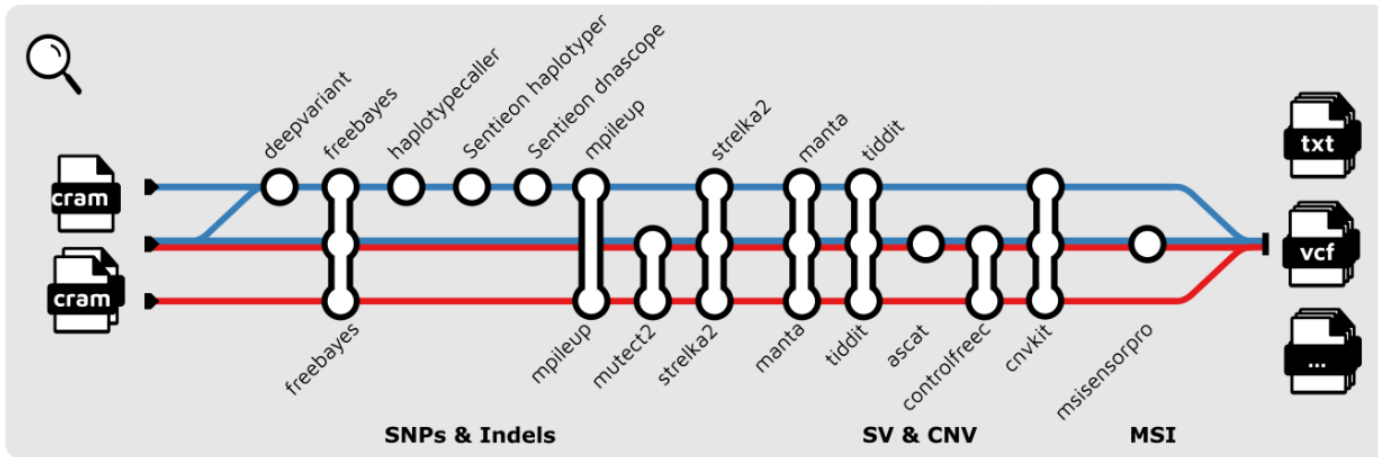
Adapted from: Fellows Yates, James A., et al. PeerJ 9 (2021).



- Mandatory
- Optional
- Optionally Sentieon accelerated
- Core workflow
- Germline variant calling
- Tumor only variant calling
- Tumor-normal pair variant calling



Example analysis pathways



□ Mandatory

○ Optional

■ Optionally Sentieon accelerated

— Core workflow

— Germline variant calling

— Tumor only variant calling

— Tumor-normal pair variant calling

Adapted from: Fellows Yates, James A., et al. PeerJ 9 (2021).



- ❖ Sample sheet with fastq pairs:
 - csv format
 - at least 3 columns
 - header line
 - specified with the `--input` parameter
- ❖ Minimal config file:

```
patient, sample, lane, fastq_1, fastq_2
```

```
patient1, test_sample, lane_1, test_R1.fastq.gz, test_R2.fastq.gz
```



- ❖ You need only a fasta file
- ❖ You can provide known variants
 - either use `--genome` and `--fasta` parameters
 - or configure a `nextflow.config` file

```
params {
  genomes {
    'Gallus_gallus-5.0_25-26' {
      fasta      = "${params.genomes_base}/Gallus_gallus.Gallus_gallus-5.0.dna.toplevel_chr25-26.fa"
      species    = 'Gallus_gallus'
      known_indels = "${params.genomes_base}/Gallus_gallus_incl_consequences_chr25-26.vcf.gz"
    }
  }
}
```



- ❖ The workflow configuration is a merge of several config files found in:
 - user home directory
 - workflow directory
 - current directory

```
$HOME/.nextflow/config  
WORKFLOW_DIR/nextflow.config  
CURRENT_DIR/nextflow.config  
-c myConfig.config
```

