Supernatural Bigdata Classification of Genome Evolution

Rama Naga Kiran Kumar. K [1], Dr. Ramesh Babu. I [2]

[1]Research Scholar, Dept of Computer Science & Engineering, Acharya Nagarjuna University, Guntur, AP.
[2]Professor, Dept of Computer Science & Engineering, Acharya Nagarjuna University, Guntur, AP.

Abstract: Genome sequencing plays a vital role in the research area to understand the DNA order to discover human genetic secrets. To study the genome sequence capacious data is available in this field.Genome sequence characterization is tedious and non-trivial task. Some of the algorithms are studied in literature study. The genome sequences include big data characteristics, so a technique that includesthe map-reduce paradigm is proposed. Trends in genome sequence can be discovered by various approaches in machine learning. Framework in distributed programming is used to support parallel processing and use Graphical Processing Units. The datasets are managed in cloud to handle with ease. A prototype application is built to describe the concept of proof. The results provide observations during genomic study.

Keywords: Genome sequence, big data, Map Reduce, spectral characterization.

1. Introduction

High-performance computing plays a vital role in analyzing data related to bioinformatics. This includes division of work on a cluster which includes file-sharing systems. The distribution of work and parallelism is implemented through message passing technique. One of the services in cloud computing is IAAS(Infrastructure As A Service) that can help to invoke data and computationally intensive applications. The software framework is installed in LINUX to analyze the data. Hadoop was first introduced by Doug Cutting in 2004. A large number of datasets are included in bioinformatics for parallel processing like Hadoop [16]. Hadoop is more suitable to work with big data which includes genome sequences.

The patterns determined in the genome sequence are used for analyzing in bioinformatics. Information hidden in the sequence can be used for making better decisions in genomic data. The sequences of DNA are analyzed for autocorrelation and auto convolution. The kit that is used to conduct experiments on genome sequence is Genome Analysis Toolkit. The framework for dividing the DNA sequence into small pieces is based on MapReduce programming technique. Section 2 provides the Literature work. experimental setup is discussed in Section 3, proposed work is presented in Section 4, experiments conducted, and the results are shown in Section 5, conclusion in Section 6.

2. Related Work

The Literature on genome sequence analysis is done in this section. Differential coverage binning method which is used to determine genome sequence which is related to bacteria that is hardly available is discussed by Albert et.al[1]. The encoding technique which is used to understand the resistance power in rice is studied by Liu et al.[2]. The mutation spectrum in breast cancer is studied by Silwal-Pandit et.al. 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 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 biosynthesis 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

Several Experiments are conducted in a distributed programming environment to analyze the genome sequence. It is GATK which is based on Map Reduce programming paradigm. In this section we provide the details of Hadoop, GATK, HDFS (Hadoop Distributed File System), Map Reduce. The distributed programming environment which supports Map Reduce programming is Hadoop.

