An accurate and variable length long read-mapping tool on Apache Spark Platform Steps to use StreamAligner

1. Install Hadoop and Apache Spark
2. Download StreamAligner from github.
3. Go to StreamAligner directory cd StreamAligner/
4. Put reference genome (text or fasta) and query data (text, fasta or fastq) file on Hadoop hadoop fs -copyFromLocal /path/to/reference.fasta / hadoop fs -copyFromLocal /path/to/query.fastq /

Index Generation

1. Transform and clean reference

cd suffix-prep

/path/to/spark/bin/spark-submit --class StreamAligner --master spark://cluster:7077 target/StreamAligner-1.0.jar hdfs://localhost:54310/reference.fasta hdfs://localhost:54310/cleaned-reference.txt reference-line-length suffix-keylength hdfs://localhost:54310/reference-chromosome-indexes

1. Create suffix array index for A

cd ..

cd suffix-a

/path/to/spark/bin/spark-submit --class StreamAligner --master spark://cluster:7077 target/StreamAligner-1.0.jar hdfs://localhost:54310/reference.fasta hdfs://localhost:54310/cleaned-reference.txt reference-line-length suffix-keylength hdfs://localhost:54310/reference-chromosome-indexes hdfs://localhost:54310/suffix-a

1. Create suffix array index for C

cd ..

cd suffix-c

/path/to/spark/bin/spark-submit --class StreamAligner --master spark://cluster:7077 target/StreamAligner-1.0.jar hdfs://localhost:54310/reference.fasta hdfs://localhost:54310/cleaned-reference.txt reference-line-length suffix-keylength hdfs://localhost:54310/reference-chromosome-indexes hdfs://localhost:54310/suffix-a hdfs://localhost:54310/suffix-c

1. Create suffix array index for G

cd ..

cd suffix-g

/path/to/spark/bin/spark-submit --class StreamAligner --master spark://cluster:7077 target/StreamAligner-1.0.jar hdfs://localhost:54310/reference.fasta hdfs://localhost:54310/cleaned-reference.txt reference-line-length suffix-keylength hdfs://localhost:54310/reference-chromosome-indexes hdfs://localhost:54310/suffix-a hdfs://localhost:54310/suffix-c hdfs://localhost:54310/suffix-g

9. Create suffix array index for T

cd ..

cd suffix-t

/path/to/spark/bin/spark-submit --class StreamAligner --master spark://cluster:7077 target/StreamAligner-1.0.jar hdfs://localhost:54310/reference.fasta hdfs://localhost:54310/cleaned-reference.txt reference-line-length suffix-keylength hdfs://localhost:54310/reference-chromosome-indexes hdfs://localhost:54310/suffix-a hdfs://localhost:54310/suffix-c hdfs://localhost:54310/suffix-g hdfs://localhost:54310/suffix-t

1. Mapping read data

cd ..

cd query

/path/to/spark/bin/spark-submit --class StreamAligner --master spark://cluster:7077 target/StreamAligner-1.0.jar hdfs://localhost:54310/reference.fasta hdfs://localhost:54310/cleaned-reference.txt reference-line-length suffix-keylength hdfs://localhost:54310/reference-chromosome-indexes hdfs://localhost:54310/suffix-a hdfs://localhost:54310/suffix-c hdfs://localhost:54310/suffix-g hdfs://localhost:54310/suffix-t reference-size hdfs://localhost:54310/query.fastq query-delimiter number-of-error-allowed read-data-partitions hdfs://localhost:54310/result.fastq 120