- ❖ To ignore any default configuration, use a single custom file with the `-C` command line option

```
nextflow run nf-core/sarek -r 3.2.2 -C myConfig.config
```

- ❖ To know the used configuration

```
nextflow config nf-core/sarek
```



nf-core/sarek: execution



Use singularity image

input and outdir

jobs launched with slurm

task working directory

complete job

failed job

parallelized tasks

```
N E X T F L O W ~ version 24.10.0
Launching `https://github.com/nf-core/sarek` [backstabbing_franklin] DSL2 - revision: 5cc30494a6
[3.4.4]
Core Nextflow options
  revision           : 3.4.4
  runName            : ecstatic_lavoisier
  containerEngine    : singularity
  launchDir          : /work/user/ccabau/nextflow-sarek
  workDir            : /work/user/ccabau/nextflow-sarek/work
  projectDir         : /home/ccabau/.nextflow/assets/nf-core/sarek
  userName           : ccabau
  profile            : genotoul
  configFiles        : /home/ccabau/.nextflow/assets/nf-core/sarek/nextflow.config

Input/output options
  input              : sample.csv
  outdir             : results_chicken

executor > slurm (8)
[0e/b9ecc6] process > get_software_versions           [ 0%] 0 of 1
[a9/d87864] process > BuildBWAindexes (Gallus_gallu... [ 0%] 0 of 1
[34/5854df] process > BuildDict (Gallus_gallus.Ga... [ 0%] 0 of 1
[a1/a2e06b] process > BuildFastaFai (Gallus_gallu... [100%] 1 of 1 ✓
[-          ] process > BuildGermlineResourceIndex    -
[6c/b9bb9c] process > BuildKnownIndelsIndex (Gall... [100%] 1 of 1, failed: 1 X
[-          ] process > BuildPonIndex                    -
[ca/d7f80b] process > BuildIntervals (Gallus_gall... [ 0%] 0 of 1
[a8/7c8106] process > BaseRecalibrator (chicken-S... [ 50%] 2 of 4
[-          ] process > TrimGalore                       [ 0%] 0 of 1
```



- ❖ All nf-core pipelines have:
 - several directories per step
 - a MultiQC output directory
 - a pipeline_info output directory
 - ...





nf-core/sarek: outputs



```
results_chicken
├── multiqc                                # main HTML reports
│   ├── multiqc_data
│   ├── multiqc_plots
│   └── multiqc_report.html
├── pipeline_info
│   ├── execution_report_2023-11-06_17-56-20.html    # HTML CPU, Memory, Time report
│   ├── execution_timeline_2023-11-06_17-56-20.html # HTML timeline
│   ├── execution_trace_2023-11-06_17-56-20.txt     # txt trace
│   ├── params_2023-11-06_18-18-20.json
│   ├── pipeline_dag_2023-11-06_17-56-20.html      # workflow graphic representation
│   └── software_versions.yml                       # version of all softwares
├── preprocessing                          # BAM, BAI
│   ├── markduplicates
│   ├── recalibrated
│   └── recal_table
├── reference                              # all indexes to reuse in other pipeline/execution
├── reports                                # TXT or HTML reports for each step
│   ├── bcftools
│   ├── fastqc
│   ├── markduplicates
│   ├── samtools
│   └── vcftools
└── variant_calling                        # calling results per caller and per sample
    ├── deepvariant
    ├── freebayes
    └── haplotypcaller
```


A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

This report has been generated by the [nf-core/sarek](#) analysis pipeline. For information about how to interpret these results, please see the [documentation](#).

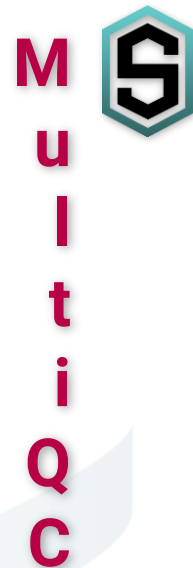
Report generated on 2023-11-06, 18:17 CET based on data in: `/work/user/ccabau/nextflow-sarek-deep/work/d8/37cc491260da8110fec506d45feedf`

Welcome! Not sure where to start? [Watch a tutorial video](#) (6:06) [don't show again](#) ✕

General Statistics

Copy table | Configure Columns | Plot | Showing 19_{r19} rows and 20_{r31} columns.

