International Journal of Engineering & Technology, 7 (2.12) (2018) 189-192

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Spectral domain characterization of genome sequences

P. Venkateswarlu [1] *, E. G. Rajan [2]

[1] Department of Computer Science University of Mysore Karnataka, India.
[2] Avatar MedVision US. LLC, USA, Pentagram Labs Inc, California, USA, Helios & Matheson Analytics, USA
**Corresponding author E-mail: venkat123.pedakolmi@gmail.com*

Abstract

Genome sequencing became an important research area for understanding order of DNA and discovering genetic secrets of humans. Fortunately voluminous data in this area is available for the study of genome sequences. Characterization of genome sequences is non-trivial and tedious task. Nevertheless, algorithms were found in the literature to study them. As the genome sequences data has characteristics of big data we proposed a technique based on MapReduce programming paradigm to attempt spectral characterization of genome sequences. A machine learning approach is used to discover trends in the genome sequences. Rationale behind using MapReduce, a distributed programming framework, is its support for parallel processing and the usage of more powerful Graphical Processing Units (GPUs). Moreover, the datasets can be maintained in cloud so as to handle it with ease. We built a prototype application to demonstrate proof of the concept. Our empirical results reveal encouraging observations in the genomic study.

Keywords: *Genome Sequence; Spectral Characterization; Big Data; Map Reduce*

1. Introduction

Cloud computing technology is capable of providing infrastructure as a service that helps in achieving data intensive and computation-intensive applications. In bioinformatics there are many huge datasets that need parallel processing frameworks like Hadoop [16]. Hadoop is one of the distributed programming frameworks widely used to process data pertaining to bioinformatics. It is suitable for processing big data which is in the form of genome sequences. Since genome data is treated as big data, it needs parallel processing in resource rich environment like cloud.

It is important to work on genome sequences for characterisation using frameworks like Hadoop. Patterns found in the sequence can provide useful insights in bioinformatics. Hidden information in the genomic data can be obtained and understood for making well informed decisions. In the literature it is found that DNA sequences are analyzed to understand periodicity and, auto convolution and autocorrelation. In addition to Hadoop Genome Analysis Toolkit (GATK) [17] is used for experiments. This is the framework used to analyze genome sequences. It breaks DNA sequence data into manageable pieces. This framework is based on MapReduce programming paradigm. The remainder of the paper is structured as follows. Section 2 provides review of literature. Section 3 presents experimental setup. Section 4 presents proposed work. Section 5 presents experimental results while Section 6 concludes the paper and provides directions for future work.

2. Related work

This section provides review of literature on genome analysis. Albertsen et al. [1] used differential coverage binning method for genome sequences related to uncultured bacteria that are rarely available. Liu et al. [2] on the other hand studied gene cluster encoding to understand resistance power in rice. Silwal-Pandit et al. [4] studied mutation spectrum in breast cancer. Dubrovinaa et al. [5] studied genes under specific stress conditions to understand their prognostic relevance. Soverini et al. [6] investigated on the understanding of the complexities involved in kinase inhibitor-resistant populations by using ultra-deep sequencing. Rabbani et al. [7] focused on medical genetics to ascertain whole-exome sequencing.

Abdel-Wahab and Dey [8] explored ASXL–BAP1 axis to understand the prognosis related to cancer and epigenetics. Rytz et al. [9] on the other hand studied Ionotropic Receptors in Drosophila. Bertsch et al. [10] and Plant [11] studied gene modifications in microbes and rice respectively. Cross et al. [12] investigated on the clinical significance of mutations pertaining to NOTCH1 and SF3B1 mutations. Craig et al. [13] focused on transcriptome and genome sequences. Li et al. [14] studied genome wide association to understand genetic architecture of oil bio synthesis in maize kernels. Smith and Simmonds [15] focused on the classification of family Hepeviridae with consensus proposals. In this paper we studied genome analysis using a distributed programming framework for efficiency as the framework supports parallel processing.

3. Experimental setup

The experiments are made in distributed programming framework for genome sequence analysis. It is nothing but GATK. It is based on MapReduce programming paradigm. Therefore, this section provides details of Hadoop, HDFS, MapReduce and GATK. Hadoop is a distributed programming framework that supports MapReduce programming model.

