Working with Alignment Data

This tutorial details how to use the OpenCGA alignment command line to run the alignment/mapping pipeline steps. The alignment pipeline outputs alignments in BAM files from raw sequence data in FastQ format files. BAM files can be used for further analysis, such as alignment statistics, coverage computation or variant calling.

Table of Contents:

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:

- Raw sequence data file: input.fastq
- Reference genome: reference.fasta

The alignment pipeline

Quality control for raw sequence data: FastQC subcommand

In order to use the input.fastq file, it has to be linked to the OpenCGA catalog:

```
$ ./opencga.sh files link -i ~/input.fastq --path test/ --parents
```

Once linked the FastQ file, you can run the FastQC command:

```
$ ./opencga.sh alignments fastqc-run --file input.fastq
```

For the *input.fastq* file, the FastQC command creates a report file called *input_fastqc.html* that can be downloaded from the OpenCGA catalog to the local directory */tmp* by using the following command:

```
$ ./opencga.sh files download --file input_fastqc.html --to /tmp
```

Here is the FastQC report file: input_fastqc.html.

Mapping raw sequences: BWA subcommand

First, link the reference.fasta file to the OpenCGA catalog:

```
$ ./opencga.sh files link -i ~/reference.fasta --path test/ --parents
```

Then, you can run the *bwa index* command to index database sequences in the FASTA format:

```
$ ./opencga.sh alignments bwa-run --command index --fasta-file reference.
fata
```

Internally, the index for the *reference.fasta* file created by the *bwa index* command consists of the following files:

```
reference.fasta.bwt
reference.fasta.pac
reference.fasta.ann
reference.fasta.amb
reference.fasta.sa
```

Once created the index, you can map the FastQ file by using the *bwa mem* command:

```
$ ./opencga.sh alignments bwa-run --command mem --index-base-file
reference.fasta --fastq1-file input.fastq --sam-file output.sam
```

In the previous command, the result alignments are saved in SAM format in the *output.sam* file .

Converting to and sorting BAM files

In order to convert a SAM file into BAM file, use the *samtools view* command with the parameter *b* (for more parameters details, see http://www.htslib.org/doc/samtools-view.1.html):

```
$ ./opencga.sh alignments samtools-run --input-file alignments.sam --
output-filename alignments.bam --command view -Db=null
```

To sort a BAM file, use the *samtools sort* command:

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

Indexing and querying BAM files

Once the BAM file is sorted, you can index. A BAM index consists of a BAI file. Use the following command to create the index file (it will be called *alignments.sorted.bam.bai*):

```
$ ./opencga.sh alignments index --file alignments.sorted.bam
```

To query for alignments use the *query* command applying a set of filters (e.g., regions, insert size, maximum number of hits of mismatches, minimum mapping quality...). The following command query alignments for a given region in chromosome 20 and from the position 500000 to 1000000:

```
$ ./opencga.sh alignments query --file alignments.sorted.bam --region 20:
500000-1000000  --study study1 --rpc REST
```

Computing and querying BAM coverage

In order to compute the BAM coverage, use the command coverage-run. The coverage is saved in a BigWig format file.

```
$ ./opencga.sh alignments coverage-run --file alignments.sorted.bam --
window-size 50
```

To query for the coverage in a given region use the command *coverage-query*. The following command query for the coverage from the position 78700 to 78900 in chromosome 20:

```
$ ./opencga.sh alignments coverage-query --file alignments.sorted.bam --
region 20:78700-78900 --study study1
```

Computing BAM statistics

In order to compute the statistics for a given BAM file, use the command *stats-run:*

```
$ ./opencga.sh alignments stats-run --file alignments.sorted.bam
```

BAM statistics can be viewed by executing the command *stats-info:*

```
$ ./opencga.sh alignments stats-info --file alignments.sorted.bam
```

In addition, OpenCGA provides the command *stats-query* to fetch BAM files according to their statistics. The following command fetch BAM files whose average mapping quality is greater than 35:

```
$ ./opencga.sh alignments stats-query --average-quality ">25"
```