1. SparkSeq

1.1 SparkSeq Dependencies 설치

1.1.1 Hadoop-BAM-6.1

다운로드 URL :
http://sourceforge.net/projects/hadoop-bam/files/?source=navbar

*** hadoop계정으로 진행**

① 다운로드 리스트에서 hadoop-bam-6.1.tar.gz 을 다운로드 받는다.

② 다운로드 받은 hadoop-bam-6.1.tar.gz 파일을 /BiO/hadoop/hadoop_tools 디렉터리에 Upload 한다.

③ 압축해제

```
# tar xvfz hadoop-bam-6.1.tar.gz
```

1.2 SparkSeq 설치

* hadoop계정으로 진행

① SparkSeq 다운로드

```
# git clone https://bitbucket.org/mwiewiorka/sparkseq.git
```

② 디렉토리 이동

```
# cd sparkseq
```

③ sparkseq-core/build.sbt 변경

```
# vi sparkseq-core/build.sbt
```

- 라인 15 수정

변경전 :

```
val DEFAULT_HADOOP_VERSION = "1.2.1"
```

변경후 :

val DEFAULT_HADOOP_VERSION = "2.6.0"

- 라인 21 수정

변경전 :

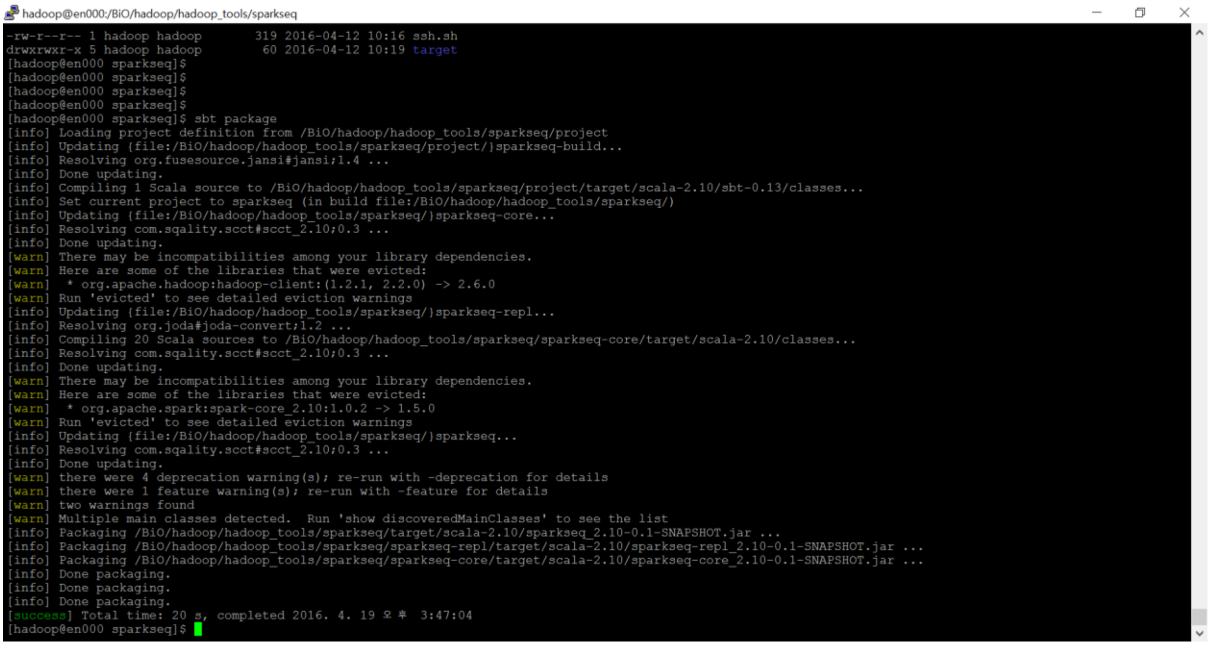
"org.apache.spark" %% "spark-core" % "1.1.0"

변경후 :

"org.apache.spark" %% "spark-core" % "1.5.0"