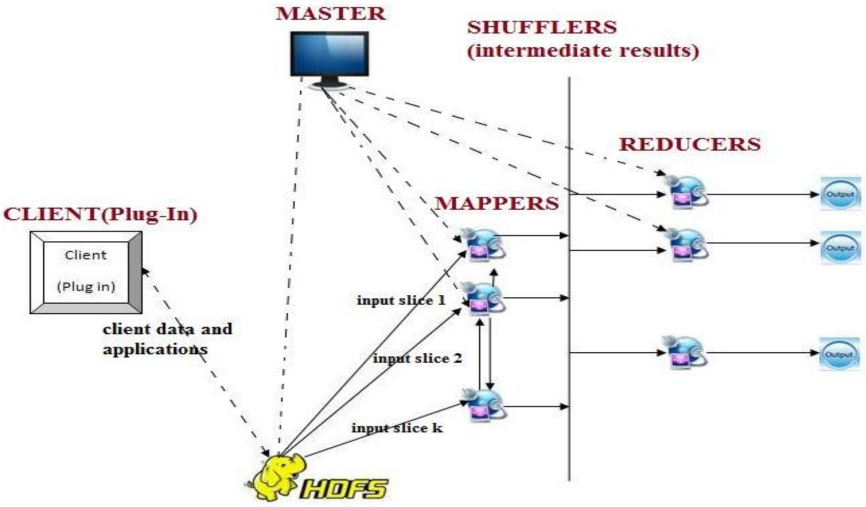

Figure 1: Map Reduce Paradigm Functionality

The above Fig 1 shows the Hadoop Distributed File system (HDFS) which supports the Map Reduce programming paradigm. Itis the file system that is associated with Hadoop framework. The input given is divided into multiple parts and were allotted to different mappers. The output of the mappers is assigned to reducers which gives the final output. So, the intermediate results are given by the mappers which are a worker node along with reducers in distributed environment.

Genome sequences are analyzed by a toolkit named GATK. It is used in various projects that aim to map the nucleotides contained in a human haploid reference genome (more than three billion). In principle, full genome sequencing can provide the raw nucleotide sequence of an individual organism's DNA. However, further analysis must be performed to provide the biological or medical meaning of this sequence, such as how this knowledge can be used to help prevent disease. The framework was constructed by genome sequencing and analysis group from Harvard University and Board Institute of MIT. GATK, pronounced "Gee Ay Tee Kay" (not "Gat-Kay"), stands for Genome Analysis Toolkit. It is a collection of command-line tools for analyzing high-throughput sequencing data with a primary focus on variant discovery. The tools can be used individually or chained together into complete workflows. It is used in Cancer genome Atlas project. The framework of GATK is shown in Fig 2.

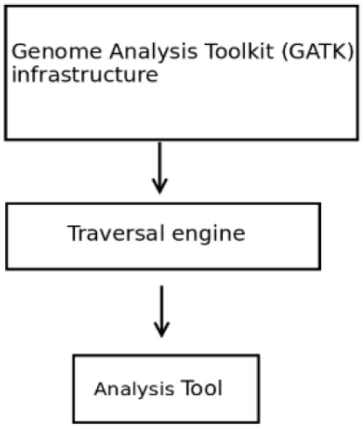

Figure 2: Structure of GATK

This framework is used to analyze different genetic variations. Genetic variation is commonly divided into three main forms:

- Single base-pair substitution, also known as single nucleotide polymorphism (SNP)
- Insertion or deletion, also known as 'indel'
- Structural variation.

The difference between these is based on frequency of occurrence. The variation discovery[3] is determined by a framework as shown in Fig 3.

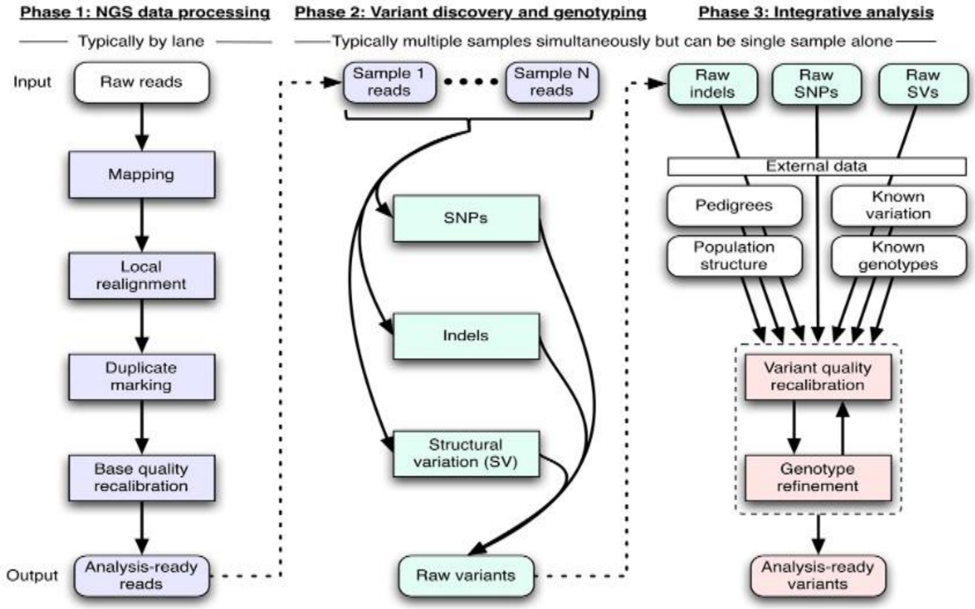

Figure 3: Framework for discovery of variations

The framework includes three phases. The first phase data processing of NGS is carried out. The genotyping and variant discovery is taken place in phase 2 and Interactive analysis in third phase.

