Calling Variants using the OpenCGA command line

As OpenCGA wraps the packages Samtools and GATK, users can easily call variants from their alignment files, i.e., BAM files. This tutorial details how to use the command line to the Samtools and GATK wrappers.

Prerequisites

A working setup of OpenCGA is required to setup a testing environment, please follow the steps on installation guide.

In addition, you need to download the following data files:

- Reference genome: ref.fasta
- Input BAM file: mother.bam

Variant Calling

Before calling variants you need to prepare your reference genome (FASTA) file and your input alignment (BAM) file. So, you need to index your FASTA file and create the sequence dictionary file for that FASTA file. In addition, you need to index your BAM file. Next sections show the OpenCGA command lines to do it.

Preparing the reference genome: index and dictionary files

In order to use the *ref.fasta* file, it has to be linked to the OpenCGA catalog:

```
$ ./opencga.sh files link -i ~/ref.fasta --path call/ --parents
```

Once linked the Fasta file, you need to index it by running the Samtools wrapper with the command *faidx*:

```
$ ./opencga.sh alignments samtools-run --command faidx --input-file ref.
fasta
```

The FASTA index file (*ref.fasta.fai)* will be created in the folder of the input FASTA file (*ref.fasta*). You can retrieve information about the created FASTA index file by running the following command:

```
$ ./opencga.sh files info --file ref.fasta.fai
```

And you need to create the sequence dictionary file for that FASTA file, again you run the Samtools wrapper with the command *dict*. The sequence dictionary output file (*ref.dict*) will be created in the folder of the input FASTA file (*ref.fasta*):

```
$ ./opencga.sh alignments samtools-run --command dict --input-file ref.
fasta --output-filename ref.dict
```

Preparing the alignment file: sort and index BAM file

In order to use the *mother.bam* file, it has to be linked to the OpenCGA catalog:

```
$ ./opencga.sh files link -i ~/mother.bam --path call/ --parents
```

Then you need to sort the BAM file, you run the Samtools wrapper with the command *sort*. The sorted BAM file (*mother.sorted.bam*) will be created in the folder of the input BAM file (*mother.bam*):

```
$ ./opencga.sh alignments samtools-run --command sort --input-file mother.
bam --output-filename mother.sorted.bam
```

Once sorted the BAM file, you can index it by running the Samtools wrapper with the command *index.* The index file (*mother.sorted.bam.bai*) will be created in the folder of the input BAM file (*mother.sorted.bam*):

```
$ ./opencga.sh alignments samtools-run --command index --input-file mother.
sorted.bam
```

Variant calling

You can call variants by running the Gatk wrapper with the command *HaplotypeCaller*:

```
$ ./opencga.sh variant gatk-run --command HaplotypeCaller --fasta-file ref.
fasta --bam-file mother.sorted.bam --vcf-filename mother.vcf
```

The variant calls are saved in the VCF file *mother.vcf* that can be downloaded from the OpenCGA catalog to the local directory */tmp* by using the following command:

```
$ ./opencga.sh files download --file mother.vcf --to /tmp
```

Here you can view the *mother.vcf* file.