1. Getting results from HDFS to local file system

hadoop fs -copyToLocal /result.fastq /path/to/output-director

Args Description

Args[0] Location of reference genome on hadoop

Args[1] Location to save cleaned and transformed reference genome on hadoop

Args[2] Number of base pairs in single line of reference genome

Args[3] Keylength- size of keys to sort suffixes

Args[4] Chromosomes with starting location in reference genome

Args[5] Suffix array for suffixes starting with A

Args[6] Suffix array for suffixes starting with C

Args[7] Suffix array for suffixes starting with G

Args[8] Suffix array for suffixes starting with N

Args[9] Suffix array for suffixes starting with T

Args[10] Total size of reference

Args[11] Query data location on Hadoop

Args[12] Query data delimiter

Args[13] Number of error/mismatches allowed

Args[14] Number of partitions for read data

Args[15] Location to save results on Hadoop

Args[16] Time for stream

Example:

1. Install Hadoop and Apache Spark
2. Download StreamAligner from github.
3. Go to StreamAligner directory cd StreamAligner/
4. Put reference genome (text or fasta) and query data (text, fasta or fastq) file on Hadoop hadoop fs -copyFromLocal data/chr1.fasta / hadoop fs -copyFromLocal data/query.fastq /

Index Generation

1. Transform and clean reference

cd suffix-prep

/path/to/spark/bin/spark-submit --class StreamAligner --master spark://cluster:7077 target/StreamAligner-1.0.jar hdfs://localhost:54310/chr1.fasta hdfs://localhost:54310/cleaned-reference.txt 5020hdfs://localhost:54310/reference-chromosome-indexes

1. Create suffix array index for A

cd ..

cd suffix-a

/path/to/spark/bin/spark-submit --class StreamAligner --master spark://cluster:7077 target/StreamAligner-1.0.jar hdfs://localhost:54310/chr1.fasta hdfs://localhost:54310/cleaned-reference.txt 5020hdfs://localhost:54310/reference-chromosome-indexes hdfs://localhost:54310/suffix-a

1. Create suffix array index for C

cd ..

cd suffix-c

/path/to/spark/bin/spark-submit --class StreamAligner --master spark://cluster:7077 target/StreamAligner-1.0.jar hdfs://localhost:54310/chr1.fasta hdfs://localhost:54310/cleaned-reference.txt 5020hdfs://localhost:54310/reference-chromosome-indexes hdfs://localhost:54310/suffix-a hdfs://localhost:54310/suffix-c

1. Create suffix array index for G

cd ..

cd suffix-g

/path/to/spark/bin/spark-submit --class StreamAligner --master spark://cluster:7077 target/StreamAligner-1.0.jar hdfs://localhost:54310/chr1.fasta hdfs://localhost:54310/cleaned-reference.txt 5020hdfs://localhost:54310/reference-chromosome-indexes hdfs://localhost:54310/suffix-a hdfs://localhost:54310/suffix-c hdfs://localhost:54310/suffix-g

1. Create suffix array index for T

cd ..

cd suffix-t

/path/to/spark/bin/spark-submit --class StreamAligner --master spark://cluster:7077 target/StreamAligner-1.0.jar hdfs://localhost:54310/chr1.fasta hdfs://localhost:54310/cleaned-reference.txt 5020hdfs://localhost:54310/reference-chromosome-indexes hdfs://localhost:54310/suffix-a hdfs://localhost:54310/suffix-c hdfs://localhost:54310/suffix-g hdfs://localhost:54310/suffix-t

1. Mapping read data

cd ..

cd query

/path/to/spark/bin/spark-submit --class StreamAligner --master spark://cluster:7077 target/StreamAligner-1.0.jar hdfs://localhost:54310/chr1.fasta hdfs://localhost:54310/cleaned-reference.txt 5020hdfs://localhost:54310/reference-chromosome-indexes hdfs://localhost:54310/suffix-a hdfs://localhost:54310/suffix-c hdfs://localhost:54310/suffix-g hdfs://localhost:54310/suffix-t 249250627hdfs://localhost:54310/query.fastq “@HWI” 2 6 hdfs://localhost:54310/result.fastq 120

1. Getting results from HDFS to local file system

hadoop fs -copyToLocal /result.fastq /home/output