Wrocław University
of Environmental
and Life Sciences

# NextFlow vs Bash

Different approaches to SNP calling parallelisation on the Whole Genome Bovine Sequence

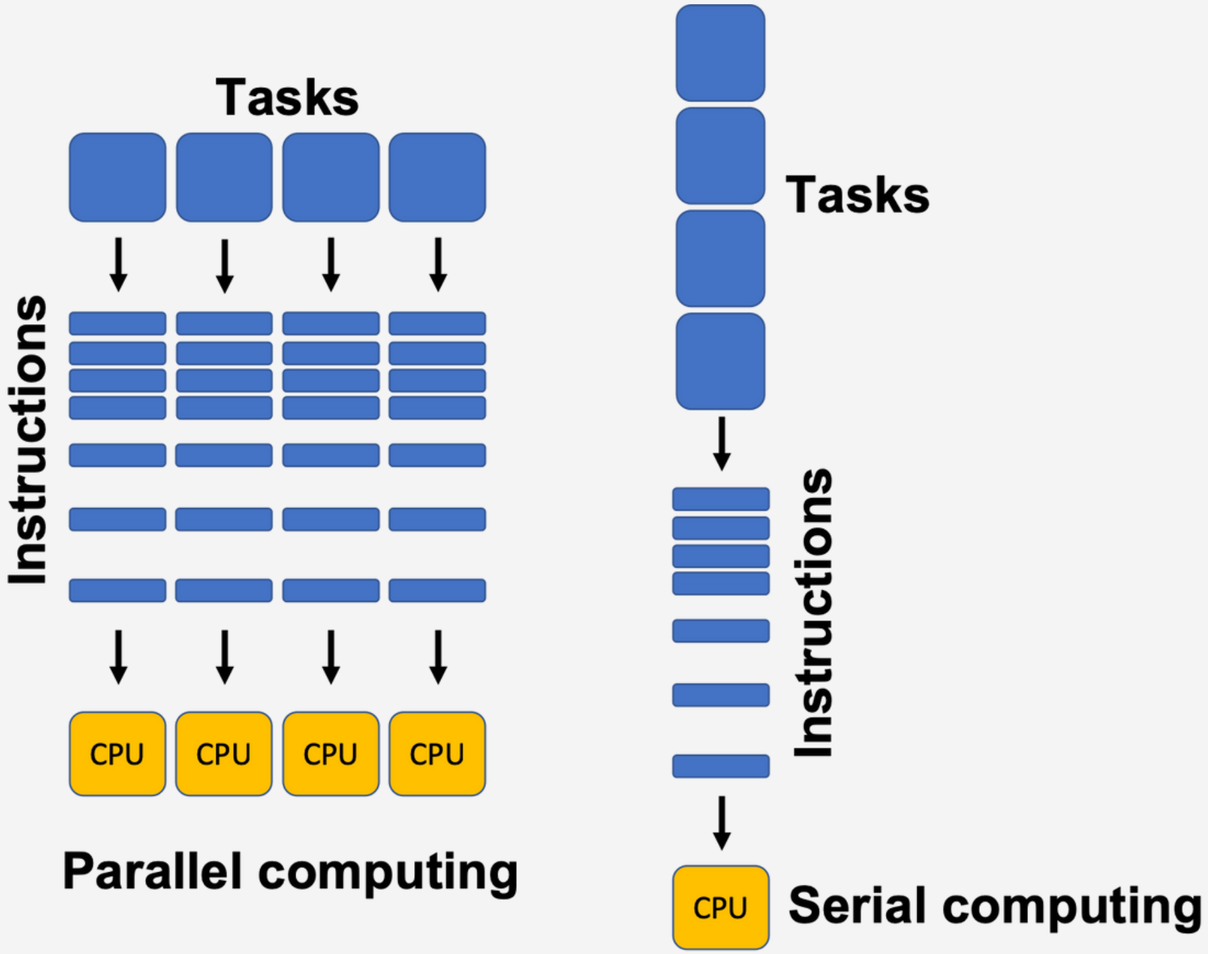
**M. Sztuka**, P. Hajduk, J. Liu, K. Kotlarz, M. Mielczarek and J. Szyda