4. Proposed Work

To characterize and analyze the genome sequence DNA sequencing dataset is used. The GATK framework which is used to conduct experiments contains two different kinds of traversals

1. Read-based
2. Locus-based

In read-based traversal, it involves a read sequencer to read the data that is involved in each iteration.The read-based traversal that is supported by GATK is Traversereads. The locus-based traversal is based on reading each position in genome . The locus based traversal that is supported by GATK is Traverse Loci. In proposed work the work is partitioned into different pieces and is assigned to the map function. The results are further provided to reduce function that finally produces the result. In TraverseLoci each base locus is read with its reference base along with its associated reference points and then forwarded to analysis walker. Proposed work involves analyzing the genome sequence. Each phase includes map function along with reducing function as shown in Fig 4.

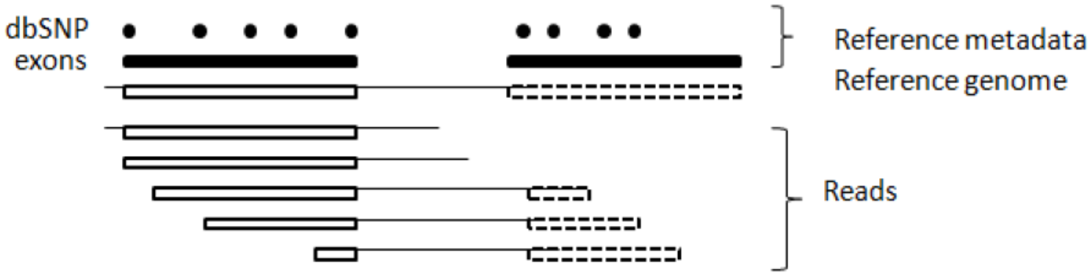

Figure 4: MapReduce over the genome

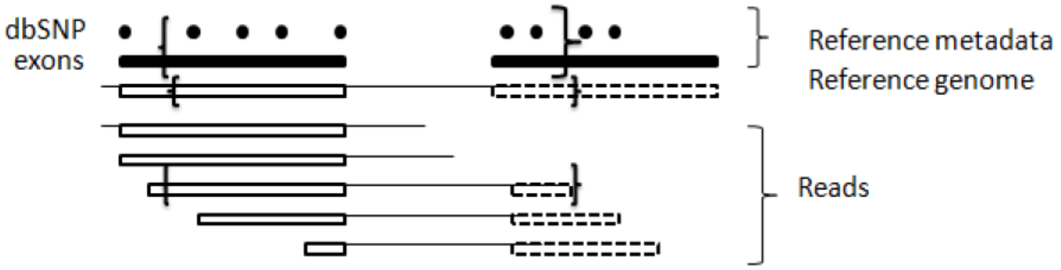

Figure 5: MapReduce by reading

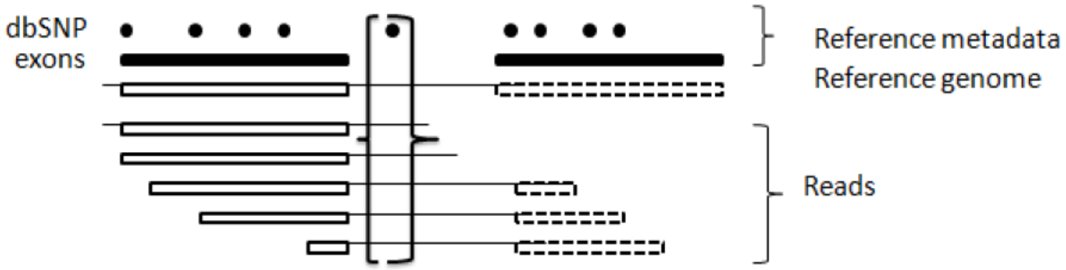

Figure 6: MapReduce by Loci

The read and locus based traversals are conducted by using map-reduce paradigm as visualized in Fig5 and Fig 6. A map function is performed for each locus based on three parameters like tracker, context, and reference. Metadata that is data about the data, context base and read base along with data structures are provided by tracker. Once the mapping procedure is completed the results are given to the reducer. Initially, the loci value is 0. The map and sum parameters are included in reducing function. The final result gives the total occurrences count and sequence of loci at all locations where they are actually matched.

