جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Core Labs and Research Infrastructure

Shaheen 101 Training

Application software example: Bioinformatics workflow

Nagarajan Kathiresan, Ph.D.,

KAUST Supercomputing Core Laboratory (KSL),

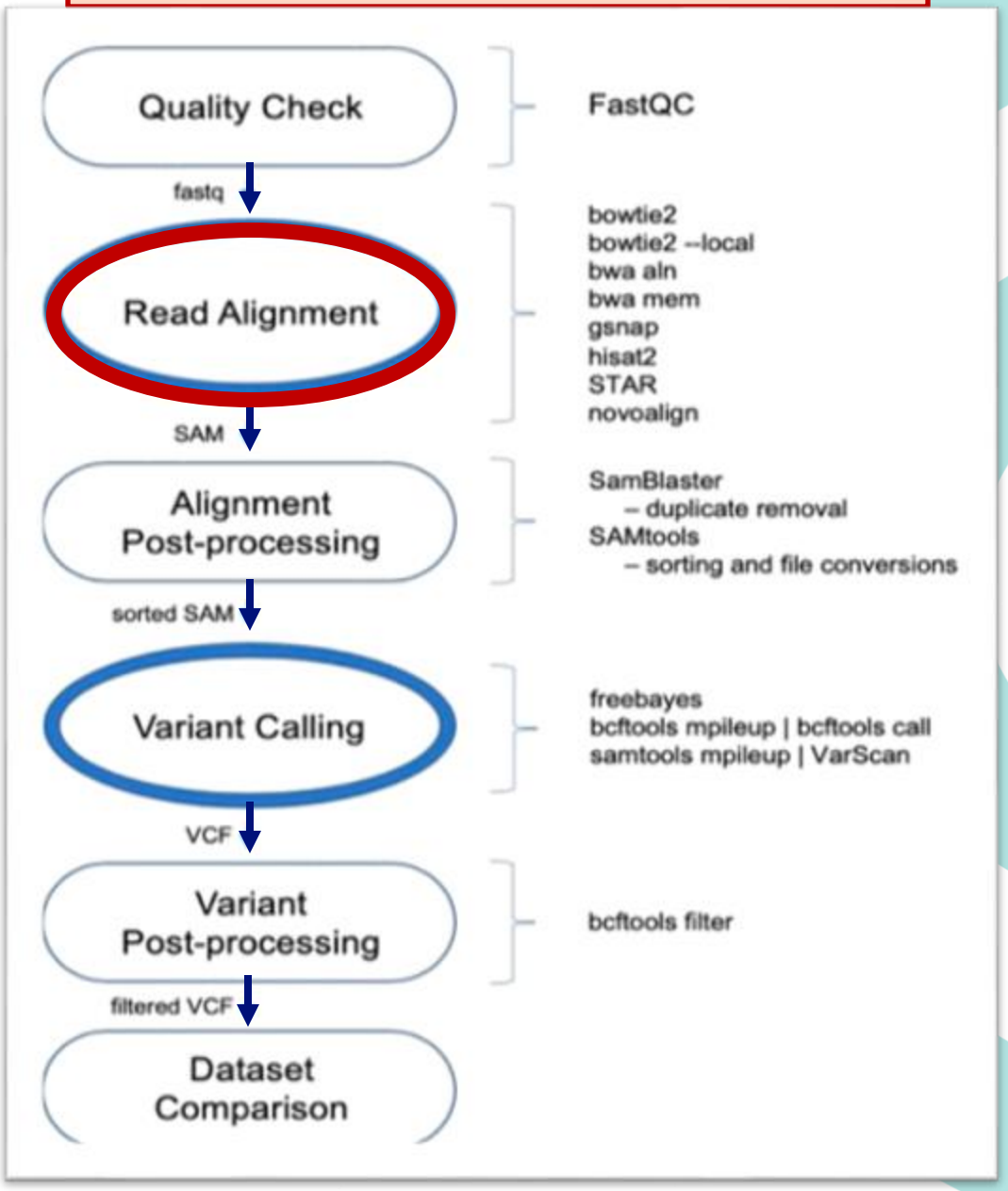
Building 1, Level 0, Office 0126

Example bioinformatics workflow

Definition[1]: A bioinformatics workflow is an organized and automated sequence of data processing steps used to analyze biological data, integrating various software tools, algorithms, and data formats to answer specific biological questions.

