#Commands used for ADAM Genomics analysis platform with specialized file formats

#https://github.com/bigdatagenomics/adam

# Requires MAVEN 3.1.0 + for installation (Requires Java 1.7)

#MAVEN installation:

# <https://maven.apache.org/download.cgi>

# Instructions:

# <http://stackoverflow.com/questions/7532928/how-do-i-install-maven-with-yum>

# Changed “export M2\_HOME=/usr/local/apache-maven“

# to “ export M2\_HOME=/usr/local/apache-maven/apache-maven-3.3.9-bin.tar.gz”

# ADAM installation

# $ git clone <https://github.com/bigdatagenomics/adam.git>   
# $ cd adam  
# $ export MAVEN\_OPTS="-Xmx512m -XX:MaxPermSize=256m"  
# $ mvn clean package -DskipTests

# save alias in ~/.bashrc, where ${ADAM\_HOME} is the directory where ADAM is found

# alias adam-submit="${ADAM\_HOME}/bin/adam-submit"  
# alias adam-shell="${ADAM\_HOME}/bin/adam-shell"

# developer’s original usage command   
# adam-submit [<spark-args> --] <adam-args>

#===========Convert Reference .fasta file into .adam file===========

# RUN ADAM, convert FASTA to ADAM file, input must exist on HDFS and output folder not exist on HDFS, ie: hdfs://quickstart.cloudera:8020/user/cloudera/mrsa/adam/refseq/ASM28453v1\_genomic.fa

# spark default directory is hdfs://quickstart.cloudera:8020/user/cloudera/

# command format  
# adam-submit fasta2adam [spark options] <input fasta file> <output directory> [options]

$ adam-submit fasta2adam mrsa/adam/refseq/ASM28453v1\_genomic.fa mrsa/adam/adam.asm28453v1\_genomic

# OUTPUT: in hdfs://quickstart.cloudera:8020/user/cloudera/mrsa/adam/adam.asm28453v1\_genomic

#===========Convert Reference .adam file into .vcf===========

# once reference adam is created, create VCF to calculate depth of coverage for sequence data

# command format  
# adam-submit [spark options] adam2vcf <input adam directory> <output directory> [options]

$ adam-submit adam2vcf mrsa/adam/adam.asm28453v1\_genomic mrsa/adam/vcf.asm28453v1\_genomic -dict mrsa/adam/refseq/ASM28453v1\_genomic.dict

# Error

# Exception in thread “main” java.lang.AssertionError: assertion failed: BCF not yet supported

# The source code row 77 indicates that ADAM currently is not able to save file as “VCF” (<https://github.com/bigdatagenomics/adam/blob/master/adam-core/src/main/scala/org/bdgenomics/adam/rdd/variation/VariationRDDFunctions.scala>).

# As a result, depth of coverage cannot be calculated using ADAM.

#===========Transform .fastq into .adam File===========

# command format   
# adam-submit transform [spark options] <input fastq> <output dir> [options]

$ adam-submit transform mrsa/adam/staphiso/ERR064898\_1.cor.fastq mrsa/adam/adamq.ERR064898/ -mark\_duplicate\_reads -paired\_fastq mrsa/adam/staphiso/ERR064898\_2.cor.fastq -sort\_reads

# OUTPUT: in hdfs://quickstart.cloudera:8020/user/cloudera/mrsa/adam/adamq.ERR064898

#===========Transform .bam into .adam File===========

# command format   
# adam-submit transform [spark options] <input bam/sam/fastq> <output dir> [options]

$ adam-submit transform mrsa/adam/staphiso/ERR064898.bam mrsa/adam/adam.ERR064898 -markduplicate\_reads -sort\_reads

# OUTPUT: in hdfs://quickstart.cloudera:8020/user/cloudera/mrsa/adam/adam.ERR064898

#===========K-mer Count===========

# command format  
# adam-submit count\_kmers <input fastq> <output directory> <K-mer length>

# k-mer length can be estimated using kmergenie on local computer  
# best k-mer length is 55 for this *S. aureus*

$ adam-submit count\_kmers mrsa/adam/adam.ERR064898 mrsa/adam/stat.ERR064898/kmer.ERR064898 55

# OUTPUT: in hdfs://quickstart.cloudera:8020/user/cloudera/mrsa/adam/stat.ERR064898/kmer.ERR064898

#===========Print ADAM Files===========

# command format  
# adam-submit print [spark options] <adam dir>

$ adam-submit print mrsa/adam/adam.ERR064898 > printERR064898.adam

# OUTPUT directory: in local file directory where ADAM is being executed from (run pwd after ADAM finishes processing)

#===========Flagstat Similar to Samtools Flagstat===========

# command  
# adam-submit flagstat [spark options] <adam dir>

$ adam-submit flagstat mrsa/adam/adam.ERR064898 > ERR064898.flagstat

# OUTPUT directory: in local file directory where ADAM is being executed from (run pwd after ADAM finishes processing)

#===========Print All Tags and Counts===========

# command  
# adam-submit print\_tags [spark options] <adam dir>

$ adam-submit print\_tags mrsa/adam/adam.ERR064898 > ERR064898.tags

# OUTPUT directory: in local file directory where ADAM is being executed from (run pwd after ADAM finishes processing)

#===========Print ADAM Sequence Dictionary===========

# command  
# adam-submit listdict [spark options] <adam dir>

$ adam-submit listdict mrsa/adam/adam.ERR064898 > ERR064898.dic

# OUTPUT directory: in local file directory where ADAM is being executed from (run pwd after ADAM finishes processing)

#===========Flatten ADAM File===========

# command  
# adam-submit flatten [spark options] <adam dir>

$ adam-submit flatten mrsa/adam/adam.ERR064898 mrsa/adam/flatten.reference

# OUTPUT: in hdfs://quickstart.cloudera:8020/user/cloudera/mrsa/adam/flatten.reference