Sentieon Quick Start

Release 201808.07

Sentieon, Inc

Jul 24, 2019

Contents

This document serves as a guide to get first-time users introduced to the Sentieon DNAseq software. For more detailed information on the software and/or its deployment in a production environment, please refer to the official software documentation. If you have any additional questions, please contact the technical support at Sentieon Inc. at support@sentieon.com.

1 What do you need to get started?

To get started using Sentieon DNAseq software, you will need the following:

1. Hardware requirements: A Linux server or a Mac computer with the following configuration:

- Linux running one of the following distributions or higher: RedHat/CentOS 6.5, Debian 7.7, OpenSUSE-13.2, or Ubuntu-14.04.
- Or a Mac running OsX 10.9 (Mavericks) or higher.
- 16GB of memory for small panel or whole exome or 64GB for whole genome.
- (Recommended) High-speed SSD drives are preferred for ideal I/O performance to get maximum CPU utilization.

2. Software requirements:

- Python 2.6.x or Python 2.7.x is required. You can check Python version by typing the following:

```
python --version
```

3. Sentieon DNAseq software release package:

- Download the package from the link provided by the technical support at Sentieon.
- Decompress the package by running the following command, where VERSION is the version you are using, for example 201808.07:

```
tar xvzf sentieon-genomics-VERSION.tar.gz
```

4. License requirements: Please see the Appendix for more details on how to set up your license. IT support may be needed.
5. Environment requirements:
   - If Python 2.6.x or Python 2.7.x is not the default Python version, you can set the following environment variable.

```
export SENTIEON_PYTHON=Python_2_X_location
```

   - If you are using a localhost license file, set the following environment variable, where LICENSE_DIR is where the license file is located, and LICENSE_FILE.lic is the license file name.

```
export SENTIEON_LICENSE=LICENSE_DIR/LICENSE_FILE.lic
```

   - If user is using a license server, set the following environmental variable, where LICSRVR_HOST and LICSRVR_PORT are the hostname and port of the license server. Please see the next section for more details.

```
export SENTIEON_LICENSE=LICSRVR_HOST:LICSRVR_PORT
```

   - For convenience, set the binary path as shown beloww, where PATH_TO_ SENTIEON_BINARY_DIRECTORY is where Sentieon binary is installed.

```
export SENTIEON_INSTALL_DIR=PATH_TO_SENTIEON_BINARY_DIRECTORY
```

   - For improved performance when using NFS storage, set the SENTIEON_TMPDIR environmental variable to point to local scratch fast storage.

```
export SENTIEON_TMPDIR=/tmp
```

   - If you are running the Sentieon DNAseq software in a MacBook laptop, we recommend that you disable the energy saving mode while running jobs, so that the computer does not go to sleep. You can disable the energy saving mode by running the following in a Terminal:

```
pmset noidle
```

2 Start your first Sentieon DNAseq job

Sentieon Inc. provides a quick start package that includes a sample script and data to help you quickly test the installation and to diagnose potential problems.

The quick start package includes data for a single chromosome, both sequence data of a sample and reference materials. The job script follows the Broad Institute's BWA-GATK HaplotypeCaller 3.x Best Practice Workflow pipeline for a set of pair-ended Illumina fastq files:

- BWA: Map reads to the reference.

- Metrics: Collect reads' statistics.
- Dedup: Remove duplicate reads.
- Indel re-aligner.
- BSQR: Re-calibrate base quality score.
- Variant caller: HaplotypeCaller variant caller.

2.1 Run the quick start package

To get started, copy the downloaded quick start package to a new directory, and unpack it by running the following:

```
tar xzvf quick_start.tar.gz
```

Here is what is included in the package:

- `sentieon_quickstart.sh`: the sample shell script that drives the entire pipeline.
- `reference`: a directory that contains human genome reference files and database files of known SNP sites.
- FASTQ files: sample sequence files.

Before running the script, you need to make sure that the environment variables are properly set as described above, including the license and path to the directory.

Then open your favorite editor to edit the user settings in `sentieon_quickstart.sh`.

```
# Update with the location of the Sentieon software package
SENTIEON_INSTALL_DIR=/home/release/sentieon-genomics-201808.07

# Update with the location of temporary fast storage and uncomment
#SENTIEON_TMPDIR=/tmp

# It is important to assign meaningful names in actual cases.
# It is particularly important to assign different read group names.
sample="sample_name"
group="read_group_name"
platform="ILLUMINA"

# Other settings
nt=16 #number of threads to use in computation
```

Note

In the user setting shell script `sentieon_quickstart.sh`:

- It is important to assign meaningful names in actual cases.
- It is particularly important to assign different read group names.

To get the number of the CPU cores, user can run nproc as shown below.

```
nproc
```

To better understand the rest of the `sentieon_quickstart.sh` script, please read the comment in each section, and the corresponding chapters in the manual.