Sample Name	% Dups	% GC	M Seqs	% Duplication	M Reads After Filtering	GC content	% PF	% Dups	Error rate	M Non-Primary
SRR7062654-1				0.9%	1.9	50.6%	94.9%			
SRR7062654-1.md								2.7%		
SRR7062654-1_1	5.1%	50%	1.0							
SRR7062654-1_2	4.7%	50%	1.0							
SRR7062654.deepvariant										
SRR7062654.md										
SRR7062654.md.cram								0.91%		0.0
SRR7062654.recal										
SRR7062654.recal.cram								0.91%		0.0
SRR7062655-1				1.6%	1.8	50.5%	94.7%			
SRR7062655-1.md								4.3%		
SRR7062655-1_1	6.7%	50%	1.0							
SRR7062655-1_2	6.1%	50%	1.0							





pipeline_info/execution_report



Nextflow Report Summary Resources Tasks

[disturbed_perlman]

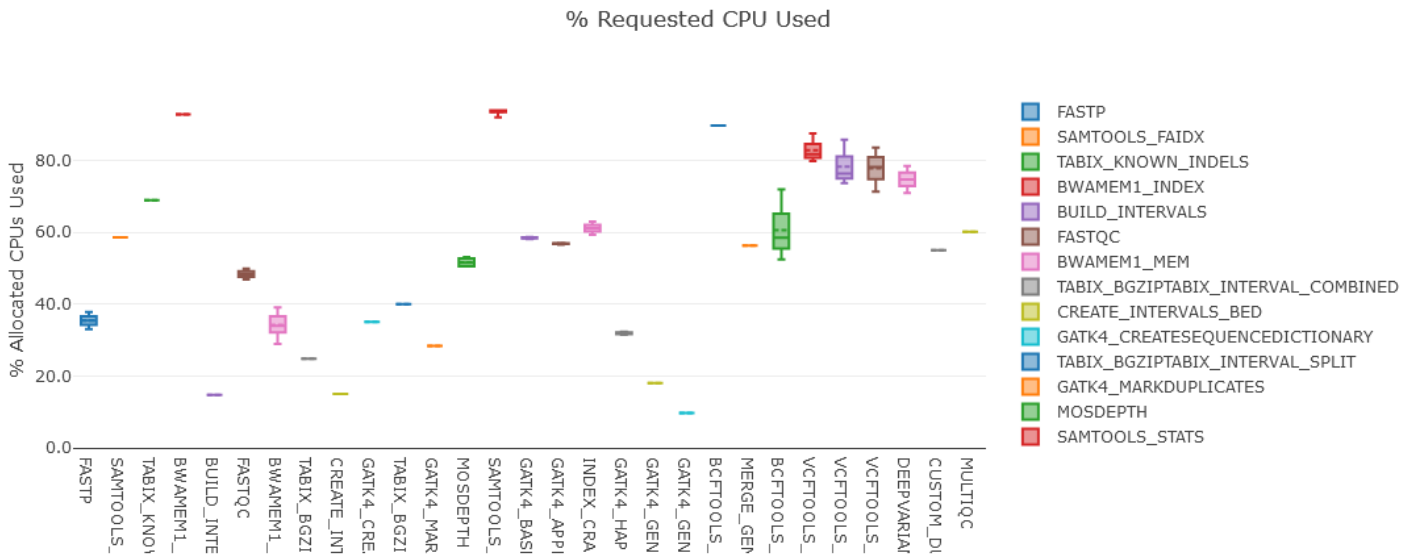
Resource Usage

These plots give an overview of the distribution of resource usage for each process.

CPU

Raw Usage

% Allocated





“That’s all Folks!”

