Alignment Engine

Table of Contents:

OpenCGA Alignment Engine provides a solution to storage and process sequence alignment data from Next-Generation Sequencing (NGS) projects. The Alignment Engine supports the most common alignment file formats, i.e.: SAM, BAM and CRAM, and takes the alignment data model specification from **GA4GH** and the implementation from **OpenCB GA4GH**. See a full description at Alignment Data Model.

We do not define or endorse any dedicated unaligned sequence data format. Instead we recommend storing such data in one of the alignment formats (SAM, BAM, or CRAM) with the unmapped flag set.

OpenCGA alignment engine provides the following analysis:

- **Index analysis**
  To index a coordinate-sorted alignment file (BAM or CRAM) for fast random access. This index is needed when region parameters are used to limit the query analysis to particular regions of interest.

- **Query analysis**
  This analysis outputs those alignments matching the specified filters, such as minimum mapping quality, maximum insert size, maximum number of mismatches in the alignment, properly paired alignments,... In addition, users may specify one or more comma-separated regions to restrict output to only those alignments which overlap the specified region(s). Note that use of region specifications requires a coordinate-sorted and indexed input file (in BAM or CRAM format).

- **Coverage analysis**
  This analysis takes a coordinate-sorted and indexed alignment file (in BAM or CRAM format) as input and generates a coverage file (in BigWig format). The coverage is calculated as the number of reads per window of a user-defined size, if window size is equal to 1, the coverage is the number of reads per position. Once coverage is computed, the read coverage over multiple genomic regions can be fetched quite quickly.

- **Statistics analysis**
  OpenCGA computes statistics for a given alignment file by using the samtools stats command. Alignment statistics are indexed in order to allow users to query for alignment files according to those statistics.

In addtion, OpenCGA provides wrappers to the following third-party alignment software packages:

- **FastQC**: a quality control tool for high throughput sequence data.
- **BWA**: a software package for mapping low-divergent sequences against a large reference genome.
- **Samtools**: a program for interacting with high-throughput sequencing data in SAM, BAM and CRAM formats.
- **deepTools**: a suite of python tools particularly developed for the efficient analysis of high-throughput sequencing data, such as ChIP-seq, RNA-seq or MNase-seq.

OpenCGA Alignment User Interfaces

OpenCGA provides two interfaces to allow users execute the alignment tools and analysis:

- Command line inteface
- RESTful web services interface

OpenCGA command line interface

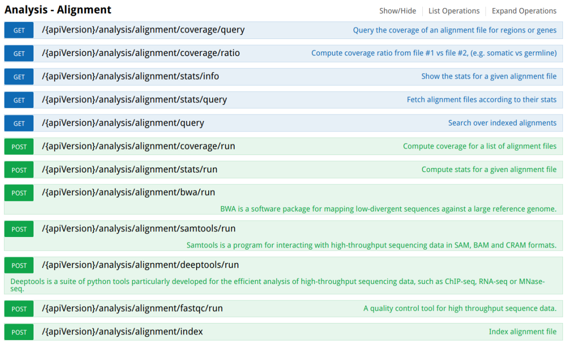
The OpenCGA command line interface to manage alignment data is accessible through the script *opencga.sh* using the command *alignments*:

```
$ ./opencga.sh alignments

Usage:   opencga.sh alignments <subcommand> [options]

Subcommands:
         index  Index alignment file
         query  Search over indexed alignments
     stats-run  Compute stats for a given alignment file
    stats-info  Retrieve stats for a given alignment file
   stats-query  Fetch alignment files according to their stats
  coverage-run  Compute coverage for a given alignemnt file
coverage-query  Query the coverage of an alignment file for regions or
genes
coverage-ratio  Compute coverage ratio from file #1 vs file #2, (e.g.
somatic vs germline)
           bwa  BWA is a software package for mapping low-divergent
sequences against a large reference genome.
      samtools  Samtools is a program for interacting with high-throughput
sequencing data in SAM, BAM and CRAM formats.
     deeptools  Deeptools is a suite of python tools particularly
developed for the efficient analysis of high-throughput sequencing data,
such as ChIP-seq, RNA-seq or MNase-seq.
        fastqc  A quality control tool for high throughput sequence data.
```

The tutorial Working with Alignment Data shows how to use the OpenCGA alignment commandline.

OpenCGA RESTful web services interface

Next image shows the OpenCGA RESTfull web services to manage alignment data:



Analysis - Alignment

GET	/{apiVersion}/analysis/alignment/coverage/query	Query the coverage of an alignment file for regions or genes
GET	/{apiVersion}/analysis/alignment/coverage/ratio	Compute coverage ratio from file #1 vs file #2, (e.g. somatic vs germline)
GET	/{apiVersion}/analysis/alignment/stats/info	Show the stats for a given alignment file
GET	/{apiVersion}/analysis/alignment/stats/query	Fetch alignment files according to their stats
GET	/{apiVersion}/analysis/alignment/query	Search over indexed alignments
POST	/{apiVersion}/analysis/alignment/coverage/run	Compute coverage for a list of alignment files
POST	/{apiVersion}/analysis/alignment/stats/run	Compute stats for a given alignment file
POST	/{apiVersion}/analysis/alignment/bwa/run	BWA is a software package for mapping low-divergent sequences against a large reference genome.
POST	/{apiVersion}/analysis/alignment/samtools/run	Samtools is a program for interacting with high-throughput sequencing data in SAM, BAM and CRAM formats.
POST	/{apiVersion}/analysis/alignment/deeptools/run	Deeptools is a suite of python tools particularly developed for the efficient analysis of high-throughput sequencing data, such as ChIP-seq, RNA-seq or MNase-seq.
POST	/{apiVersion}/analysis/alignment/fastqc/run	A quality control tool for high throughput sequence data.
POST	/{apiVersion}/analysis/alignment/index	Index alignment file