④ 컴파일 실행

# sbt package

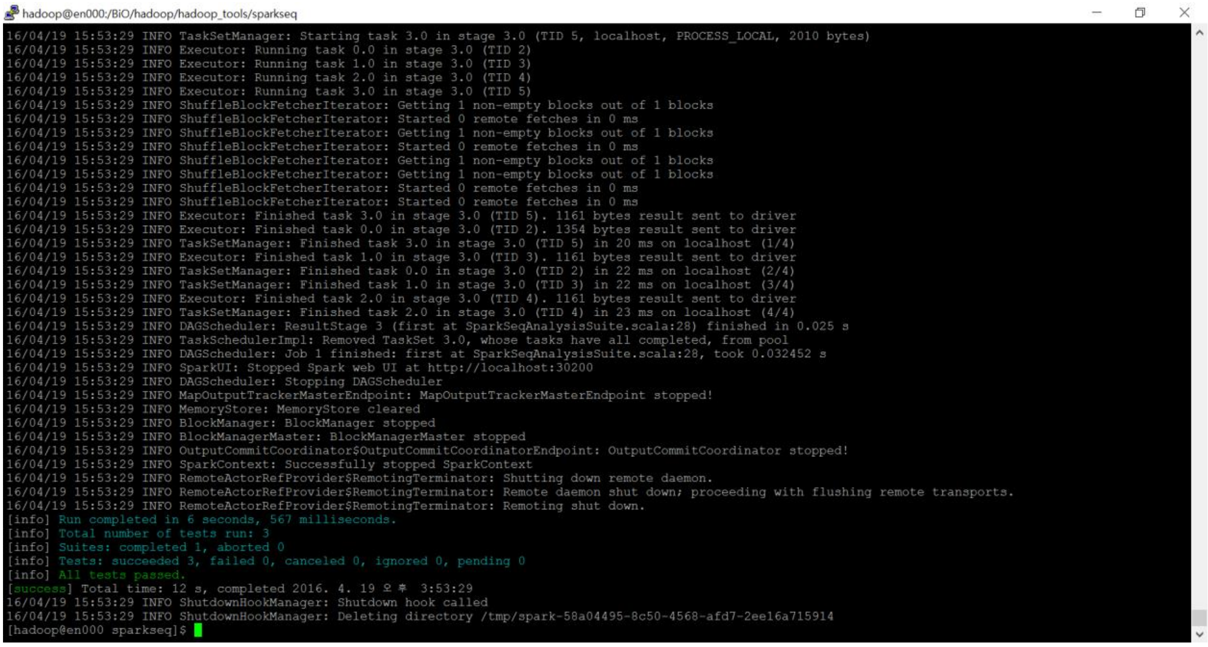


hadoop@en000:/BiO/hadoop/hadoop_tools/sparkseq

```
-rw-r--r-- 1 hadoop hadoop       319 2016-04-12 10:16 ssh.sh
drwxrwxr-x 5 hadoop hadoop        60 2016-04-12 10:19 target
[hadoop@en000 sparkseq]$
[hadoop@en000 sparkseq]$
[hadoop@en000 sparkseq]$
[hadoop@en000 sparkseq]$
[hadoop@en000 sparkseq]$ sbt package
[info] Loading project definition from /BiO/hadoop/hadoop_tools/sparkseq/project
[info] Updating {file:/BiO/hadoop/hadoop_tools/sparkseq/project/}sparkseq-build...
[info] Resolving org.fusesource.jansi#jansi;1.4 ...
[info] Done updating.
[info] Compiling 1 Scala source to /BiO/hadoop/hadoop_tools/sparkseq/project/target/scala-2.10/sbt-0.13/classes...
[info] Set current project to sparkseq (in build file:/BiO/hadoop/hadoop_tools/sparkseq/)
[info] Updating {file:/BiO/hadoop/hadoop_tools/sparkseq/}sparkseq-core...
[info] Resolving com.sqality.scct#scct_2.10;0.3 ...
[info] Done updating.
[warn] There may be incompatibilities among your library dependencies.
[warn] Here are some of the libraries that were evicted:
[warn]  * org.apache.hadoop:hadoop-client:(1.2.1, 2.2.0) -> 2.6.0
[warn] Run 'evicted' to see detailed eviction warnings
[info] Updating {file:/BiO/hadoop/hadoop_tools/sparkseq/}sparkseq-repl...
[info] Resolving org.joda#joda-convert;1.2 ...
[info] Compiling 20 Scala sources to /BiO/hadoop/hadoop_tools/sparkseq/sparkseq-core/target/scala-2.10/classes...
[info] Resolving com.sqality.scct#scct_2.10;0.3 ...
[info] Done updating.
[warn] There may be incompatibilities among your library dependencies.
[warn] Here are some of the libraries that were evicted:
[warn]  * org.apache.spark:spark-core_2.10:1.0.2 -> 1.5.0
[warn] Run 'evicted' to see detailed eviction warnings
[info] Updating {file:/BiO/hadoop/hadoop_tools/sparkseq/}sparkseq...
[info] Resolving com.sqality.scct#scct_2.10;0.3 ...
[info] Done updating.
[warn] there were 4 deprecation warning(s); re-run with -deprecation for details
[warn] there were 1 feature warning(s); re-run with -feature for details
[warn] two warnings found
[warn] Multiple main classes detected.  Run 'show discoveredMainClasses' to see the list
[info] Packaging /BiO/hadoop/hadoop_tools/sparkseq/target/scala-2.10/sparkseq_2.10-0.1-SNAPSHOT.jar ...
[info] Packaging /BiO/hadoop/hadoop_tools/sparkseq/sparkseq-repl/target/scala-2.10/sparkseq-repl_2.10-0.1-SNAPSHOT.jar ...
[info] Packaging /BiO/hadoop/hadoop_tools/sparkseq/sparkseq-core/target/scala-2.10/sparkseq-core_2.10-0.1-SNAPSHOT.jar ...
[info] Done packaging.
[info] Done packaging.
[info] Done packaging.
[success] Total time: 20 s, completed 2016. 4. 19 오후  3:47:04
[hadoop@en000 sparkseq]$ 
```