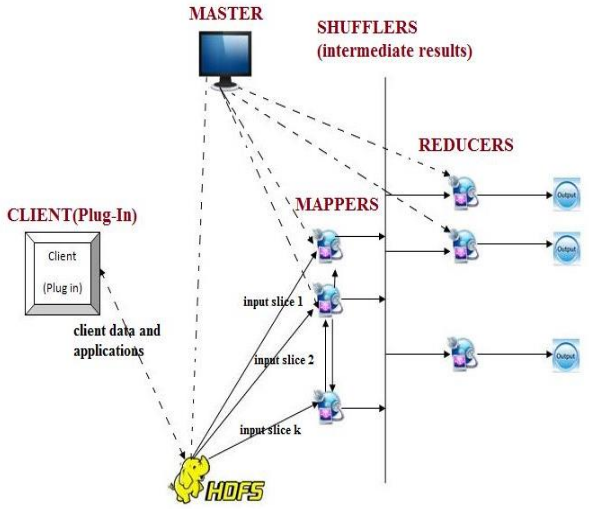

Fig. 1: Functionality of Mapreduce Paradigm.

As shown in Figure 1, it is understood that the Hadoop framework has associated file system known as Hadoop Distributed File System (HDFS). It has support for MapReduce programming paradigm. The input is split into multiple chunks and assigned to various mappers. The work of mappers is sent to reducers where the final output is generated. The mappers produce intermediate output. Both mappers and reducers are worker nodes in distributed environment. GATK is the toolkit that can be used to analyse genome sequences. It is used by different projects such as next generation DNA sequencing project and next generation sequencing project to mention few. The framework was developed by Genome Sequencing and Analysis Group from Harvard University and Broad Institute of MIT. GATK is used in Cancer Genome Atlas project. The tool can help in breaking terabases of sequence data into shards. The structure of GATK framework is as shown in Figure 2.

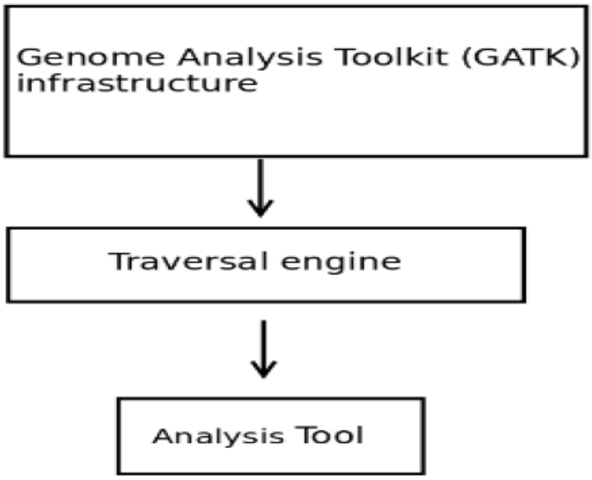

Fig. 2: The Structure of GATK.

The framework can be used for discovering genetic variations. There are different types of genetic variations. They are known as Single Nucleotide Aberrations, Short Insertions or Deletions (indels), Larger Structural Variations (SVs). The Single Nucleotide Aberrations are of two types namely Single Nucleotide Variations (SNVs) and Single Nucleotide Polymorphisms (SNPs). The differences among them are matter of frequency of occurrence. There is a framework for variation discovery [3] as shown in Figure 3.

