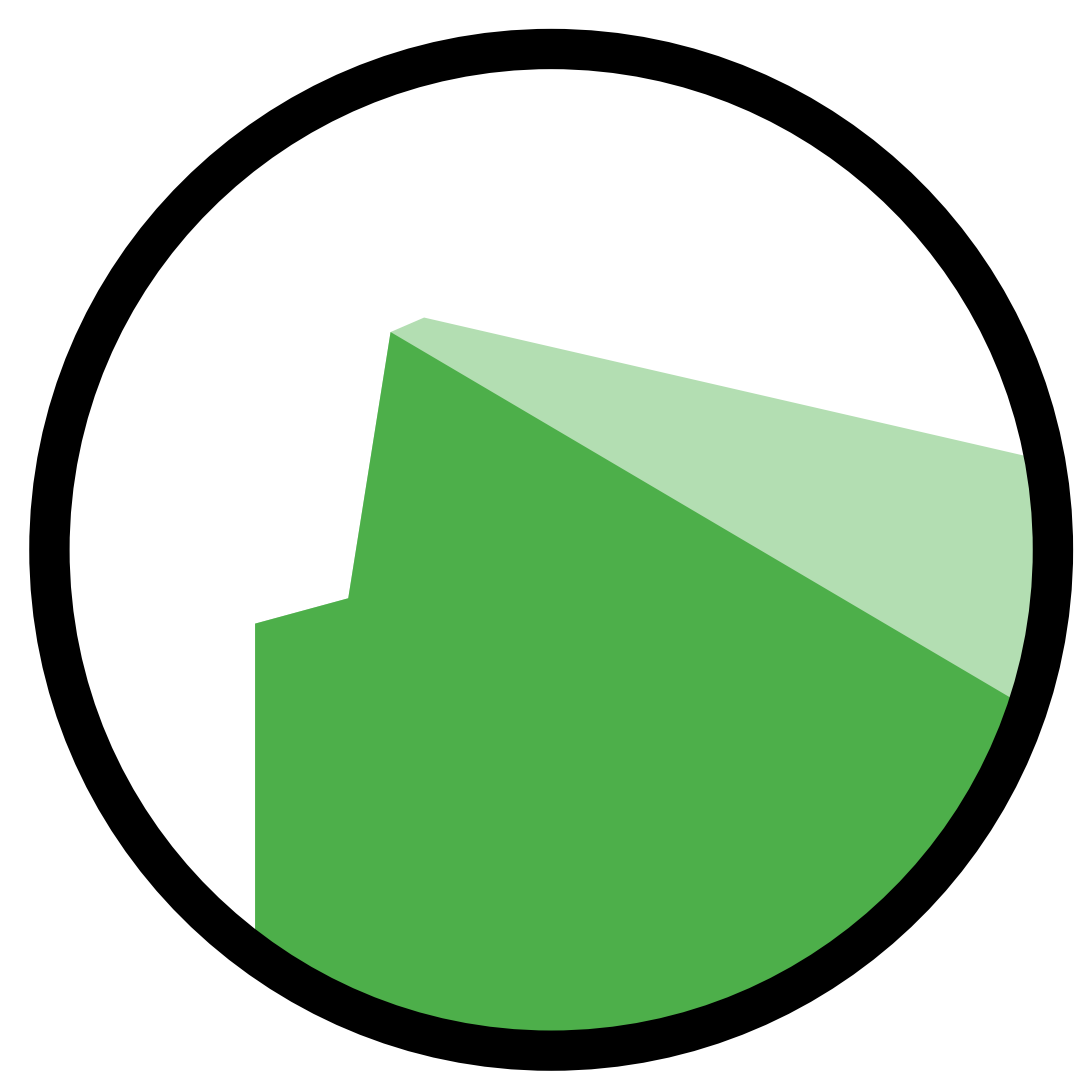


This work was supported by the AWS Cloud Credits for Research program in the form of a computational credit grant received by Philip A Ewels.

1. School of Engineering Sciences in Chemistry, Biotechnology and Health, Royal Institute of Technology KTH, Sweden
2. Department of Oncology, Karolinska Institute, Stockholm, Sweden
3. Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany
4. University of Southern Denmark, Odense, Denmark
5. A*STAR Genome Institute of Singapore, Bioinformatics Core Unit, Singapore, Singapore
6. Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden

* National Genomics Infrastructure Stockholm, Science for Life Laboratory



Sarek

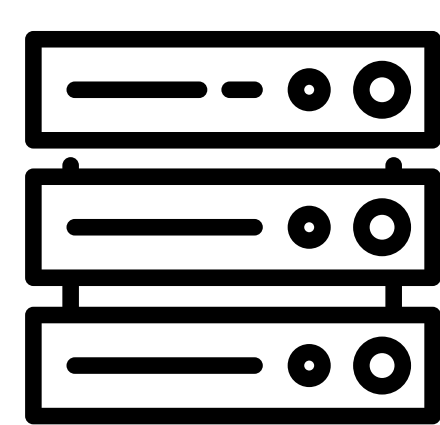
Sarek is a whole genome sequencing analysis pipeline implemented in NextFlow.