⑤ Unit 테스트

# sbt test



```
16/04/19 15:53:29 INFO TaskSetManager: Starting task 3.0 in stage 3.0 (TID 5, localhost, PROCESS_LOCAL, 2010 bytes)
16/04/19 15:53:29 INFO Executor: Running task 0.0 in stage 3.0 (TID 2)
16/04/19 15:53:29 INFO Executor: Running task 1.0 in stage 3.0 (TID 3)
16/04/19 15:53:29 INFO Executor: Running task 2.0 in stage 3.0 (TID 4)
16/04/19 15:53:29 INFO Executor: Running task 3.0 in stage 3.0 (TID 5)
16/04/19 15:53:29 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
16/04/19 15:53:29 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/04/19 15:53:29 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
16/04/19 15:53:29 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/04/19 15:53:29 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
16/04/19 15:53:29 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
16/04/19 15:53:29 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/04/19 15:53:29 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/04/19 15:53:29 INFO Executor: Finished task 3.0 in stage 3.0 (TID 5). 1161 bytes result sent to driver
16/04/19 15:53:29 INFO Executor: Finished task 0.0 in stage 3.0 (TID 2). 1354 bytes result sent to driver
16/04/19 15:53:29 INFO TaskSetManager: Finished task 3.0 in stage 3.0 (TID 5) in 20 ms on localhost (1/4)
16/04/19 15:53:29 INFO Executor: Finished task 1.0 in stage 3.0 (TID 3). 1161 bytes result sent to driver
16/04/19 15:53:29 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 2) in 22 ms on localhost (2/4)
16/04/19 15:53:29 INFO TaskSetManager: Finished task 1.0 in stage 3.0 (TID 3) in 22 ms on localhost (3/4)
16/04/19 15:53:29 INFO Executor: Finished task 2.0 in stage 3.0 (TID 4). 1161 bytes result sent to driver
16/04/19 15:53:29 INFO TaskSetManager: Finished task 2.0 in stage 3.0 (TID 4) in 23 ms on localhost (4/4)
16/04/19 15:53:29 INFO DAGScheduler: ResultStage 3 (first at SparkSeqAnalysisSuite.scala:28) finished in 0.025 s
16/04/19 15:53:29 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
16/04/19 15:53:29 INFO DAGScheduler: Job 1 finished: first at SparkSeqAnalysisSuite.scala:28, took 0.032452 s
16/04/19 15:53:29 INFO SparkUI: Stopped Spark web UI at http://localhost:30200
16/04/19 15:53:29 INFO DAGScheduler: Stopping DAGScheduler
16/04/19 15:53:29 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
16/04/19 15:53:29 INFO MemoryStore: MemoryStore cleared
16/04/19 15:53:29 INFO BlockManager: BlockManager stopped
16/04/19 15:53:29 INFO BlockManagerMaster: BlockManagerMaster stopped
16/04/19 15:53:29 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
16/04/19 15:53:29 INFO SparkContext: Successfully stopped SparkContext
16/04/19 15:53:29 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
16/04/19 15:53:29 INFO RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.
16/04/19 15:53:29 INFO RemoteActorRefProvider$RemotingTerminator: Remoting shut down.
[info] Run completed in 6 seconds, 567 milliseconds.
[info] Total number of tests run: 3
[info] Suites: completed 1, aborted 0
[info] Tests: succeeded 3, failed 0, canceled 0, ignored 0, pending 0
[info] All tests passed.
[success] Total time: 12 s, completed 2016. 4. 19 오후 3:53:29
16/04/19 15:53:29 INFO ShutdownHookManager: Shutdown hook called
16/04/19 15:53:29 INFO ShutdownHookManager: Deleting directory /tmp/spark-58a04495-8c50-4568-afd7-2ee16a715914
[hadoop@en000 sparkseq]$
```