Now, launch the script by simply running `sentieon_quickstart.sh`, and watch the result unfold. The entire run takes about 3 - 5 minutes on a typical Linux server. Actual time varies depending on the computation environment.

```
sh sentieon_quickstart.sh &
```

2.2 Understanding the results

Below is a list of the files, their meaning and references. For more details, please refer to documentation.

1. Quick start test output files

File name	Description
sorted.bam	Coordinate-sorted BAM file after alignment with Sentieon BWA mem.
aln_metrics.txt	Alignment and general statistics of the two pair sequence reads.
gc_summary.txt	GC bias statistics summary.
gc_metrics.txt, qc-report.pdf	GC bias statistics data file and report PDF.
qd_metrics.txt, qd-report.pdf	Base quality score distribution data file and report PDF.
mq_metrics.txt, mq-report.pdf	Cycle-dependence of the mean quality score data file and report PDF.
is_metrics.txt, is-report.pdf	Insert size distribution data file and report PDF.
deduped.bam	Output BAM file of Dedup stage, with duplicated reads removed.
recal_data.table	Calibration data table.
recal_plots.pdf	BQSR report PDF.
output-hc.vcf.gz	Output VCF file of haplotype variant caller stage.

3 Description of the Sentieon tools

The table below shows the different Sentieon products and tools and their purpose. It is also noted if a tool implements functionality equivalent to an existing GATK pipeline tool.

Sentieon product	Sentieon tool	Typical use	Equivalent G
Sentieon BWA	Sentieon BWA	Read alignment and mapping	BWA
DNAseq	Genotyper	Germline SNV/Indel calling, non haplotype based	UnifiedGeno
DNAseq	Haplotyper	Germline SNV/Indel calling	HaplotypeC
DNAseq	GVCFtyper	Joint calling of cohorts, demonstrated up to 200,000 samples	GenotypeGV
DNAseq	VarCal	Calculate Variant Quality Score Recalibration	VariantReca
DNAseq	ApplyVarCal	Apply Variant Quality Score Recalibration	ApplyRecali
DNAscope	DNAscope	Improved germline SNV/Indel/SV calling	
RNAseq	RNASplitReadsAtJunction	RNA SNV/Indel calling	SplitNCigarR
RNAseq	Haplotyper	RNA SNV/Indel calling	HaplotypeC
TNseq	TNsnv	Somatic SNV calling, non haplotype based	MuTect
TNseq	TNhaplotyper	Somatic SNV/Indel calling	MuTect2
TNseq	TNhaplotyper2 + tnhapfilter	Somatic SNV/Indel calling	Mutect2 and
TNscope	TNscope	Improved somatic SNV/Indel/SV calling	
General tools	Dedup and LocusCollector	Perform deduplication	Picard Mark
General tools	Realigner	Perform Indel relaignment for non-haplotype based callers	RealignerTa
General tools	QualCal	Perform Base Quality Score Recalibration	BaseRecalib
General tools	ReadWriter	Create BAM files	PrintReads
General tools	AlignmentStat	QC metrics	Picard Colle
General tools	BaseDistributionByCycle	QC metrics	Picard Colle
General tools	CollectVCMetrics	QC metrics	Picard Colle
General tools	ContaminationAssessment	QC metrics	ContEst
General tools	CoverageMetrics	QC metrics	DepthOfCov
General tools	GCBias	QC metrics	Picard Colle
General tools	HsMetricAlgo	QC metrics	Picard Colle
General tools	InsertSizeMetricAlgo	QC metrics	Picard Colle

Table 3.1 – continued from previous page

Sentieon product	Sentieon tool	Typical use	Equivalent C
General tools	MeanQualityByCycle	QC metrics	Picard Mean
General tools	QualDistribution	QC metrics	Picard Quali
General tools	QualityYield	QC metrics	Picard Colle
General tools	SequenceArtifactMetricsAlgo	QC metrics	Picard Colle
General tools	WgsMetricsAlgo	QC metrics	Picard Colle

4 Appendix - Set up license

Sentieon DNAseq software is a license-controlled software. The user is required to properly set up the license in order to run the software.

We provide two types of the licenses:

- Single machine evaluation license: this license is used for evaluating the Sentieon DNAseq software in a single machine. It allows new users to get quickly started on using the software without requiring help from the IT department. In order to use this license, the computer where you plan on running the Sentieon DNAseq software requires external Internet access.
- Cluster license: this license is used in a cluster environment. With this license, a floating license server lightweight process is running on one node in the cluster, serving licenses though TCP to all other nodes that have network connection to the license server. This license server is running in a special non-computing node on the cluster periphery that has unrestricted access to the outside world through HTTPS, and serves the licenses to the rest of the nodes in the cluster by listening to a specific TCP port that needs to be open within the cluster.

4.1 Setting up a single machine evaluation license

To use the single machine evaluation license, the computing node needs have access to the Internet. This allows Sentieon software to validate the license.