5. Experimental Results

The genome sequence is analyzed and configured by GATK framework. Results are determined based on the execution time in seconds. The comparison of proposed method with other methods such as TabRec+UnifGen and IndelRealinger are shown in Table 1. The total number of cores used for evaluation are 8,16,32,64.

	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

Table 1: Comparision with different virtual cores

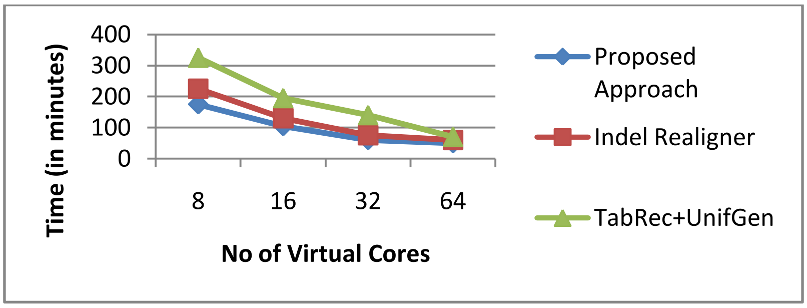

Figure 7: Effect of the number of virtual cores on execution time

Two observations are viewed in the results as shown in Fig 7. In the first observation, we see that execution time is effected by total number of virtual cores.The genome sequence analysis takes more time when the virtual cores are less.The other observation is the proposed method has improvement over other two methods.

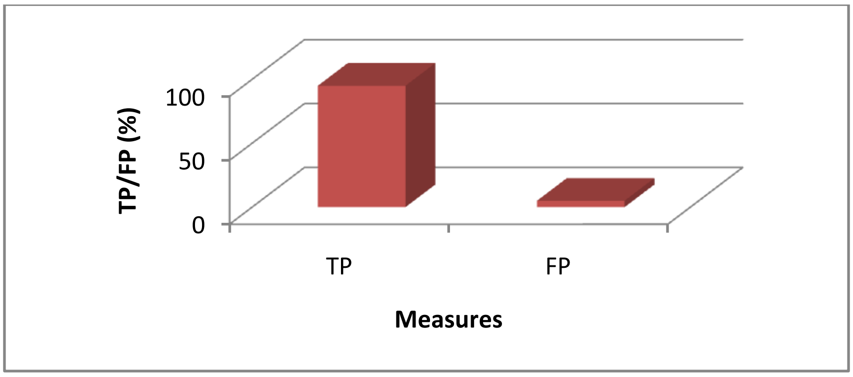

Figure 8: Evaluation of the proposed method

We can see that 95% truepositive rate and a 5% false-positive rate in proposed method. So, the proposed method is used for genome analysis. True positive include the correctly identified classes on the other hand false positive rate determine the classes that are identified incorrectly.

6. Conclusion and Future Work

GATK is the distributed framework that is used with map-reduce technique to process the dataset which is associated with given genome sequence is discussed in this paper. The framework includes a method that divides the sequence into small pieces and performs read-based and locus-based traversals. The data that is involved in genome sequence is more we use Hadoop based tool GATK. The experimental environment support parallel processing used to analyze data and provide results. The results of proposed method, when compared to other methods like TabRec+UnifGen and IndelRealinger, are better. This approach can be used to analyze genome sequences by considering various output variables in future.

References

[1]. MadsAlbertsen, 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]. LaxmiSilwal-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 Vitisamurensis. *Journal of Plant Physiology*, p1491-1500.

[6]. Simona Soverini, Caterina De Benedittis, Katerina MachovaPolakova 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]