Nextflow Data Processing & Configuration

⚠️ Our pipeline configuration is still in-development, and the contents of this document are subject to change.

https://nf-co.re/sarek/usage v2.7.1			
Data Type	**Method**	**Output**	**Tried yet?**
WES or WGS	DeepVariant	Germline SNV, INDEL	Yes
WES or WGS	Strelka, Mutect2, Freebayes	Somatic SNV, INDEL	
WES or WGS	TBD	Germline and Somatic Structural Variants	
WES or WGS	TBD	Germline and Somatic CNV	
WES or WGS	TBD	Tumor MSI	
SNV, INDEL variants	TBD	Annotated Variants	

https://nf-co.re/rnaseq v3.5			
Data Type	**Method**	**Output**	**Tried yet?**
RNA-Seq	Salmon	Gene expression counts	Yes

bam/cram to fastq conversion

When fastq files are not available, cram/bam files are converted to fastq using this pipeline: GitHub - qbic-pipelines/bamtofastq (v1.2.0).

WES and WGS Variant Calling (SNV & INDEL)

Germline SNV + INDEL

This involves transformation of WES `fastq` or `cram` files to variant call files in VCF format (`.vcf` files).

As of Jan 2022, the `reference genome` used is `GRCh38`.

The processing steps include the following:

- Raw fastq files uploaded to Synapse by researcher in a folder with name format `experiment_name_rnaseq_fastq_date` . No white space should be present in the filenames (all filenames should have `_` for whitespaces.
- All experiment and sample related annotations need to be added on Synapse before processing can start. This is a required step so that a sample sheet can be generated to trigger the processing workflow
- The sample sheet should contain the following information in the following format (saved as a `.txt` file) :

sample	subject	status	sex	file_1	file_2	lane	parentId
Synapse specimenID	Synapse individualID	1 (Tumor = 1, Normal 0)	XX or XY	synID	synID	Lane information	SynapseID of parent folder

- The files are pulled into NextFlow workflow setup and processed using the following versions of software:

```
nf-core/sarek  v2.7.1
Nextflow   v21.10.5
BWA    0.7.17
```

```
GATK   v4.1.7.0
FreeBayes  v1.3.2
samtools   v1.9
Strelka    v2.9.10
Manta  v1.6.0
TIDDIT v2.7.1
AlleleCount    v4.0.2
ASCAT  v2.5.2
Control-FREEC  vv11.6
msisensor  v0.5
SnpEff v4.3t
VEP    v99.2
MultiQC    v1.8
FastQC v0.11.9
bcftools   v1.9
CNVkit v0.9.6
htslib v1.9
QualiMap   v2.2.2-dev
Trim Galore    v0.6.4_dev
vcftools   v0.1.16
R  v4.0.2
```

Commands used for running JHU samples on DeepVariant:

All files and sample sheets are first staged in S3 buckets linked to NFTower. then the following command are used to launch the processing pipeline.

`Params`:

```
input: s3://jhu-biobank-nf-project-tower-bucket/jobs/02-sage-sarek-2.7.1-deepvariant/inputs/sample-sheet.tsv
outdir: s3://jhu-biobank-nf-project-tower-bucket/jobs/02-sage-sarek-2.7.1-deepvariant/outputs/
genome: GRCh38
igenomes_base: s3://sage-igenomes/igenomes
model_type: WES
tools: "deepvariant"
```

`Config`:

```
process {
  errorStrategy = 'retry'
  maxRetries = 3

  withLabel:deepvariant {
    container = "google/deepvariant:1.1.0"
    cpus = 24
  }
}
```

`Pre-run Script`:

```
export NXF_VER=21.10.5
```

`Profiles`:

```
aws_tower
```

Somatic SNV + INDEL

TBD

Annotated Variants

Currently, germline variant calls in VCF format are being processed manually using VEP and `vcf2maf`

RNA SEQUENCING DATA QUANTIFICATION

Processing RNA-seq files involve transformation of raw data (`fastq` files) to transcript counts (`quants.sf` files).

