BigBWA

BigBWA is a tool to run the Burrows-Wheeler Aligner--BWA on a Hadoop cluster. The current version of BigBWA (2.1, november 2016) supports the following BWA algorithms:

- BWA-MEM
- BWA-backtrack
- BWA-SW

All of them work with paired and single-end reads.

If you use **BigBWA**, please cite this article:

> José M. Abuin, Juan C. Pichel, Tomás F. Pena and Jorge Amigo. "BigBWA: approaching the Burrows–Wheeler aligner to Big Data technologies". Bioinformatics 31(24), pp. 4003-4005, 2015.

A version for Apache Spark is available here.

Structure

Since version 2.0 the project keeps a standard Maven structure. The source code is in the *src/main* folder. Inside it, we can find two subfolders:

- **java** - Here is where the Java code is stored.
- **native** - Here the BWA native code (C) and the glue logic for JNI is stored.

Getting started

Requirements

Requirements to build **BigBWA** are the same than the ones to build BWA, with the only exception that the *JAVA_HOME* environment variable should be defined. If not, you can define it in the */src/main/native/Makefile.common* file.

It is also needed to include the flag *-fPIC* in the *Makefile* of the considered BWA version. To do this, the user just need to add this option to the end of the *CFLAGS* variable in the BWA Makefile. Considering bwa-0.7.15, the original Makefile contains:

```
CFLAGS=     -g -Wall -Wno-unused-function -O2
```

and after the change it should be:

```
CFLAGS=     -g -Wall -Wno-unused-function -O2 -fPIC
```

Additionaly, and as **BigBWA** is built with Maven since version 0.2, also have it in the user computer is needed.

Building

The default way to build **BigBWA** is:

```
git clone https://github.com/citiususc/BigBWA.git
cd BigBWA
mvn package
```

This will create the *target* folder, which will contain the *jar* file needed to run **BigBWA**:

- **BigBWA-2.1.jar** - jar file to launch with Hadoop.

Configuring

Since version 2.0 there is no need of configuring any Hadoop parameter. The only requirement is that the YARN containers need to have at least 7500MB of memory available (for the human genome case).

Running BigBWA

BigBWA requires a working Hadoop cluster. Users should take into account that at least 7500MB of free memory per map are required (each map loads into memory the bwa index). Note that **BigBWA** uses disk space in the Hadoop ***tmp*** directory.

Here it is an example of how to run**BigBWA** using the BWA-MEM paired algorithm. This example assumes that our index is stored in all the cluster nodes at */Data/HumanBase/* . The index can be obtained with BWA, using "bwa index".

First, we get the input Fastq reads from the 1000 Genomes Project ftp:

```
wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/NA12750/sequence_read/ERR000589_1.filt.fastq.gz
wget ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/NA12750/sequence_read/ERR000589_2.filt.fastq.gz
```

Next, the downloaded files should be uncompressed:

```
gzip -d ERR000589_1.filt.fastq.gz
gzip -d ERR000589_2.filt.fastq.gz
```

and prepared to be used by BigBWA:

```
python src/utils/Fq2FqBigDataPaired.py ERR000589_1.filt.fastq ERR000589_2.filt.fastq ERR000589.fqBD

hdfs dfs -copyFromLocal ERR000589.fqBDP ERR000589.fqBDP
```

Finally, we can execute **BigBWA** on the Hadoop cluster:

```
yarn jar BigBWA-2.1.jar com.github.bigbwa.BigBWA -D mapreduce.input.fileinputformat.split.minsize=123641127
-D mapreduce.input.fileinputformat.split.maxsize=123641127
-D mapreduce.map.memory.mb=7500
-w "-R @RG\tID:foo\tLB:bar\tPL:illumina\tPU:illumina\tSM:ERR000589 -t 2"
-m -p --index /Data/HumanBase/hg19 -r ERR000589.fqBDP ExitERR000589
```

Options:

- **-m** - Sequence alignment algorithm.
- **-p** - Use paired-end reads.
- **-w "args"** - Can be used to pass arguments directly to BWA (ex. "-t 4" to specify the amount of threads to use per instance of BWA).
- **--index index_prefix** - Index prefix is specified. The index must be available in all the cluster nodes at the same location.
- The last two arguments are the input and output HDFS files.

If you want to check all the available options, execute the command:

```
yarn jar BigBWA-2.1.jar com.github.bigbwa.BigBWA -h
```

The commands are:

```
BigBWA performs genomic alignment using bwa in a Hadoop/YARN cluster
 usage: yarn jar --class com.github.bigbwa.BigBWA BigBWA-2.1.jar
       [-a | -b | -m] [-h] [-i <Index prefix>]   [-n <Number of
       partitions>] [-p | -s] [-r]  [-w <"BWA arguments">]
       <FASTQ file> <SAM file output>
Help options:
  -h, --help                                   Shows this help

Input FASTQ reads options:
  -p, --paired                                  Paired reads will be used as input FASTQ reads
  -s, --single                                 Single reads will be used as input FASTQ reads

BWA algorithm options:
  -a, --aln                                   The ALN algorithm will be used
  -b, --bwasw                                    The bwasw algorithm will be used
  -m, --mem                                       The MEM algorithm will be used

Index options:
  -i, --index <Index prefix>                         Prefix for the index created by bwa to use - setIndexPath(string)

Spark options:
  -n, --partitions <Number of partitions>         Number of partitions to divide input - setPartitionNumber(int)

Reducer options:
  -r, --reducer                                 The program is going to merge all the final results in a reducer phase

BWA arguments options:
  -w, --bwa <"BWA arguments">                      Arguments passed directly to BWA
```

After the execution, to move the output to the local filesystem use:

```
hdfs dfs -copyToLocal ExitERR000589/part-r-00000 ./
```

In case there is no reducer, the output will be split into several pieces. In order to put it together users could use one of our Python utils or "samtools merge":

```
hdfs dfs -copyToLocal ExitERR000589/Output* ./
python src/utils/FullSam.py ./ ./OutputFile.sam
```

Frequently asked questions (FAQs)

1. I can not build the tool because *jni_md.h* or *jni.h* is missing.

1. I can not build the tool because *jni_md.h* or *jni.h* is missing.

You need to set correctly your *JAVA_HOME* environment variable or you can set it in Makefile.common.

INFORMACIÓN

Investigadores
José Manuel Abuín Mosquera
Juan Carlos Pichel Campos
Tomás Fernández Pena
Jorge Amigo Lechuga

Licenza

DESCARGAR

- Repositorio Gitlab
- Descargar de Gitlab
- Repositorio Github

PUBLICACIÓNS

BigBWA: Approaching the Burrows-Wheeler Aligner to Big Data Technologies
Bioinformatics, 2015

PROXECTOS DE INVESTIGACIÓN

SHSCAP: Solucións hardware e software para a computación de altas prestacións