To use a single machine evaluation license, follow the steps below:

1. Copy the license file to the computing node. For example, the license file **LICENSE_FILE.lic** is now located at **LICENSE_DIR** .
2. Set up environment variable as below:

```
export SENTIEON_LICENSE=LICENSE_DIR/LICENSE_FILE.lic
```

4.2 Setting up license server

As shown in Fig. 4.1, license server requires the following:

1. The license server should have access to the Internet to perform license validation.
2. The computing nodes should have access to the license server via a host name **LICSRVR_HOST**
3. The machine the license server is running has an open port for the license services to listen on, and the computing nodes have access to that port. Here we assume the available port is **LICSRVR_PORT**

You may need IT support to get **LICSRVR_HOST:LICSRVR_PORT**, and confirm that the above requirements are met.

Note: If the license server is behind a firewall, separated from the computing nodes through a NAT, the license server's hostname/IP visible to the nodes may be different from its actual hostname/IP. If this is the case, you will need to bind

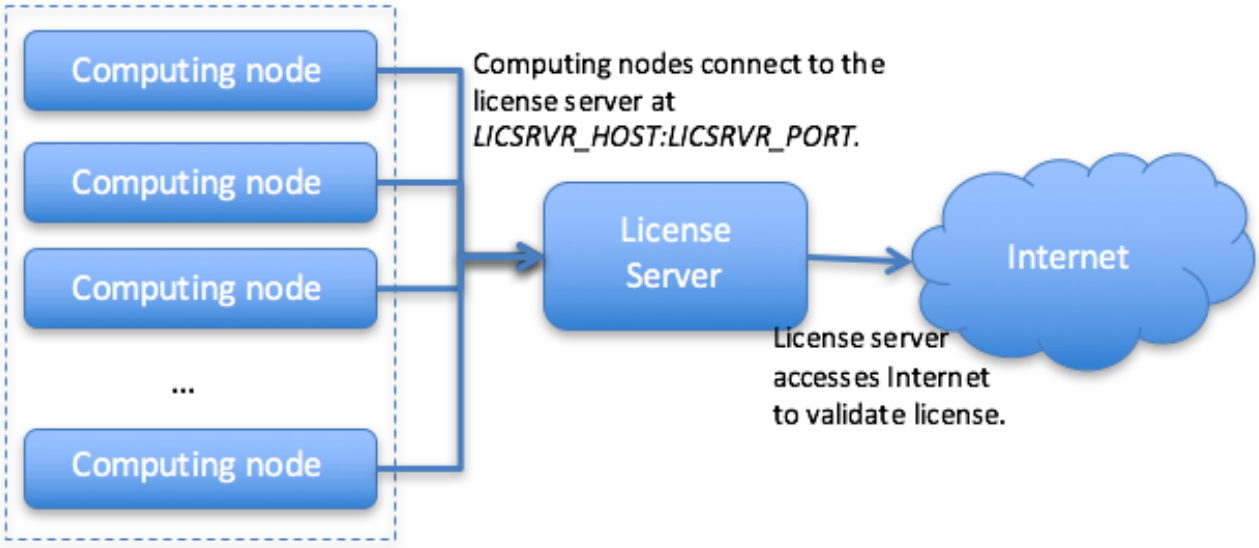

Fig. 4.1: Topology of the computing nodes and license server

the license server on the actual IP address, while the compute node requests license from the IP address after NAT. Please contact Sentieon support for more details.

Follow these steps to obtain license file, set up and test the license server:

1. Send the following information to Sentieon to receive the license file:

- FQDN (fully qualified domain name) **LICSRVR_HOST** of the designated machine to run license service.
- The designated port **LICSRVR_PORT** to Sentieon to receive the license file.

2. Copy the received license file to the license server LICSRVR_HOST. We assume the license file is located in **LICENSE_PATH/LIENSE_FILE**. Run the following command *on the license server* to start the license server process:

```
<SENTIEON_INSTALL_DIR>/bin/sentieon licsrvr --start --log LOG_FILE LICENSE_PATH/LIENSE_FILE
```

3. Alternatively, you can follow the instructions in section *8.5 - Running the license server (LICSRVR) as a system service* in the Sentieon Genomics Manual, to configure and start the license server as a system daemon.
4. Go to the Sentieon installation directory. Run the following commands *on the license server* to confirm the license server is up and running.

```
<SENTIEON_INSTALL_DIR>/bin/sentieon licclnt ping s LICSRVR_HOST:LICSRVR_PORT
```

   If the command returns without an error message, the license server is up and running.
5. Login to one of the computing node, go to the Sentieon installation directory, and run the above command again:

```
<SENTIEON_INSTALL_DIR>/bin/sentieon licclnt ping s LICSRVR_HOST:LICSRVR_PORT
```

   If the command returns without an error message, the computing node now can access the license server, too.
6. Set up the following environment variable and you are good to go.

```
export SENTIEON_LICENSE=LICSRVR_HOST:LICSRVR_PORT
```

465 Fairchild Drive, Suite 135, Mountain View CA 94043
www.sentieon.com