⑥ SparkSeq용 Lib디렉토리 생성 및 Library 카피

```
# mkdir /opt/cloudera/parcels/CDH/lib/sparkseq

# 5.1.1 Hadoop-BAM 내용 참고
# cp /BiO/hadoop/hadoop_tools/hadoop-bam-6.1/hadoop-bam-6.1.jar /opt/cloudera/parcels/CDH/lib/sparkseq/

# cp /home/hadoop/.m2/repository/org/seqdoop/hadoop-bam/7.1.0/hadoop-bam-7.1.0.jar /opt/cloudera/parcels/CDH/lib/sparkseq/

# cp /home/hadoop/.m2/repository/org/apache/commons/commons-jexl/2.1.1/commons-jexl-2.1.1.jar /opt/cloudera/parcels/CDH/lib/sparkseq/

# cp /home/hadoop/.ivy2/cache/org.seqdoop/htsjdk/jars/htsjdk-1.118.jar /opt/cloudera/parcels/CDH/lib/sparkseq/

# wget http://hadoop-bam.sourceforge.net/maven/picard/picard/1.93/picard-1.93.jar -P /opt/cloudera/parcels/CDH/lib/sparkseq/

# wget http://downloads.sourceforge.net/project/picard/sam-jdk/1.93/sam-
```

```
1.93.jar?r=https%3A%2F%2Fsourceforge.net%2Fprojects%2Fpicard%2Ffiles%2Fsam
-jdk%2F1.93%2F&ts=1461049989&use_mirror=heanet -P
/opt/cloudera/parcels/CDH/lib/sparkseq/
```

```
# cp /BiO/hadoop/hadoop_tools/sparkseq/sparkseq-core/target/scala-
2.10/sparkseq-core_2.10-0.1-SNAPSHOT.jar
/opt/cloudera/parcels/CDH/lib/sparkseq/
```

```
# wget http://hadoop-bam.sourceforge.net/maven/tribble/tribble/1.93/tribble-
1.93.jar  -P /opt/cloudera/parcels/CDH/lib/sparkseq/
```

```
# wget
http://codenav.org/code.html?project=/edu/berkeley/cs/amplab/adam/adam-
cli/0.7.1&path=/Dependencies/variant-1.93.jar -P
/opt/cloudera/parcels/CDH/lib/sparkseq/
```