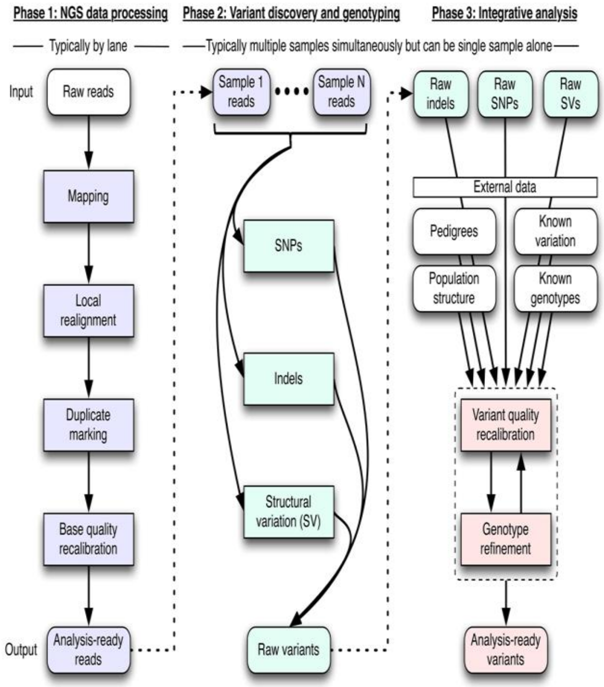

Fig 3: Framework for Discovery of Variations.

There are three phases involved in the framework. In the first phase NGS data processing is carried out. The second phase takes care of genotyping and variant discovery while third phase performs integrative analysis.

4. Proposed work

Next generation DNA sequencing dataset is used to analyze and characterize genome sequences. The GATK framework is used for experiments. It contains traversals of two kinds namely read-based and locus-based. Read base traversal involves a sequencer read and its data associated in every iteration of traversal. TraverseReads is the read-based traversal supported by GATK. On the other hand, locus-based traversals provide the pileup of the read based at the given locus (given location at DNA sequence), reference ordered data and actual reference base. TraverseLoci is one of the locus-based traversal supported by GATK. In the proposed implementation, the workload is divided into many independent pieces and they are assigned to map function. Then the results of map function are provided to reduce function which in turn produces final result.

In case of TraverseLoci, each base locus is read along with referenced data and reference base before it is passed to an analysis walker. In other words, TraverseLoci reads every data that covers a single base function in the genome. The proposed work is to analyze genome sequences. In every phase both map and reduce functions are involved. The MapReduce over genome appears as shown in Figure 4.

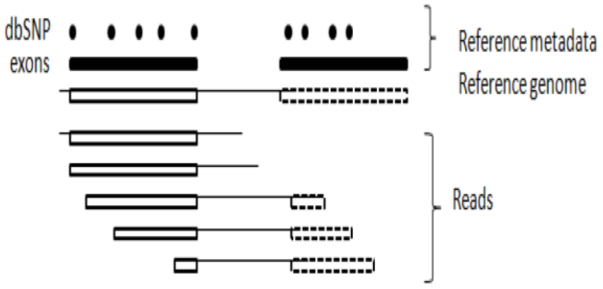

Fig. 4: Map Reduce over genome.

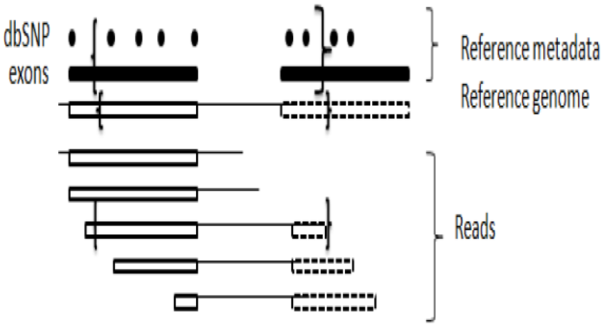

Fig. 5: Mapreduce by Read.

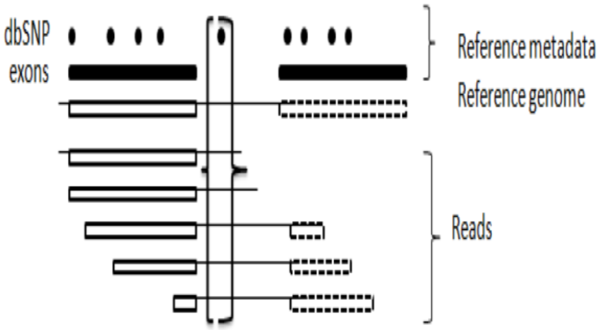

Fig. 6: Mapreduce by Loci.

