Contents

NF-IAP PIPELINE

NF-IAP is a Nextflow Illumina Analysis Pipeline meant for processing of DNA (WGS/WES) data.

https://github.com/UMCUGenetics/NF-IAP

The pipeline performs the following tasks.

- Read QC (FastQC)
- Mapping (BWA)
- QC (GATK)
- PCR duplicate detection (Sambamba MarkDup)
- Variant calling and filtering (GATK)
- Variant annotation (snpEff and GATK)
- CNV calling (Control-FREEC)
- SV calling (Manta)
- QC report (MultiQC)

Download and information

The pipeline can be downloaded from the link below:

https://github.com/UMCUGenetics/NF-IAP

Follow the "Installing & Setup" section to set up the pipeline for own use.

Output description

Descriptions of the contents of the resulting processed data folder.

- **.nextflow/**
  Nextflow cache and logging files, not of importance for the user but kept for completeness
- **BAMS/**
  The resulting aligned data in .bam format with corresponding .bam.bai

index files. These can be viewed in for example the IGV genome browser or used for further analysis

- **CNV/**
  The Copy Number Variation results, when requested, for the corresponding tools in this pipeline
  - **CNV/FREEC/**
    Control-FREEC CNV results
- **configs/**
  A copy various general configuration files used for the pipeline, containing parameters, resources, etc
- **log/**
  Pipeline logs and nextflow report files (regarding resource usage, time taken, etc)
- **QC/**
  Various QC logs and reports
  - **QC/fastqc/**
    FastQC reports on the raw fastq files
  - **QC/multiple_metrics/**
    Collection of metrics as collected by CollectMultipleMetrics (Picard), such as: gc bias, base qualityscore distribution, etc. See https://gatk.broadinstitute.org/hc/en-us/articles/360037594031-CollectMultipleMetrics-Picard-
  - **QC/summary/**
    MultiQC collects all information from the various QC modules as well as other tools (STAR, sambamba) and shows this in a summary overview in order to get a good and quick overview of the statistics of the various steps in the pipeline for this run
  - **QC/wgs_metrics/**
    Collection of metrics about coverage and performance of WGS experiments as collected by CollectWgsMetrics (Picard). See: https://gatk.broadinstitute.org/hc/en-us/articles/360037269351-CollectWgsMetrics-Picard-
- **SV/**
  The Structural Variants results, when requested, for the corresponding tools in this pipeline
  - **SV/MANTA/**
    Manta SV results
- **VCFS/**
  The variants called in both VCF and GVCF format for the samples in this run

  - **VCFS/VCF/**
    The variants called with GATK in VCF format, both filtered and annotated versions when requested

  - **VCFS/GVCF/**
    GVCF file for each sample

- **work/**
  The Nextflow working directory containing logs and script from the various cached steps. Kept for completeness and not of importance for the user

MultiQC information

MultiQC generates a summary report using the logs and reports generated by a selection of the various tools in the piplines.

General Statistics The General Statistics provide an overview of the various modules, by hoovering over the column headers one can see from which specific tool this column was gathered. The Sample Name column lists both the actual sample (with info regarding the aligned bam file) as well as the various fastq files containing the raw reads for that sample (and the corresponding fastqc information).

Picard These metrics are generated using the GATK (picard) toolkit, see the relevant documentation for more indepth information on these modules:
CollectWgsMetrics (Picard)
CollectMultipleMetrics (Picard)

FastQC Please refer to the FastQC documentation for the interpretation of the various modules:
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Software

Tools used:

- Samtools value_samtools
  http://www.htslib.org/doc/samtools.html

- BWA v0.7.17
  http://bio-bwa.sourceforge.net/

- Control-FREEC v11.5
  http://boevalab.inf.ethz.ch/FREEC/

- FastQC v0.11.5
  https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

- GATK v4.1.3.0
  https://gatk.broadinstitute.org/hc/en-us

- Manta v1.6.0
  https://github.com/Illumina/manta

- MultiQC
  https://multiqc.info/

- Sambamba v0.6.8
  https://lomereiter.github.io/sambamba/docs/sambamba-flagstat.html

- snpEff v4.3.t
  http://pcingola.github.io/SnpEff/

Parameters

When parameters are not specifically specified, the default parameters are used

```
// Custom settings of tools.
params.freec_ploidy = 2
params.freec_window = 1000
params.freec_telocentromeric = 50000
params.freec_maxlevel = 4

params.bwa.optional = '-M -c 100'
params.bwaindex.optional = '-a bwtsw'

params.haplotypecaller.optional = '-ERC GVCF'
params.collectmultiplemetrics.optional = '--PROGRAM CollectAlignmentSummaryMetrics --PROGRAM
params.markdup.optional = '--overflow-list-size=2000000'

params.snpefffilter.optional = 'GRCh37.75 -hgvs -lof -no-downstream -no-upstream -no-interge
params.snpsiftsbnsfp.optional = '-f hg38_chr,hg38_pos,genename,Uniprot_acc,Uniprot_id,Unipro
params.snpsiftannotate.optional = '-tabix -name GoNLv5 -info AF,AN,AC'
```

Resources

The following resources and pipeline/nextflow versions were used for this run:

- *Pipeline Version:*
  value__pipeline

- *Nextflow Version:*
  value_nextflow

- *Genome:*
  value_genome

- *Genome fasta:*
  value_fasta

Material and Methods

NF-IAP analysis

Quality control on the sequence reads from the raw FASTQ files was done with FastQC (v0.11.5). Reads were aligned to the reference genome (value_genome) using BWA-mem (v0.7.17) after which they were sorted and index with Samtools (value_samtools). Duplicates were marked using Sambamba (v0.6.8). Followup QC on the mapped (bam) files was done using GATK (v4.1.3.0) CollectMultipleMetrics and CollectWGSMetrics. Variant calling is done using GATK best practises, including BaseQualityScoreRecalibration (BQSR), HaplotypeCaller, VariantSelection and VariantFiltration. Annotation on the resulting VCF file was done using GATK (v4.1.3.0) and/or snpEff (v4.3.t), depending on the organism/genome. Control-FREEC (v0.11.5) was used for CNV analysis, while Manta (v1.6.0) was used for SV analysis. Finally a summary report was created using MultiQC (v1.5).