The quantification software of choice is `Salmon`.

As of Jan 2022, the `reference genome` used is `GRCh38`.

Processing involves the following steps:

- Raw fastq files uploaded to Synapse by researcher in a folder with name format `experiment_name_rnaseq_fastq_date` . No white space should be present in the filenames (all filenames should have `_` for whitespaces.
- All experiment and sample related annotations need to be added on Synapse before processing can start. This is a required step so that a sample sheet can be generated to trigger the processing workflow
- The sample sheet should contain the following information in the following format (saved as a `.csv` file):

sample	single_end	fastq_1	fastq_2	strandedness
Synapse specimenID	0 (1 if paired-end)	synID	synID	reverse or forward

- The files are pulled into NextFlow workflow setup and processed using the following versions of software:

```
BEDTOOLS_GENOMECOV:
bedtools: 2.30.0
CAT_FASTQ:
cat: 8.3
CUSTOM_DUMPSOFTWAREVERSIONS:
python: 3.9.5
yaml: 5.4.1
DESEQ2_QC_STAR_SALMON:
bioconductor-deseq2: 1.28.0
r-base: 4.0.3
DUPRADAR:
bioconductor-dupradar: 1.18.0
r-base: 4.0.2
FASTQC:
fastqc: 0.11.9
GET_CHROM_SIZES:
samtools: 1.1
GTF_GENE_FILTER:
python: 3.8.3
PICARD_MARKDUPLICATES:
picard: 2.25.7
PRESEQ_LCEXTRAP:
preseq: 3.1.1
```

```
QUALIMAP_RNASEQ:
qualimap: 2.2.2-dev
RSEM_PREPAREREFERENCE_TRANSCRIPTS:
rsem: 1.3.1
star: 2.7.6a
RSEQC_BAMSTAT:
rseqc: 3.0.1
RSEQC_INFEREXPERIMENT:
rseqc: 3.0.1
RSEQC_INNERDISTANCE:
rseqc: 3.0.1
RSEQC_JUNCTIONANNOTATION:
rseqc: 3.0.1
RSEQC_JUNCTIONSATURATION:
rseqc: 3.0.1
RSEQC_READDISTRIBUTION:
rseqc: 3.0.1
RSEQC_READDUPLICATION:
rseqc: 3.0.1
SALMON_QUANT:
salmon: 1.5.2
SALMON_SE_GENE:
bioconductor-summarizedexperiment: 1.20.0
r-base: 4.0.3
SALMON_TX2GENE:
python: 3.8.3
SALMON_TXIMPORT:
bioconductor-tximeta: 1.8.0
r-base: 4.0.3
SAMPLESHEET_CHECK:
python: 3.8.3
SAMTOOLS_FLAGSTAT:
samtools: 1.13
SAMTOOLS_IDXSTATS:
samtools: 1.13
SAMTOOLS_INDEX:
samtools: 1.13
SAMTOOLS_SORT:
samtools: 1.13
SAMTOOLS_STATS:
samtools: 1.13
STAR_ALIGN:
star: 2.6.1d
STRINGTIE:
stringtie: 2.1.7
TRIMGALORE:
cutadapt: 3.4
trimgalore: 0.6.7
UCSC_BEDCLIP:
ucsc: 377
UCSC_BEDGRAPHTOBIGWIG:
ucsc: 377
Workflow:
Nextflow: 21.10.5
nf-core/rnaseq: '3.4'
```

Command used to process JHU Biobank samples:

`Params`:

```
input: s3://jhu-biobank-nf-project-tower-bucket/jobs/01-nfcore-rnaseq-3.4/inputs/sample-sheet.csv
outdir: s3://jhu-biobank-nf-project-tower-bucket/jobs/01-nfcore-rnaseq-3.4/outputs/
genome: GRCh38
igenomes_base: s3://sage-igenomes/igenomes
```

`Config`:

```
process {
  errorStrategy = 'retry'
  maxRetries = 3
}
```

`Pre-run script`:

```
export NXF_VER=21.10.5
```

`Profile`:

```
aws_tower
```