As shown in Figure 5 and Figure 6 it is evident that read based and locus based traversals are performed using MapReduce programming paradigm. The map function runs once per each locus with three arguments such as tracker, reference and context. Tracker provides metadata, reference base and context are locus-based reference base and a data structure to hold reads respectively. Once mapping is completed, the intermediate results of mapper are given to reducer. In the first step loci value is 0. The reduce function has parameters such as result of map and sum. The final result is the total count of occurrences and sequence loci of all locations where match is found.

5. Experimental results

GATK framework is configured and used in this work to analyze gnome sequences. The results are observed in terms of execution time in seconds. The proposed method is compared with that of TabRec+UnifGen and Indel Realinger approaches. The execution time is observed against the number of virtual cores.

Table 1: Performance Comparison with Different Virtual Cores

	Execution Time (sec)			
No. of Virtual Cores	8	16	32	64
Proposed Approach	175	105	60	50
Indel Realinger	225	130	75	60
TabRec+UnifGen	325	195	140	70

As shown in Table 1, it is evident that the execution time taken for genome analysis is recorded in seconds. The number of cores used in the experiments includes 8, 16, 32, and 64. The observations are presented for different approaches such as TabRec+UnifGen, Indel Realinger and the proposed approach.

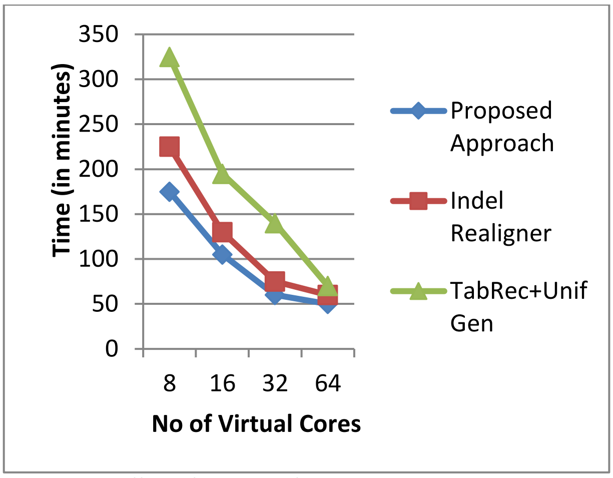

Fig. 7: Effect Of Number Of Virtual Cores On Execution Time.

As shown in Figure 7, it is evident that there are two trends observed in the results. The first trend indicates that the number of virtual cores is influencing the execution time. When number of virtual cores is less, it is taking more time to complete genome analysis. The second trend is that, the proposed mehtod has performnace improvement over other two methods due to systematic approch and the usage of the framework.

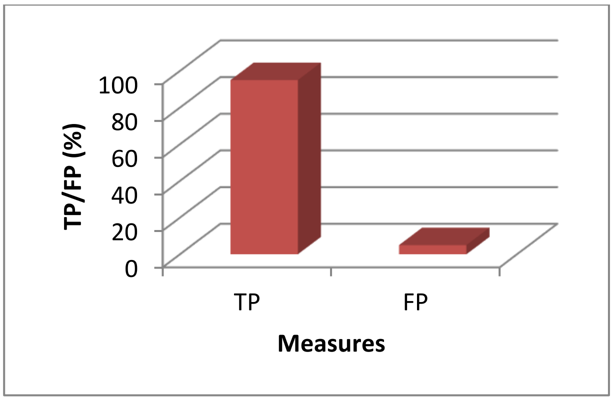

Fig. 8: Evaluation of the proposed method.

The proposed method showed 95% true positive rate and 5% false positive rate. This can effectively show that the method is useful and can be used for genome analysis. True positives are the correctly identified classes with respect to genomes while false positives indicated incorrectly identified classes. The experiment related to analysis is made for 100 times and the results are evaluated to know true positives and false positives.

6. Conclusions and future work

In this paper a distributed framework known as GATK is used along with MapReduce programming paradigm to process given dataset on genome sequences. The framework is used as per a systematic methodology to break the sequences and perform read-based and locus-based traversals. Since genome sequences pertain to big data, Hadoop based tool named GATK is employed to analyze genome sequences. The environment created for the experiments supports parallel processing in order to have better possibilities in analyzing data and produce results. The analysis time is observed with different number of virtual cores used in the experiments and the same is compared with other state of the art approaches found. The results revealed the utility of the proposed method. In future we extend our approach into a framework that can be used to do experiments further on analyzing genome sequences keeping more output variables in mind.

