**Implementing genomics workflows using docker**

This workflow uses multiple docker images that take \*.fastq files and a reference genome file, and create a \*.vcf file with a list of variants.

Here, I’ll use a test dataset and implement a basic genomics workflow to highlight the use of docker technology in DNA sequence analysis.

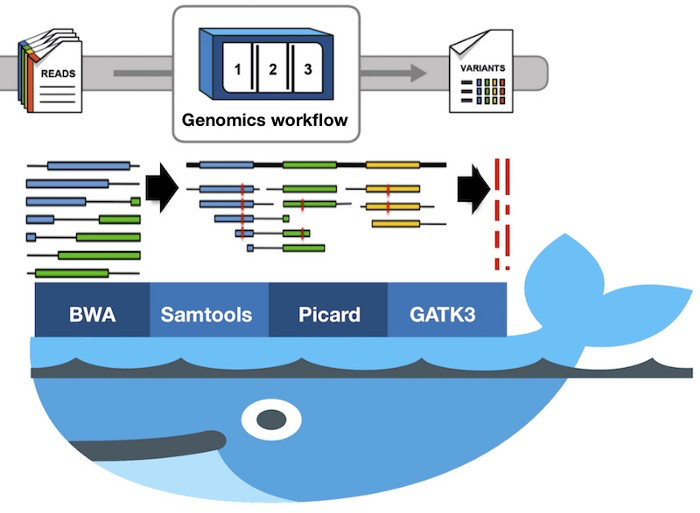


Figure 1. Sequence analysis and variant calling workflow using BWA, Samtools, Picard and GATK3 docker images

**Datasets:**

* Input

7859\_GPI.read1.fq and 7859\_GPI.read2.fq FASTQ files   
- [from GATK google drive folder — tutorial\_7859.tar.gz](https://drive.google.com/drive/folders/1aBcbV_Hlyg0wOOmZDDSBeIc0uw1r3f_w" \t "_blank)

* Reference sequence

chr19\_KI270866v1\_alt   
- [two contigs from human GRCh38/hg38: chr19](https://software.broadinstitute.org/gatk/documentation/article?id=7859" \t "_blank)

**Download docker images and initiate relevant containers**

* [BWA](http://bio-bwa.sourceforge.net/bwa.shtml" \t "_blank)

> sudo docker pull biocontainers/bwa:v0.7.17–3-deb\_cv1

> sudo docker pull biocontainers/bwa

* [Samtools](https://github.com/samtools/samtools" \t "_blank)

> sudo docker pull biocontainers/samtools:v1.7.0\_cv3

* [Picard](http://broadinstitute.github.io/picard/" \t "_blank)

> sudo docker pull biocontainers/picard:v2.3.0\_cv3

* [GATK](https://software.broadinstitute.org/gatk/" \t "_blank)

***Download latest GATK release***> sudo docker pull broadinstitute/gatk

***Download a specific GATK release***> sudo docker pull quay.io/biocontainers/gatk:3.6–7 ## specific GATK release

***Run and initiate the gatk\_3.6 image and follow GATK3 setup (download a licensed copy of GATK from the Broad Institute and copy GATK into your conda environment)***

#sudo docker run — name gatk\_3.6 -it -v <host directory>:/data <IMAGE ID>  
sudo docker container exec -it gatk\_3.6 bash ## Follow GATK3 setup

**Run basic steps in a sequence analysis workflow using docker**

* Indexing the reference genome

***Initiate bwa container... and***

> sudo docker run — name bwa\_v0.7.17 -it -v <host directory>:/data <IMAGE ID> bash

***Index the reference***

> bwa index chr19\_KI270866v1\_alt.fasta  
 *# Create* BWA index with  
  *\*fasta.amb,  
 \*fasta.ann,  
 \*fasta.bwt,   
 \*fasta.pac,   
 \*fasta.sa files*

***Create \*fasta.fai and \*dict indexes for GATK analysis***

>sudo docker run — name samtools\_v1.7 -it -v <host directory>:/data samtools faidx chr19\_KI270866v1\_alt.fasta

*> sudo docker run -v* <host directory>:/data --*rm <IMAGE ID> picard CreateSequenceDictionary R=chr19\_KI270866v1\_alt.fasta O=chr19\_KI270866v1\_alt.dict*

* Read-mapping using BWA mem
* ***Execute BWA container... and***

> sudo docker container exec -it bwa\_v0.7.17 bash***Align \*fq files to the reference***> bwa mem -R '@RG\tID:t1\tSM:t1' chr19\_KI270866v1\_alt.fasta 7859\_GPI.read1.fq 7859\_GPI.read2.fq > 7859\_GPI.aln\_pe.sam  
 # Created 7859\_GPI.aln\_pe.sam file

* Sorting SAM file, marking duplicated reads and indexing BAM files
* ***Sort sam file and create a bam file, then index the bam file***

>sudo docker run -v <host directory>:/data --rm <IMAGE *ID>* picard SortSam INPUT=7859\_GPI.aln\_pe.sam OUTPUT=7859\_GPI.aln\_pe.bam SORT\_ORDER=coordinate

> sudo docker run -v <host directory>:/data --rm *<IMAGE ID>* picard BuildBamIndex VALIDATION\_STRINGENCY=LENIENT I=7859\_GPI.aln\_pe.bamMark duplicated reads in the indexed bam file and create a new marked-bam file> docker run -v <host directory>:/data --rm <IMAGE ID> picard MarkDuplicates VALIDATION\_STRINGENCY=LENIENT AS=true REMOVE\_DUPLICATES=true I=7859\_GPI.aln\_pe.bam O=7859\_GPI.aln\_pe.md.bam M=7859\_GPI.aln\_pe.md.metricsIndex the marked-bam file> docker run -v <host directory>:/data --rm <IMAGE ID> picard BuildBamIndex VALIDATION\_STRINGENCY=LENIENT I=7859\_GPI.aln\_pe.md.bam

* GATK3 analysis and calling variants

***Execute GATK3 container... and***

> sudo docker container exec -it gatk\_3.6 bash***Define intervals to target for local realignment***> java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator \  
-nt 4 -R chr19\_KI270866v1\_alt.fasta -I 7859\_GPI.aln\_pe.md.bam \  
-o 7859\_GPI.aln\_pe.intervals***Perform local realignment of reads around indels***> java -jar GenomeAnalysisTK.jar -T IndelRealigner \  
-R chr19\_KI270866v1\_alt.fasta -I 7859\_GPI.aln\_pe.md.bam \  
-targetIntervals 7859\_GPI.aln\_pe.intervals -o 7859\_GPI.aln\_pe.md.bam.realigned.bam***Call variants using HaplotypeCaller***> java -jar GenomeAnalysisTK.jar -T HaplotypeCaller \  
-R chr19\_KI270866v1\_alt.fasta -ERC GVCF \  
-I 7859\_GPI.aln\_pe.md.bam.realigned.bam -o 7859\_GPI.aln\_pe.md.bam.realigned.g.vcf