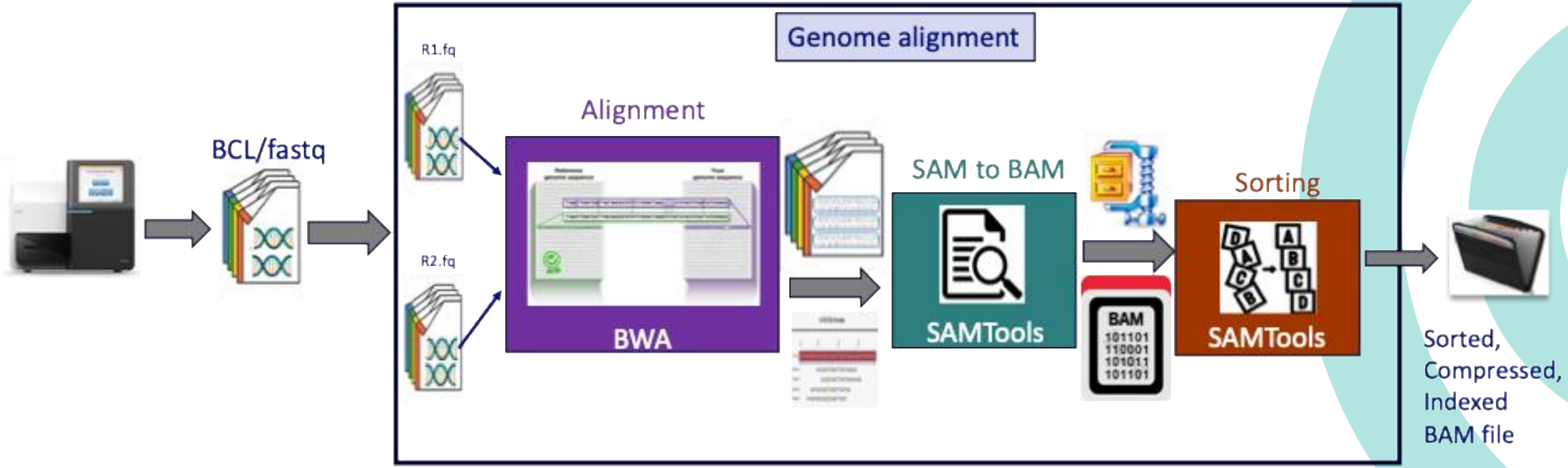
[1]Source: BMC Bioinformatics, 6:87. https://doi.org/10.1186/1471-2105-6-87

Schematic of the variant calling pipeline



Workflow Source: International Journal of Molecular Sciences, 22(19), 10400 https://doi.org/10.3390/ijms221910400

Genome Alignment!



Best practices of GATK: *Bioinformatics.* 2013;43:11.10.1-11.10.33
DOI:10.1002/0471250953.bi1110s43

Script!

Step #1:

```
bwa mem -t $CORES $REF $FR.fastq.gz $RR.fastq.gz > aligned_reads.sam
```

Step #2:

```
samtools view -bS -t $CORES aligned_reads.sam > aligned_reads.bam
```

Step #3:

```
samtools sort aligned_reads.bam -o sorted_aligned_reads.bam
```

Fast and accurate short read alignment with Burrows–Wheeler transform: *Bioinformatics.* 2009;25(14):1754-1760. DOI:10.1093/bioinformatics/btp324.
The Sequence Alignment/Map format and SAMtools: *Bioinformatics.* 2009;25(16):2078-2079. DOI:10.1093/bioinformatics/btp352.

All-in-One!

```
$ bwa mem -t $CORES $REF $FR.fastq
$RR.fastq | samtools view -@ $CORES
-b -S - | samtools sort - >
sorted_aligned_reads.bam
```