# Variant Calling

- Aligment to reference genome
- Polymorphing DNA Variant Calling
- Big data

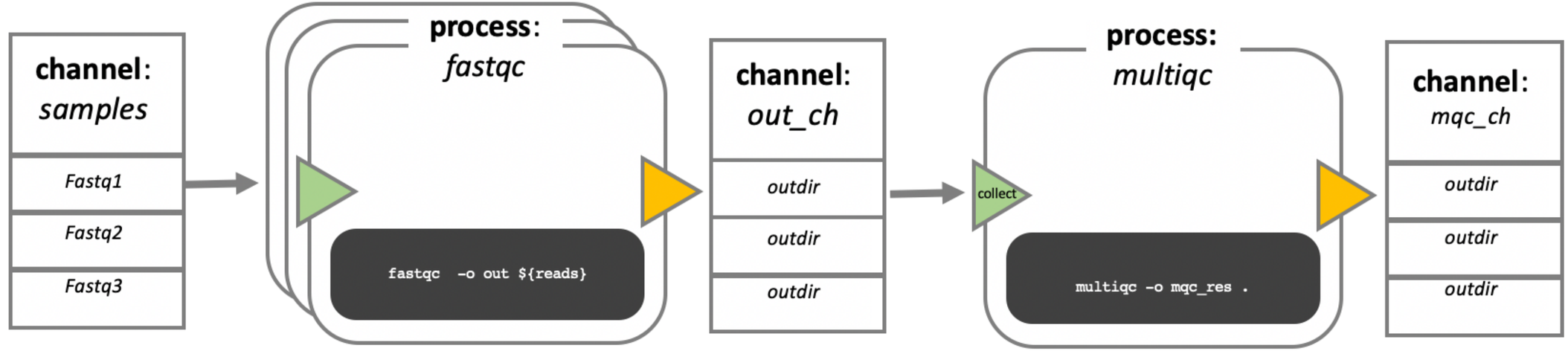
# Parallel computing

- Modern CPU's have multiple cores
- Each core is individual computing machine
- Significant execution time difference
- Not everything can be parallelised



# NextFlow

- Workflow creation and management system
- Open source platform
- Scalable and replicable pipelines