References

[1] Mads Albertsen, Philip Hugenholtz, Adam Skarshewski, Kåre L Nielsen, Gene W Tyson AND Per H Nielsen. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature Biotechnology. 31 (6), p533-542.

[2] Yuqiang Liu, Han Wu, Hong Chen AND Yanling Liu. (2014). A gene cluster encoding lectin receptor kinases confers broad-spectrum and durable insect resistance in rice. Nature Biotechnology, p1-8.

[3] DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-8. PMID: 21478889 (2011).

[4] Laxmi Silwal-Pandit, Hans Kristian Moen Vollan, Suet-Feung Chin, Oscar M. Rueda, Steven McKinney, Tomo Osako, David A. Quigley, Vessela N. Kristensen, Samuel Aparicio. (2014). TP53 mutation spectrum in breast cancer is subtype specific and has distinct prognostic relevance. American Association for Cancer, p1-30.

[5] Alexandra S. Dubrovinaa, Konstantin V. Kiseleva AND Valeriya S. Khristenkoa. (2013). Expression of calcium-dependent protein kinase (CDPK) genes under abiotic stress conditions in wild-growing grapevine Vitis amurensis. Journal of Plant Physiology, p1491-1500.

[6] Simona Soverini, Caterina De Benedittis, Katerina Machova Polakova AND Adela Brouckova. (2013). Unraveling the complexity of tyrosine kinase inhibitor-resistant populations by ultra-deep sequencing of the BCR-ABL kinase domain, p1-37.

[7] Bahareh Rabbani, Mustafa Tekin and Nejat Mahdieh. (2014). the promise of whole-exome sequencing in medical genetics. Journal of Human Genetics, p6-15.

[8] O Abdel-Wahab and A Dey. (2013). The ASXL–BAP1 axis, new factors in myelopoiesis, cancer and epigenetics, p11-15.

[9] Raphael Rytz, Vincent Croset AND Richard Benton. (2013). Ionotropic Receptors (IRs), Chemosensory ionotropic glutamate receptors in Drosophila and beyond. Insect Biochemistry and Molecular Biology, p1-10.

[10] David Bertsch, Jo¨ rg Rau, Marcel R. Eugster, Martina C. Haug, Paul A. Lawson, Christophe Lacroix and Leo Meile. (2013). Listeria fleischmannii sp. nov., isolated from cheese. International Journal of Systematic and Evolutionary Microbiology, p527-532.

[11] Molecular Plant. (2013). Rapid and Efficient Gene Modification in Rice and Brachypodium Using TALENs. .. 6 (4), p1365-1368.

[12] Nicholas C. P. Cross, Daniel Catovsky and Jonathan C. Strefford Gomez, Jade Forster, Helen Parker, Anton Parker, Anne Gardiner, Andrew Collins AND Monica Else,. (2013). the clinical significance of NOTCH1 and SF3B1 mutations in the UK LRF. Bloodjournal.hematologylibrary.org at HEALTH SERVICES, p468-475.

[13] David W. Craig, Joyce A. O'Shaughnessy, Jeffrey A. Kiefer AND et al. (2012). Genome and Transcriptome Sequencing in Prospective Metastatic, p104-118.

[14] Hui Li, Zhiyu Peng, Xiaohong Yang, Weidong Wang, Junjie Fu, Jianhua Wang, Yingjia Han AND Yuchao Chai. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels, 45 (1), p43-52.

[15] Donald B. Smith AND Peter Simmonds. (2014). Consensus proposals for classification of the family Hepeviridae. Journal of General Virology, p2223-2232.

[16] Ronald Taylor. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics, 11(12), 2010.

[17] GATK (2018). Genome Analysis Toolkit. Available online at https://software.broadinstitute.org/gatk/download/ [accessed: 10 Dec 2017].