ADD JARS 및 SPARK CLASSPATH 설정

```
# vi ~/.bashrc
export SPARKSEQ_LIB="/opt/cloudera/parcels/CDH/lib/sparkseq"
export ADD_JARS="${SPARKSEQ_LIB}/htsjdk-1.118.jar,${SPARKSEQ_LIB}/hadoop-
bam-6.1.jar,${SPARKSEQ_LIB}/hadoop-bam-7.1.0.jar,${SPARKSEQ_LIB}/picard-
1.93.jar,${SPARKSEQ_LIB}/sam-1.93.jar,${SPARKSEQ_LIB}/variant-
1.93.jar,${SPARKSEQ_LIB}/tribble-1.93.jar,${SPARKSEQ_LIB}/commons-jexl-
2.1.1.jar,${SPARKSEQ_LIB}/sparkseq-core_2.10-0.1-SNAPSHOT.jar"
export SPARK_CLASSPATH="${SPARKSEQ_LIB}/htsjdk-
1.118.jar:${SPARK_CLASSPATH}:${SPARKSEQ_LIB}/hadoop-bam-
6.1.jar:${SPARKSEQ_LIB}/hadoop-bam-7.1.0.jar:${SPARKSEQ_LIB}/picard-
1.93.jar:${SPARKSEQ_LIB}/sam-1.93.jar:${SPARKSEQ_LIB}/variant-
1.93.jar:${SPARKSEQ_LIB}/tribble-1.93.jar:${SPARKSEQ_LIB}/commons-jexl-
2.1.1.jar:${SPARKSEQ_LIB}/sparkseq-core_2.10-0.1-SNAPSHOT.jar"
```

```
# source ~/.bashrc
```

⑦ 모든 DataNode에 jar파일 및 .bashrc 파일 배포

# scp -R /opt/cloudera/parcels/CDH/lib/sparkseq dn000~dn025: /opt/cloudera/parcels/CDH/lib/

# scp ~/.bashrc dn000~dn025:~/.bashrc

1.3 SparkSeq 예제 실행

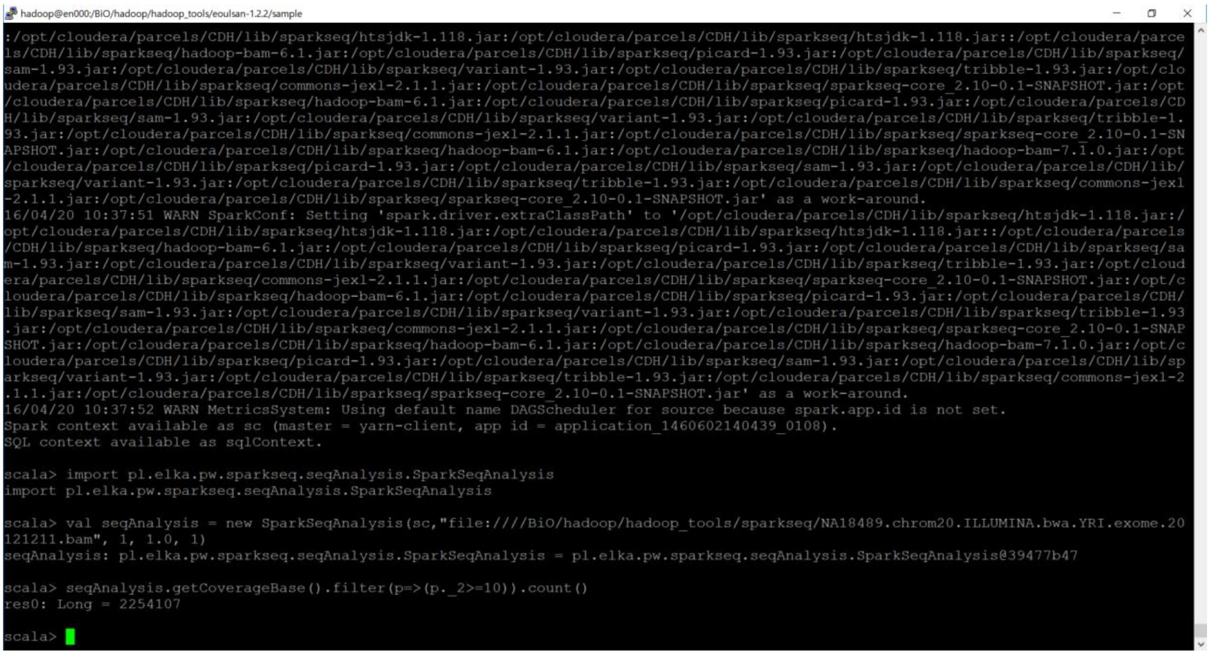
① Spark shell을 yarn-client로 실행

# spark-shell --master yarn-client --driver-memory 4G

# import pl.elka.pw.sparkseq.seqAnalysis.SparkSeqAnalysis

# val seqAnalysis = new SparkSeqAnalysis(sc,"file:////BiO/hadoop/hadoop_tools/sparkseq/NA18489.chrom20.ILLUMINA.bwa.YRI.exome.20121211.bam", 1, 1.0, 1)