Application modules

$ export MODULEPATH=/scratch/project/software/ex111genoa/modulefiles:$MODULEPATH

$ module av -S bio

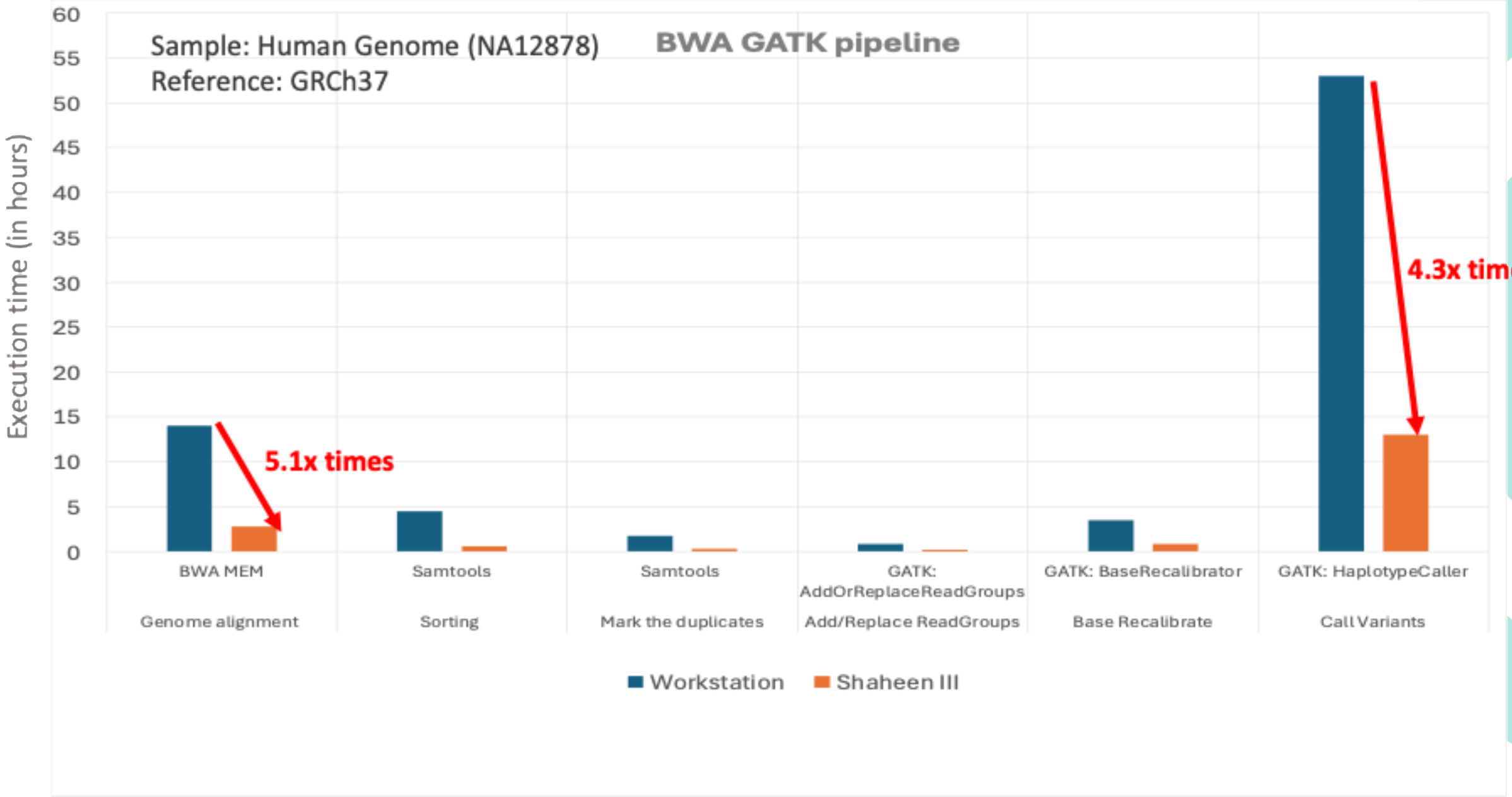
--------- /scratch/project/software/ex111genoa/modulefiles ----------

bio/blast+/2.15.0 bio/gatk/4.1.6 bio/java/21.0.2 bio/plink/2.0a bio/samtools/1.18
bio/bwa/0.7.17 bio/java/8u401 bio/plink/1.9b bio/samtools/1.8
bio/tabix/0.2.6

Example job script

```
#!/bin/bash
#SBATCH -N 1
#SBATCH -J bwa_mem
#SBATCH --time=23:00:00
#SBATCH -o genome_alignment.%j_%N.out
#SBATCH -e genome_alignment.%j_%N.err
### Input
export SAMPLE=NA12878
export INPUT=/scratch/kathirn/work/Shaheen3_101/input;
export OUTPUT=/scratch/kathirn/work/Shaheen3_101/output;
export REFERENCE=/scratch/kathirn/work/Shaheen3_101/ref;
## Software environment
export
MODULEPATH=/scratch/project/software/ex111genoa/modulefiles:$MODULEPATH
module load bio/bwa/0.7.17 bio/samtools/1.18
## Genome alignment
time -p bwa mem -t 16 $REFERENCE/GRCh37.fasta $INPUT/${SAMPLE}_1.fastq.gz
$INPUT/${SAMPLE}_2.fastq.gz | samtools view -@ $CORES -b -S -h -q 30 - |
samtools sort - > $SAMPLE.sorted_aligned_reads.bam
```

Example: Performance optimization



Thank you very much for your attention and valuable feedback!