nextflow

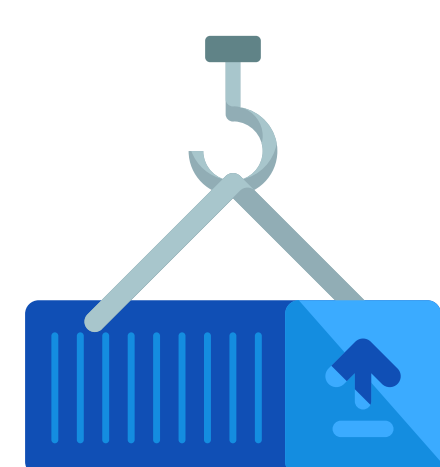
NextFlow has native support for major cloud providers, including AWS Batch.



AWS Batch executes batch jobs much like a HPC scheduler.



To scale storage space with instance size, m5d instances with physical ssd storage were used.

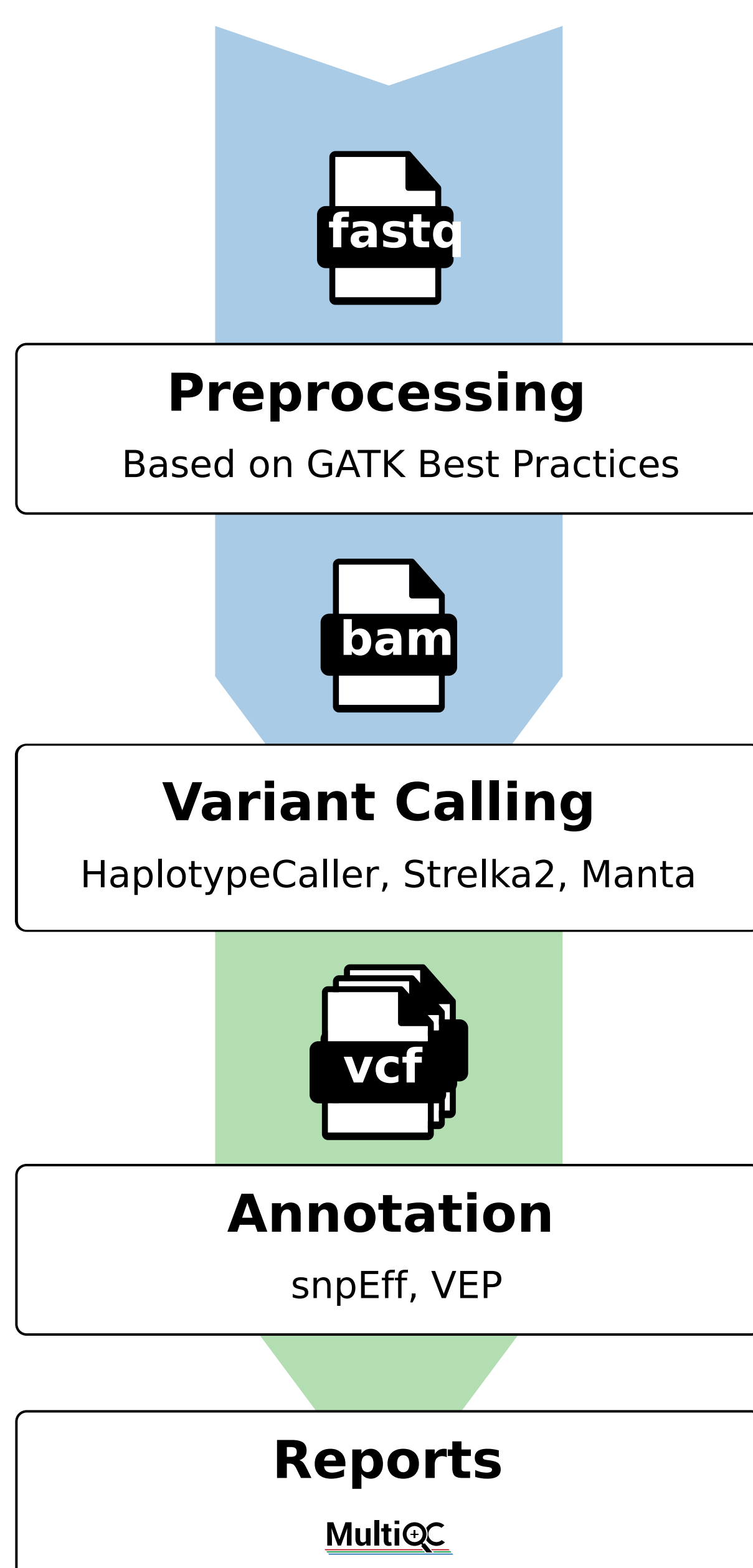


A container (Amazon machine image) was constructed to automatically mount the physical ssd drives.



The analysis finishes in roughly 30 hours.