# seqAnalysis.getCoverageBase().filter(p=>(p._2>=10)).count()



hadoop@en000:/BiO/hadoop/hadoop_tools/eoulsan-1.2.2/sample

```
:/opt/cloudera/parcels/CDH/lib/sparkseq/htsjdk-1.118.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/htsjdk-1.118.jar::/opt/cloudera/parce
ls/CDH/lib/sparkseq/hadoop-bam-6.1.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/picard-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/
sam-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/variant-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/tribble-1.93.jar:/opt/clo
udera/parcels/CDH/lib/sparkseq/commons-jexl-2.1.1.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/sparkseq-core_2.10-0.1-SNAPSHOT.jar:/opt
/cloudera/parcels/CDH/lib/sparkseq/hadoop-bam-6.1.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/picard-1.93.jar:/opt/cloudera/parcels/CD
H/lib/sparkseq/sam-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/variant-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/tribble-1.
93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/commons-jexl-2.1.1.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/sparkseq-core_2.10-0.1-SN
APSHOT.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/hadoop-bam-6.1.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/hadoop-bam-7.1.0.jar:/opt
/cloudera/parcels/CDH/lib/sparkseq/picard-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/sam-1.93.jar:/opt/cloudera/parcels/CDH/lib/
sparkseq/variant-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/tribble-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/commons-jexl
-2.1.1.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/sparkseq-core_2.10-0.1-SNAPSHOT.jar' as a work-around.
16/04/20 10:37:51 WARN SparkConf: Setting 'spark.driver.extraClassPath' to '/opt/cloudera/parcels/CDH/lib/sparkseq/htsjdk-1.118.jar:/
opt/cloudera/parcels/CDH/lib/sparkseq/htsjdk-1.118.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/htsjdk-1.118.jar::/opt/cloudera/parcels
/CDH/lib/sparkseq/hadoop-bam-6.1.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/picard-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/sa
m-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/variant-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/tribble-1.93.jar:/opt/cloud
era/parcels/CDH/lib/sparkseq/commons-jexl-2.1.1.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/sparkseq-core_2.10-0.1-SNAPSHOT.jar:/opt/c
loudera/parcels/CDH/lib/sparkseq/hadoop-bam-6.1.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/picard-1.93.jar:/opt/cloudera/parcels/CDH/
lib/sparkseq/sam-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/variant-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/tribble-1.93
.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/commons-jexl-2.1.1.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/sparkseq-core_2.10-0.1-SNAP
SHOT.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/hadoop-bam-6.1.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/hadoop-bam-7.1.0.jar:/opt/c
loudera/parcels/CDH/lib/sparkseq/picard-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/sam-1.93.jar:/opt/cloudera/parcels/CDH/lib/sp
arkseq/variant-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/tribble-1.93.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/commons-jexl-2
.1.1.jar:/opt/cloudera/parcels/CDH/lib/sparkseq/sparkseq-core_2.10-0.1-SNAPSHOT.jar' as a work-around.
16/04/20 10:37:52 WARN MetricsSystem: Using default name DAGScheduler for source because spark.app.id is not set.
Spark context available as sc (master = yarn-client, app id = application_1460602140439_0108).
SQL context available as sqlContext.

scala> import pl.elka.pw.sparkseq.seqAnalysis.SparkSeqAnalysis
import pl.elka.pw.sparkseq.seqAnalysis.SparkSeqAnalysis

scala> val seqAnalysis = new SparkSeqAnalysis(sc,"file:////BiO/hadoop/hadoop_tools/sparkseq/NA18489.chrom20.ILLUMINA.bwa.YRI.exome.20
121211.bam", 1, 1.0, 1)
seqAnalysis: pl.elka.pw.sparkseq.seqAnalysis.SparkSeqAnalysis = pl.elka.pw.sparkseq.seqAnalysis.SparkSeqAnalysis@39477b47

scala> seqAnalysis.getCoverageBase().filter(p=>(p._2>=10)).count()
res0: Long = 2254107

scala>
```