Gene Resequencing with Myrna on Intel® Distribution of Hadoop

V1.00

Intel Corporation

Authors

Abhi Basu

Contributors

Terry Toy

Gaurav Kaul

TABLE OF CONTENTS

1.Executive Summary

Genome resequencing allows us to understand how genetic differences affect health and cause diseases. This is an important step in detecting anomalies associated with many genetically inherited diseases like Heart Disorders, Down Syndrome, Cystic Fibrosis and Chromosomal Abnormalities [1]. Next Generation Sequencing (NGS) technologies running on High Performance Computing (HPC) architectures have enabled the sequencing on DNA at groundbreaking speeds [2]. However the storage, analysis and management of the massive DNA sequence datasets produced as a result of NGS research, is a new challenge. Hadoop and Mapreduce technologies come into play here by allowing parallel read-mapping algorithms to scale effectively and resulting in shorter execution times and lower costs (from software execution and hardware). Among other areas Hadoop technologies may be useful are data storage, data management, statistical analysis and statistical association between various data sources. Organizations are now able to store large datasets in Hadoop Distributed File Systems (HDFS) and are able to use real-time analytics software to access data directly from HDFS bypassing any data migration headaches. Software packages like Myrna, developed by Ben Langmead, Kasper Hansen and Jeff Leek (John Hopkins University) is one such tool that allows the calculation of differential gene expressions in RNA-seq datasets on cloud (Amazon Elastic Map Reduce) or Hadoop clusters [3].

Innovative companies like Intel ® Corporation are interested in collaborating with various key partners in the Life Sciences area in an effort to accelerate such work. Intel® wants to provide businesses with an open enterprise Hadoop platform alternative for next generation analytics and life sciences, called the Intel® Distribution for Apache Hadoop Software, which provides better manageability and performance – optimized for Intel Xeon processors [4].

In this paper, we demonstrate how to install and configure Myrna and its required components – Bowtie, R/Bioconductor and SRA toolkit within the Intel® Hadoop Distribution.

2.Introduction

2.1. Goal

The goal of this paper is to demonstrate how to install a very specific gene resequencing software like Myrna on Intel® Distribution of Hadoop (IDH) running on Cent OS 6.3 operating system.

2.2. Objectives

Specifically the following are the key objectives of this paper:

1. Define the operating system, software stack and tools.
2. Define the setup and configuration of Crossbow and components.
3. Run some tests from Myrna to validate that the installation of all components were successful.

2.3.Audience

Software developers and technologists can use this document to install Myrna software.

2.4.Terminology

Term	Description
IDH	Intel® Distribution of Apache Hadoop Software
HDFS	Hadoop Distributed File System
JDK	Java Development Kit
BKM	Best Known Method
HPC	High Performance Computing
NGS	Next Generation Sequencing

3.System Setup and Configuration

3.1. Components

Component	Details
Hardware (each node – 3 node cluster used)	• Intel® Xeon DP Sandy Bridge-EP E5-2680 FC-LGA10 2.7GHz 8.0GT/s 20MB 130W 8 Cores CM8062107184424 (2-CPU) • 128 GB RAM (1333 Reg ECC 1.5V DDR3 Romley) • KDK – Grizzly Pass 2U 12x3.5 SATA 2x750W 2xHS Rails Intel R2312GZ4GC4 • 300GB SSD 2.5in SATA 3Gb/s 25nm Intel Lyndonville SSDSA2BZ300G301 710 Series • 2TB HDD 3.5in SATA 6Gb/s 7200RPM 64MB Seagate Constellation ES ST2000NM0011 • NIC - Niantic X520-SR2 10GBase-SR PCI-e Dual Port E10G42BFSR or E10G42BFSRG1P5 Duplex Fiber Optic • LSI HBA LSI00194 (9211-8i) 8 port 6Gb/s SATA +SAS PCIe 2.0 Raid LP
Operating System	Cent OS 6.3 (Developers Workstation version)
Application Software	1. IDH 2.3 (Hadoop 1.0.3-Intel) running a 3 Node Cluster 2. Myrna 1.21 3. Bowtie 0.12.8 4. SRA Toolkit 2.3.2-5 5. R 3.0.1 6. Bioconductor 2.12

3.2.Software Setup

These instructions assume that a fully-functional version of IDH 2.3 has already been installed on Cent OS 6.3 (Developers Workstation Version) and is ready for use. All installations were performed logged in as *root* user to the system. For alternate Linux distributions, take a look at the Appendix (section 6).

3.2.1. Myrna

Myrna is used for calculating differential in genome expression in large RNA-seq datasets in the cloud [5]. It is available for the Linux OS. The following are the steps to download and install Myrna on Cent OS.