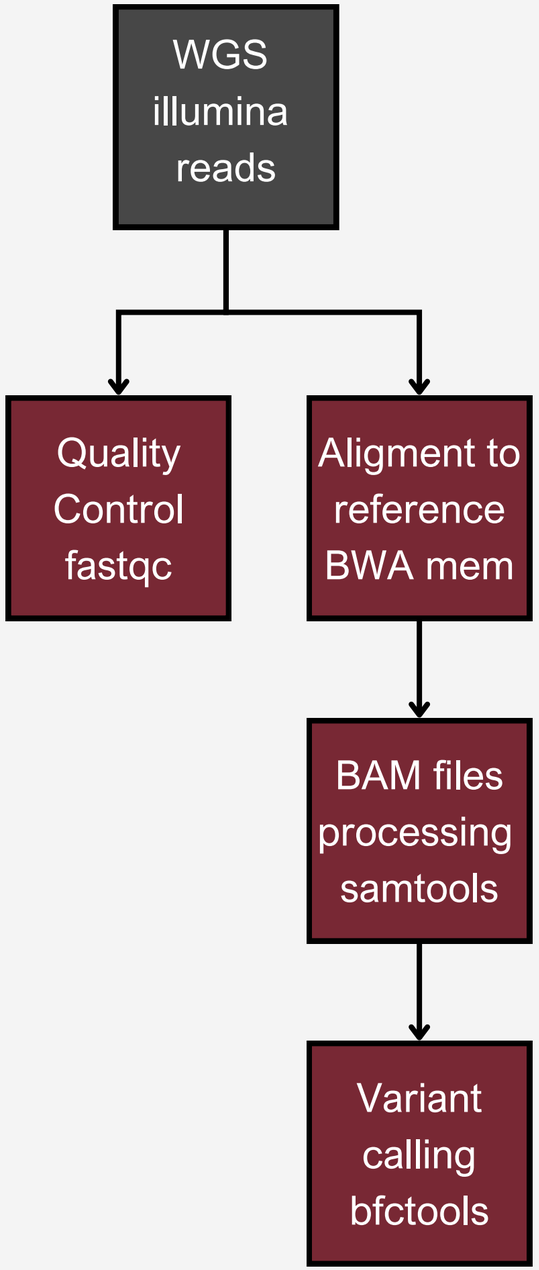


# Material

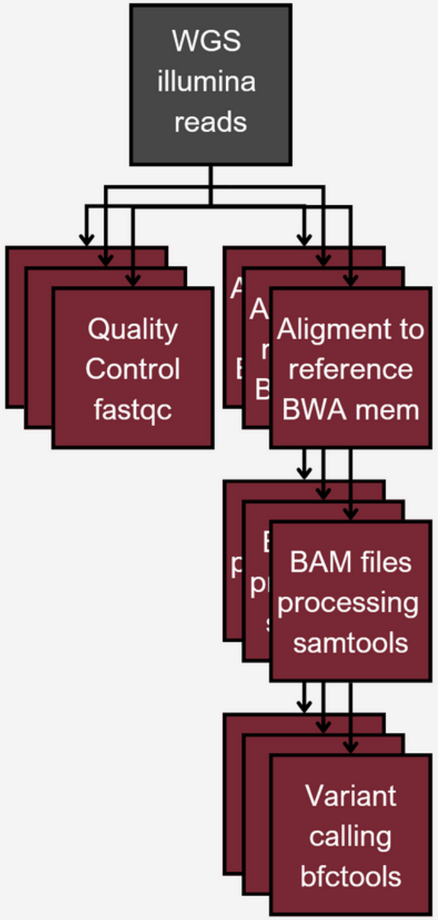
- WGS sequencing of 5 Holstein-Friesian cows
- Only Chromosome 25 was used (BTA 25) - 3,450,967 to 3,603,816 reads
- Illumina Hiseq 2000
- Computing device :
  - 44 Cores
  - 88 Threads
  - 2.2GHz
  - 188 Gb RAM

# Pipeline

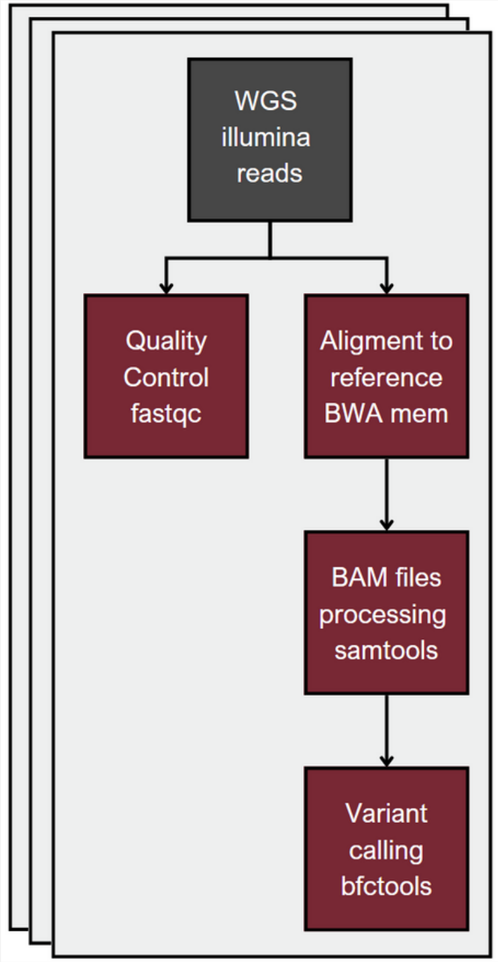
- Quality control (fastQC)
- Alignment (bwa mem)
- Post alignment (samtools)
- Variant calling (bcftools)