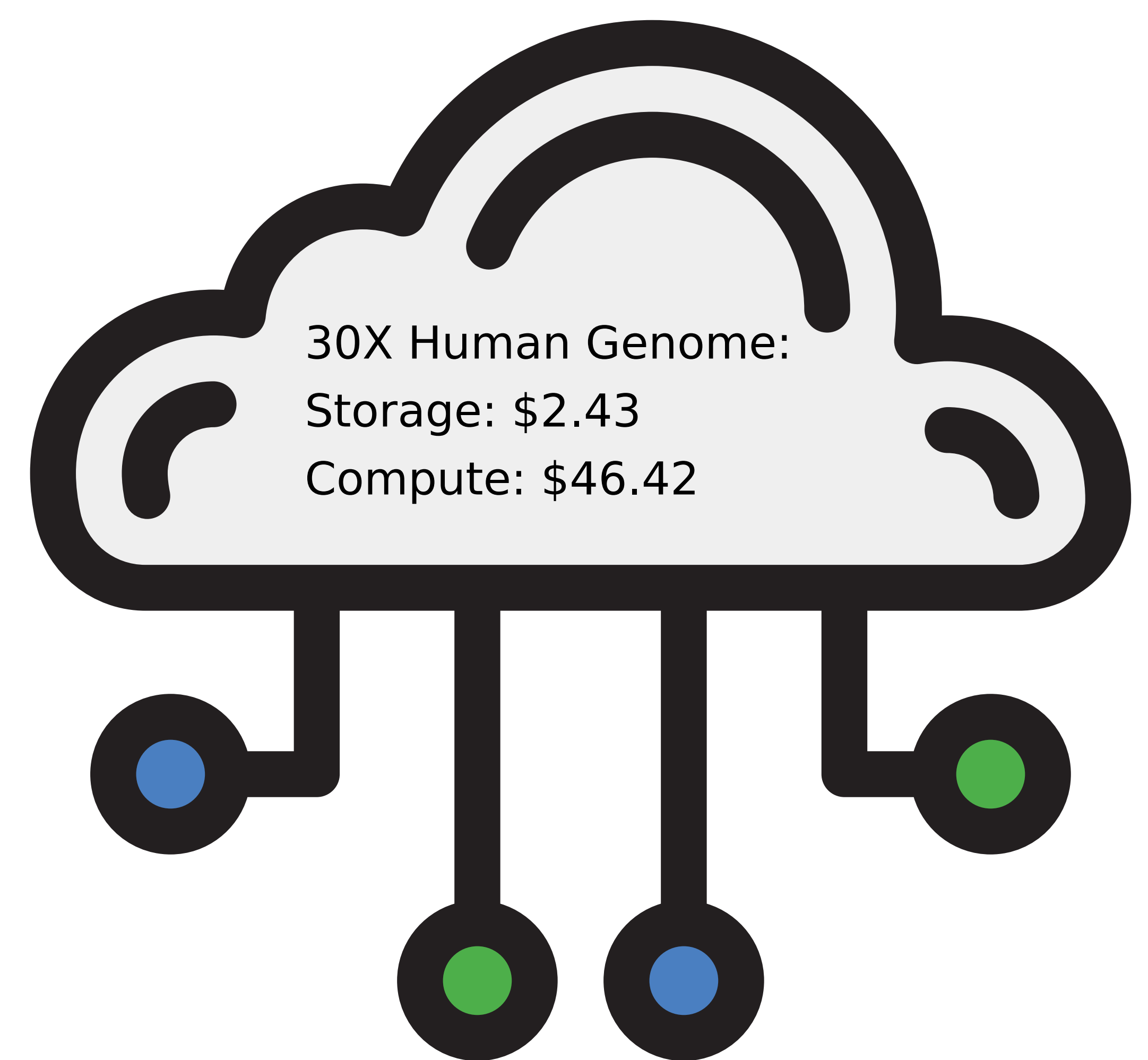
Analysis steps



Summary

While large high-performance computing might be already available to larger research groups, especially smaller research groups would greatly benefit from a publicly accessible commercial alternative.

Cloud computing offers a complete computational infrastructure without upfront infrastructure investment. Using Sarek on the AWS cloud infrastructure; mapping, germline variant calling and annotation for a human WGS sample (30X coverage) was performed for less than US-\$ 50.



Commands outline

```

$ aws batch create-compute-environment --region eu-west-1 \
--compute-environment-name $COMPUTE_ENV \
--compute-resources type=SPOT,instanceTypes="m5d",bidPercentage=50

$ aws batch create-job-queue --region eu-west-1 --job-queue-name $AWS_QUEUE \
--compute-environment-order "order=1,computeEnvironment=$COMPUTE_ENV"

$ nextflow run Sarek/main.nf -profile awsbatch --awsqueue $AWS_QUEUE
  
```

Caveat

Beware that storing sensitive data with US companies might be violating GDPR due to the "cloud act". If this is the case, other local cloud providers might be more appropriate.