1. Download the Myrna 1.2.1 binary from – http://sourceforge.net/projects/bowtie-bio/files/myrna/.
2. Copy file to /usr/local/bin.
3. Unzip file.
4. The folder /myrna1.2.1 will contain all binaries that we require.

Note : All of the above needs to be installed on each node of the Hadoop Cluster.

3.2.2. Bowtie

Bowtie is a fast memory-efficient short read aligner for alignment of DNA sequences [6]. The use of Crossbow allow Bowtie to be run in a distributed enviroment like Hadoop. The following are the steps to download and install Bowtie on Cent OS.

1. Download the Bowtie 0.12.8 binary from – https://sourceforge.net/projects/bowtie-bio/files/bowtie/0.12.8/ [This version is customized for Crossbow].
2. Copy file to /usr/local/bin.
3. Unzip file.
4. The folder /bowtie0.12.8 will contain all binaries that we require.

Note : All of the above needs to be installed on each node of the Hadoop Cluster.

3.2.3. R/Bioconductor

Bioconductor is an open-source analytic and genomic data comprehension tool which is built on top of R [7]. The already downloaded Myrna 1.2.1 package includes a bash script that allows R and Bioconductor to be downloaded and installed correctly and all required R packages installed.

1. Navigate to the Myrna home folder (/usr/bin/local/myrna1.2.1 in our case).
2. Change directory to /R (/usr/bin/local/myrna1.2.1/R).
3. Run the command ./build_r.

Note : All of the above needs to be installed on each node of the Hadoop Cluster.

3.2.4. SRA Toolkit

The Sequence Read Archive (SRA) is a resource offered by the National Center for Biotechnical Information (NCBI) for sequence data storage [8] . The following are the steps to install SRA Toolkit on Cent OS.

1. Download SRA toolkit (v2.3.2-5) from – http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software.
2. Copy file to /usr/local/bin .
3. Untar the file.
4. The folder /sratoolkit.2.3.2-5-centos_linux64 will contain all binaries that we require.

Note : All of the above needs to be installed on each node of the Hadoop Cluster.

3.2.5.Setting Environment Variables

The following environment variables need to be set. Ensure that the pathnames and filenames match those in your environment.

We added them to the /root/.bashrc file :

export JAVA_HOME=/usr/java/latest

export PATH=$JAVA_HOME/bin:$PATH

export HADOOP_HOME=/usr/lib/hadoop

export PATH=$HADOOP_HOME/bin:$PATH

export HADOOP_CMD=/usr/bin/hadoop

export HADOOP_CONF=/usr/lib/hadoop/conf

export HADOOP_STREAMING=/usr/lib/hadoop/contrib/streaming/hadoop-streaming-0.20.jar

export MYRNA_HOME=/usr/local/bin/myrna-1.2.1

export MYRNA_BOWTIE_HOME=/usr/local/bin/bowtie-0.12.8

export MYRNA_SRATOOLKIT_HOME=/usr/local/bin/sratoolkit.2.3.2-5-centos_linux64/bin

export MYRNA_FASTQ_DUMP_HOME=/usr/local/bin/sratoolkit.2.3.2-5-centos_linux64/bin

We need to modify the /etc/wgetrc file with the following :

http_proxy = http://proxy.fm.intel.com:911/

ftp_proxy = http://proxy.fm.intel.com:911/

3.2.6.Setting Permissions

Give permissions for ***mapred*** user to be able to read, write and execute on the following folders :

/tmp

/usr/bin/local/myrna1.2.1

/usr/bin/local/bowtie0.12.8

/usr/local/bin/sratoolkit.2.3.2-5-centos_linux64

4.Testing the Install

4.1.Myrna Test

1. Log into the Hadoop cluster primary node.
2. Run command: ***$MYRNA_HOME/myrna_hadoop –test***
3. Expected output:

Searching for 'bowtie' binary...

Specified via --bowtie?....no

$MYRNA_BOWTIE_HOME specified?....YES (/usr/local/bin/bowtie-0.12.8)

Runnable?....YES

Searching for 'Rscript' binary...

Specified via --Rhome?....no

$MYRNA_RHOME specified?....no

Checking /usr/local/bin/myrna-1.2.1/bin...

Scanning directory: /usr/local/bin/myrna-1.2.1/bin/linux32

Scanning directory: /usr/local/bin/myrna-1.2.1/bin/linux64

Scanning directory: /usr/local/bin/myrna-1.2.1/bin/mac32

Scanning directory: /usr/local/bin/myrna-1.2.1/bin/mac64

I'm searching for R or Rscript, so scanning directory: /usr/local/bin/myrna-1.2.1/R/bin/Rscript

Checking whether R has appropriate R/Bioconductor packages...

[1] "Found required package lmtest"

[1] "Found required package multicore"

[1] "Found required package IRanges"

[1] "Found required package geneplotter"

[1] "All packages found"

Settling on /usr/local/bin/myrna-1.2.1/R/bin/Rscript

Searching for 'fastq-dump' binary...