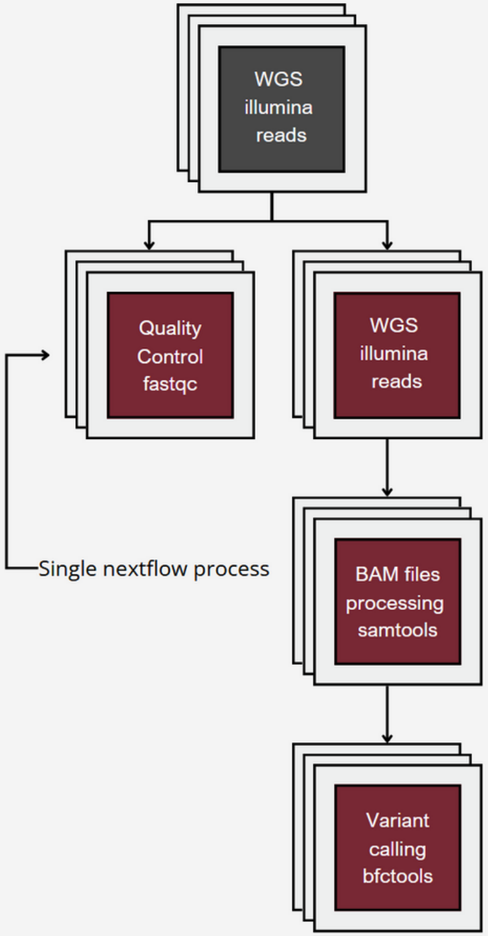
# 3 approaches to parallelisation




Bash loop




Single process nextflow

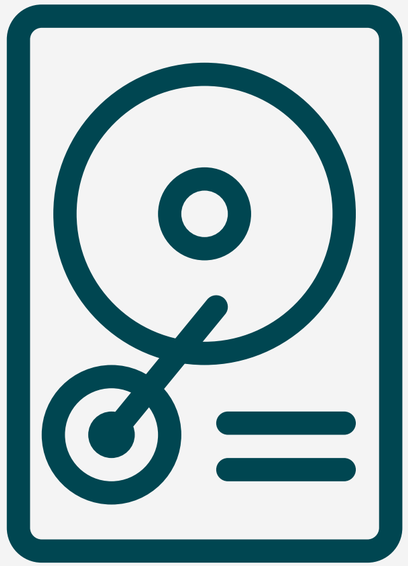



Multi-process Nextflow

# Comparison

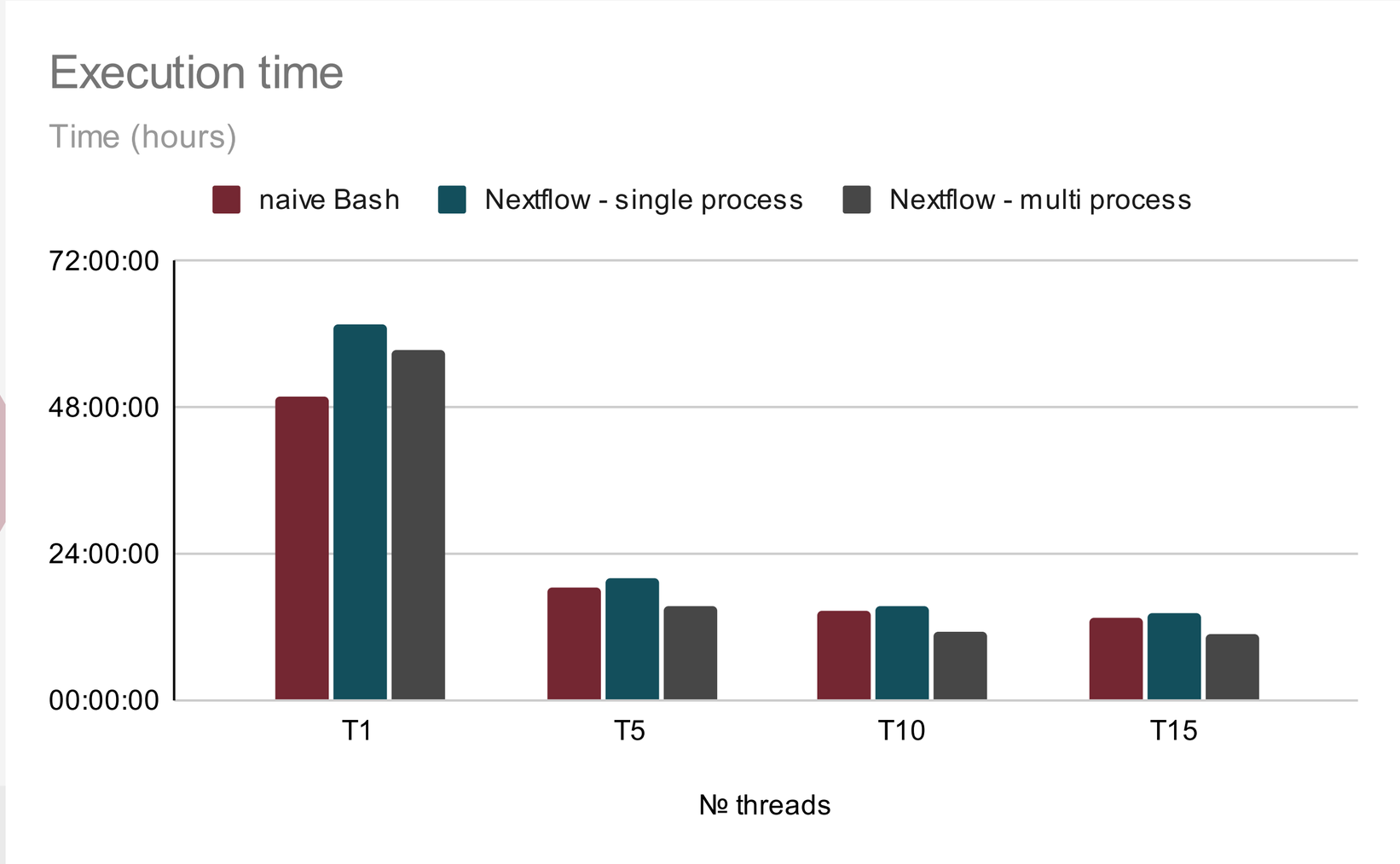
Execution time


Memory usage

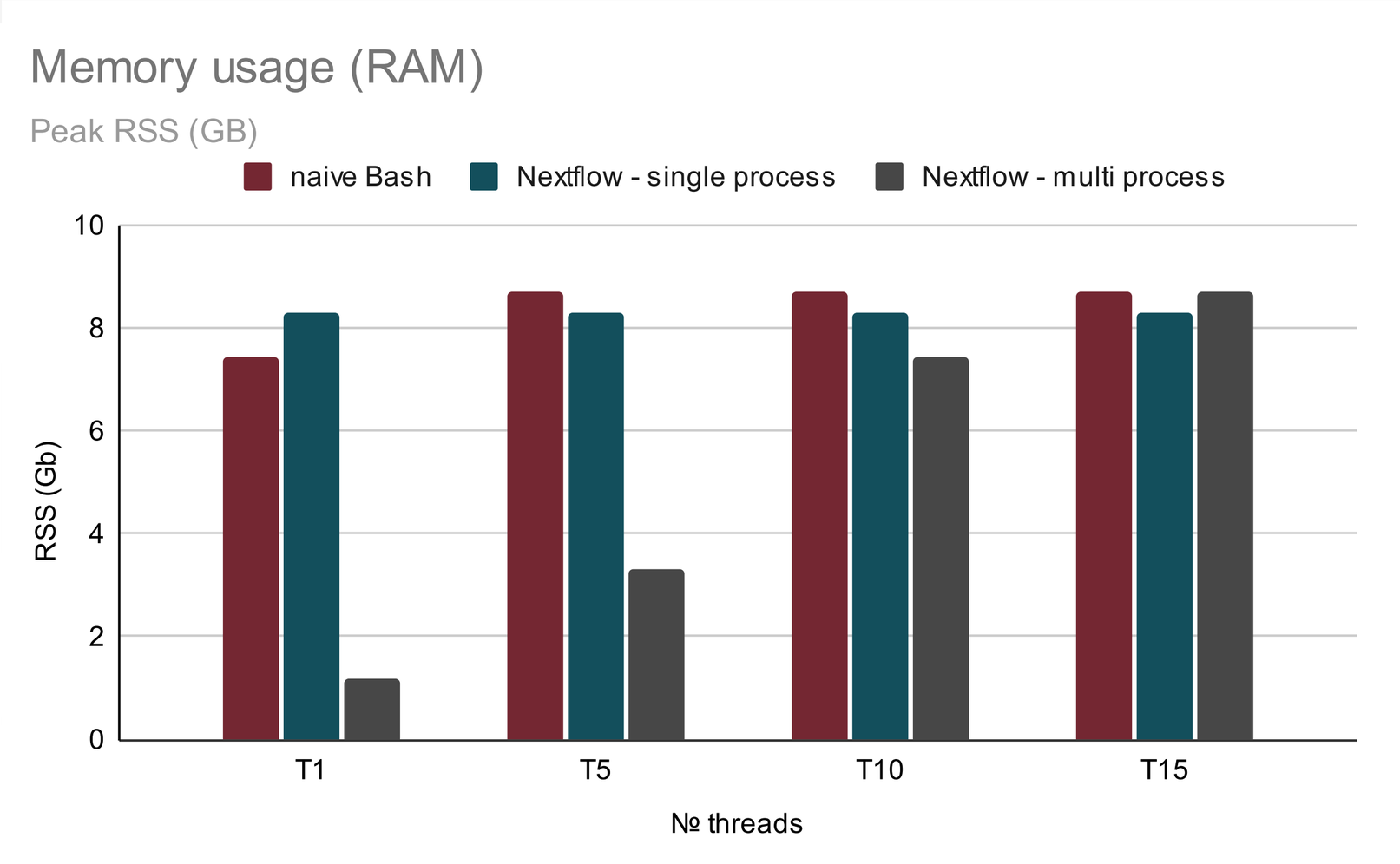


Hard drive space usage

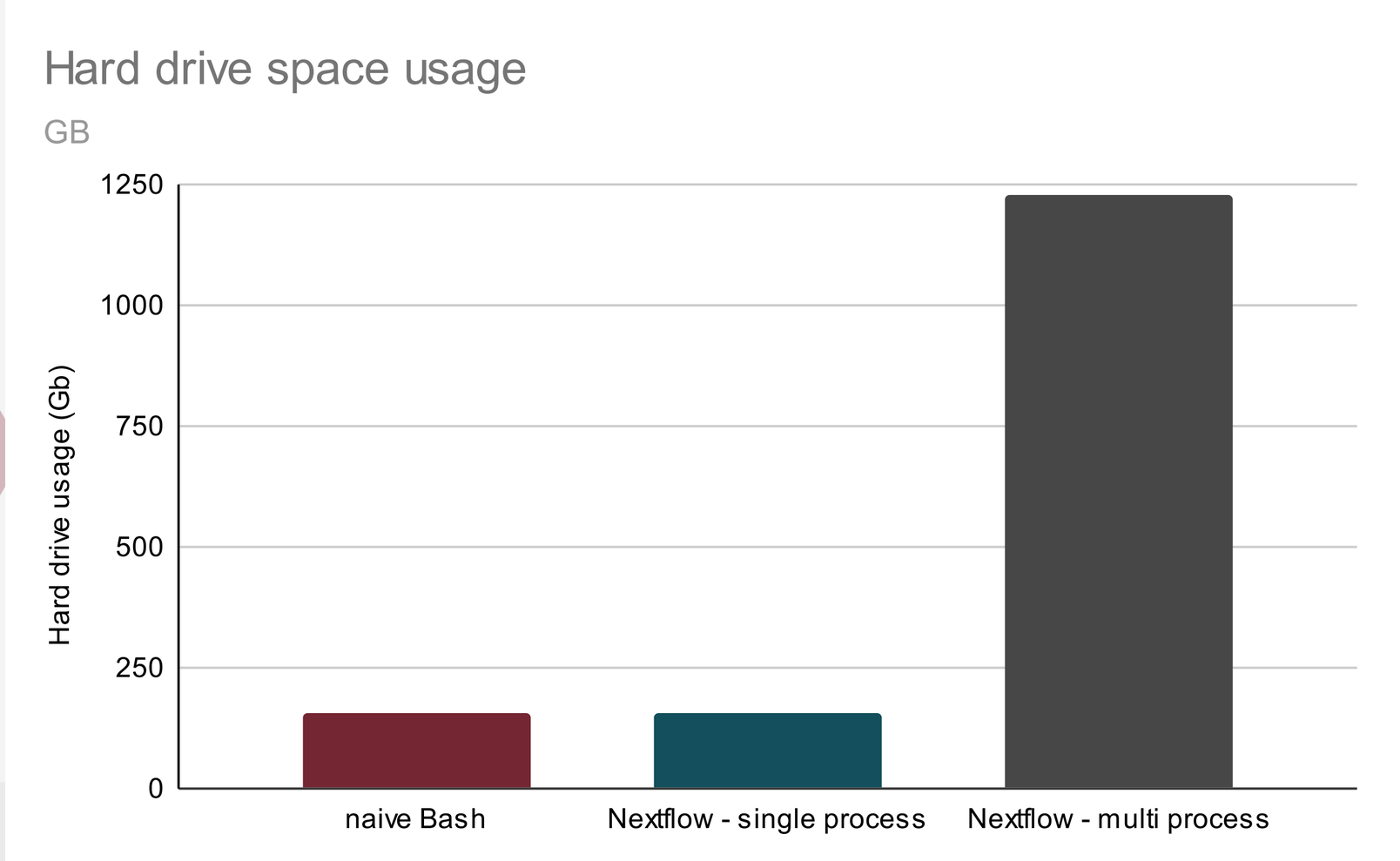
# Results - execution time



# Results - memory usage



# Results - Hard drive space usage

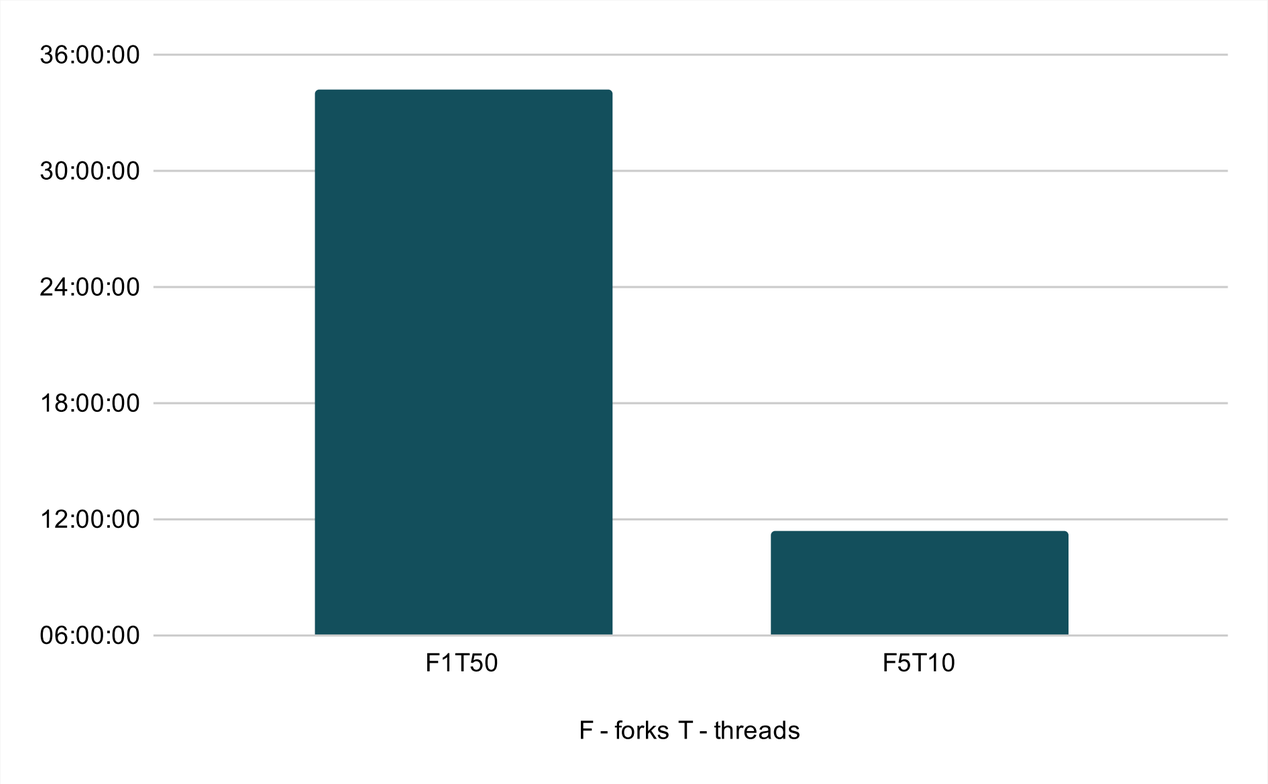


# Internal vs NF parallelisation

Time based comparison

Two Configurations:

- F1T50
- F5T10



# Conclusions

- Each pipeline generated VCF files with same number of SNP’s
- HTML files with quality control reports
- In almost every configuration Multi process nextflow was the fastest
- Memory usage was similar for larger amounts of threads (10-15)
- Nextflow parallel approach was 3 times faster than sequentional approach
- Nextflow advantage is user friendly approach to workflow creation and management

nextflow

Fin

# Leading Research Group **THETA**

## THE BIOSTATISTIC GROUP

LEADER

**PROFESSOR JOANNA SZYDA**

WROCŁAW UNIVERSITY
OF ENVIRONMENTAL
AND LIFE SCIENCES

BIOSTATISTICS GROUP
WROCLAW, POLAND