Specified via --fastq-dump?....no

$MYRNA_SRATOOLKIT_HOME specified?....YES (/usr/local/bin/sratoolkit.2.3.2-5-centos_linux64/bin)

Runnable?....YES

Summary:

bowtie: INSTALLED at /usr/local/bin/bowtie-0.12.8/bowtie

R: INSTALLED with RHOME at /usr/local/bin/myrna-1.2.1/R/bin/Rscript

Hadoop note: executables must be runnable via the SAME PATH on all nodes.

PASSED install test

If you see the above, your cluster is configured correctly for Myrna!

At this time, you can run a more involved test with yeast data on your hadoop cluster by following the instructions available here - http://bowtie-bio.sourceforge.net/myrna/manual.shtml#hadoop, or your custom workload.

5.Conclusion and Future Work

The future of Hadoop in bioinformatics looks very promising. The ability to store and process massive datasets of any kind allows Hadoop technologies to be the predominant platform for Life Sciences and Analytics where research is leading to huge amounts of data being generated and subsequently analyzed.

Our future work will include installing and configuring various software tools (especially genomics software like Cloudburst and Contrail) to make IDH more productive for our customers [9].

6. Appendix

6.1. Red Hat Enterprise Linux ver 6.1

Confirm the version of the Linux distribution you are running

```
[crick1@iswhdpims ~]$ lsb_release -i -r
Distributor ID: RedHatEnterpriseServer
Release:	6.1
```

```
################################################

## shell script to install R 3.0 and Bioconductor packages in one compute node

################################################

# Setup some initial packages for R and BioConductor

####


sudo yum -y install samtools

sudo yum -y install curl-devel

sudo yum -y install git

sudo yum -y install cmake

sudo yum -y install tetex

sudo yum -y install texinfo

sudo yum -y install libxml2

sudo yum -y install libxml2-devel


##Set up R repo for yum to work

su -c 'rpm -Uvh http://download.fedoraproject.org/pub/epel/6/i386/epel-release-6-8.noarch.rpm'

sudo yum update
```

```
##Make sure the texinfo-tex and libjpeg-turbo RPMs are present in the current library

cp /storage/CrickTraining/Software/R/texinfo-tex-4.13a-8.el6.x86_64.rpm
/storage/CrickTraining/Software/R/libjpeg-turbo-1.2.1-1.el6.x86_64.rpm .

rpm -ivh texinfo-tex-4.13a-8.el6.x86_64.rpm


rpm -e --nodeps --allmatches libjpeg-6b-46.el6.x86_64

rpm -ivh libjpeg-turbo-1.2.1-1.el6.x86_64.rpm


##Now install R package

yum -y install R


##this should be run from the R shell separately from above

 install.packages('data.table', repos='http://cran.us.r-project.org')

      install.packages('xtable', repos='http://cran.us.r-project.org')

      install.packages('R.utils', repos='http://cran.us.r-project.org')

      install.packages('ggplot2', repos='http://cran.us.r-project.org')

      install.packages('scales', repos='http://cran.us.r-project.org')

                  install.packages("geneplotter")

                  install.packages("lmtest")

                  install.packages("multicore")

      install.packages('multicore', repos='http://cran.us.r-project.org')

      source("http://bioconductor.org/biocLite.R")

      biocLite()
```

The Myrna install test can be run after this to verify if installation was successful.

7.Further Reading

1. Resequencing. Available at http://www.ncbi.nlm.nih.gov/projects/genome/probe/doc/TechResequencing.shtml. Accessed on August 5, 2013.
2. Impact of next-generation sequencing on genomics. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3076108/. Accessed on August 30, 2013.
3. B. Langmead, K. Hansen and J. Leek. Cloud-scale RNA-sequencing differential expression analysis with Myrna. Available at http://genomebiology.com/2010/11/8/r83. Accessed August 5, 2013.
4. Intel® Distribution for Apache Hadoop Software page. Available at http://hadoop.intel.com/products/distribution. Accessed on July 15, 2013.
5. Myrna Software page. Available at http://bowtie-bio.sourceforge.net/myrna/index.shtml. Accessed on August 26, 2013.
6. Bowtie Project. Accessed at http://bowtie-bio.sourceforge.net/index.shtml. Accessed on August 1, 2013.
7. Bioconductor Software page. Available at http://www.bioconductor.org/. Accessed on August 26, 2013.
8. SRA Toolkit Page. Available at http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software. Accessed August 1, 2013.
9. R. Menon, G. Bhat and M. Schatz. Rapid Parallel Genome Indexing with MapReduce. Available at http://schatzlab.cshl.edu/publications/2011-GenomeIndexingMapReduce.pdf. Accessed August 5, 2013.

NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm