An Imminent Approach for Genome Sequence and Analysis using Map Reduce

C.J. Kavithapriya
Assistant Professor,
Jeppiaar Institute of Technology,
Sriperumbudur,

ABSTRACT

The recent trend of BigData in Healthcare is overpowering and necessity increasing rapidly because of its data type diversity in addition to its volume, managing speed and leads to improving care even at the lowest cost. Cancer prevails as a challenging issue because of its different mutations. Identification of the each tumor's root for mutations and mapping of their evolution of genetics that leads to growth in the conflict against the cancer disease, "GenomeAnalysis "plays an important role. In order to accumulate and categorize the enormous revenue of information from genome analysis, research field coalesced with a data Platform ApacheHadoop supporting parallelization, composability for extremely huge upsurge in activity of sequencing data. By aggregating all aids of BigData Analytics Tools and EHR, this proposal presents a study about how to incorporate the Hadoop Tool integrated with GATK(Genome Analysis Tool Kit) through MapReduce to map cancer genomic data problems with the conscious of financially low cost and high speed of accessing data.

General Terms

Big Data Using Hadoop Map Reduce Analytics

Keywords

BigData, EHR, Genome Analysis, ApacheHadoop, MapReduce, GATK.

1. INTRODUCTION

Nowadays, every organization has been working with huge amount of data which leads to the new concept of "Digitized Universe" in which the usage of the data was expected to reach 2.7 Zeta Bytes (ZB) by 2012 and also reaches above 8ZB by 2015. In past years, these health care information were stored in the hard copy, but now the evolution of "Digitization of Universe" paved a way for Electronic Health Records (EHR) which stores Peta Bytes (PB) of health information of patients and can reduce hospital readmissions to data mining rules that look at patient charts from previous visits to fill gaps in current charts.

1.1. Big Data in Genomics

1.1.1. Cancer Genome

Almost all the cells of the body contain a complete copy of the instructions which are needed to make every tissue, cell and organ in the body. These set of instructions has been coined as "Genome" which is comprised of 3 billion of 4 letters as "A", "C", "T", and "G" as shown in Fig. 1.2. Also it has been divided into 23 volumes called Chromosomes. Each chromosome contains sentence of instruction that are used to instruct the cell how to make protein for the development of the organs. The major benefits such as detection of diseases at initial stages of their occurrence when they can be preserved more simply and effectively; also manages the managing definite individual health and detect their health care of a specific individual and detection of their health care deception more rapidly and efficiently. In order to manage these huge data as shown in Fig. 1.1 the help of Big Data Analytics has been influenced in which it addresses the major issues as Velocity, Volume and Variety.

1.2. Big Data Analytics in Medicine

Traditionally, large amounts of data has been generated by the industry of healthcare, which had been driven by he following such as record keeping, regulatory requirements, compliance and patient care [8]. When most of the data had been stored in the form of hard copy, but the recent trend is concerning about the digitization of the generation of huge amounts of data much rapidly. Normally, DNA of the Human consists of about 3 billion base pairs each personal genome that has been representing roughly about 100 gigabytes (GB) of data, which corresponds to about 102,400 photos. The estimation of capacity of global sequence was about 13 quadrillion bases and also counting which is considered to be more than enough data for a stack of DVD to be filled by 2011. The total count of transistors that are placed on an IC(Integrated Circuit) board is exponentially increasing, with a time that is doubling for 18 months roughly [3]. Similar phenomena have been noted for the capacity of hard disks which is given by Kryder's Law [11] and network bandwidth given by the Nielsen's Law and Butter's Law [5]. This trend had been true for more than 40 years, until the completion of Human Genome Projects at the year of 2003. As the hospitals have been digitized the records of patients and huge amount of data revolved to the top companies such as Microsoft, SAS,IBM (IBM), , Dell (DELL), and Oracle(ORCL) since expertise in data-mining, which would help to the providers of medical field and also performs the work of detective and additionally improving the care for the people. According to research from the organization of IDC, the concept of Big Data business has been already infused the other industries and more than $30 billion of data has been generated in the last year's revenue and also it has been expected the growth to be about $34 billion in the present year due to the increased usage in the industry of health care. The private healthcare data is critical to big data's success, it doesn't mean that your private data will become public, these suggestions has been stressed by Spradlin. But this genome are "Dynamic" which changes under the circumstances of X- Rays, sunlight, chemicals which results in an abnormal growth of the tissues and leads to cancer.

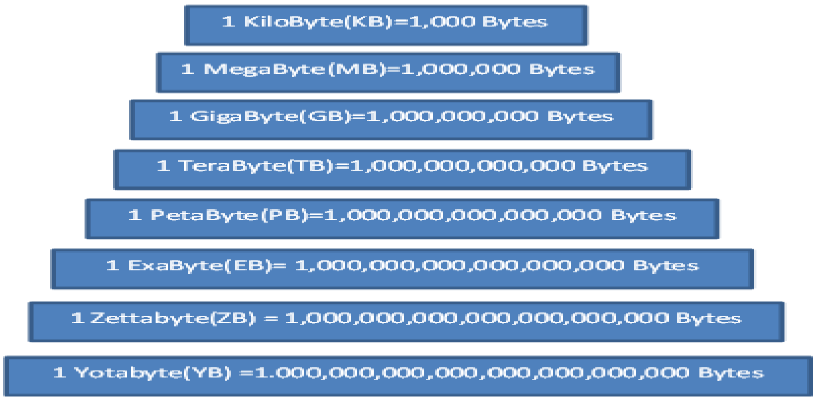

1 KiloByte(KB)=1,000 Bytes
1 MegaByte(MB)=1,000,000 Bytes
1 GigaByte(GB)=1,000,000,000 Bytes
1 TeraByte(TB)=1,000,000,000,000 Bytes
1 PetaByte(PB)=1,000,000,000,000,000 Bytes
1 ExaByte(EB)= 1,000,000,000,000,000,000 Bytes
1 Zettabyte(ZB) = 1,000,000,000,000,000,000,000 Bytes
1 Yotabyte(YB) =1.000,000,000,000,000,000,000,000 Bytes

Fig. 1.1. This Shows the figure which represents the hierarchy of the data formats. Starting from KiloBytes to YotaByte in terms of number of Bytes

Hence, as the first step in Genome Project, the genomes have been sequenced by researchers for analyzing and diagnosing the diseases. In order to provide a detailed description about cancer genome, the changes occur in it, different types of data being generated, where they can be accessed and various technologies, tools, methods for analyzing gene functions has been discussed in this work. Variations of nucleotide sequence include the genome alterations which causes cancer. In addition, the major known somatic alterations in genome include the causes such as nucleotide mutations, chromosomal rearrangement and nucleic acid from origin of foreign cells.

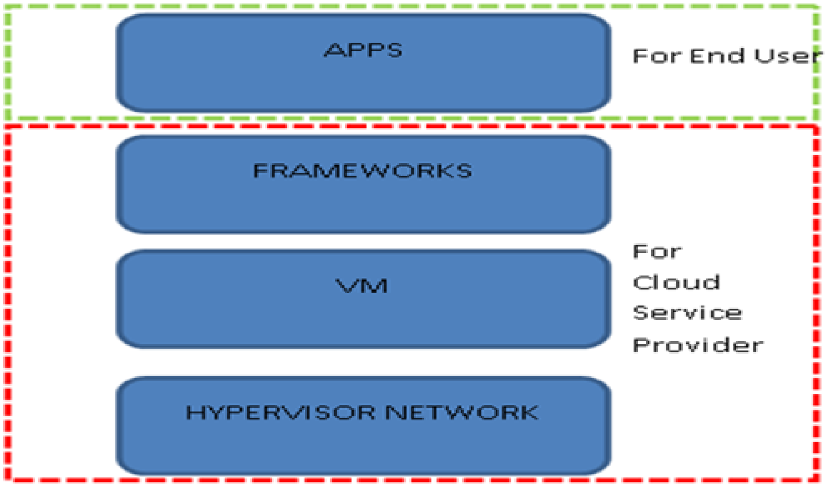


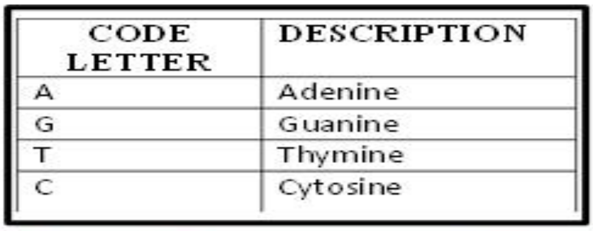
Fig. 1.2. Represents the Complete Gene with their Codes inside the genome and describes about the description of corresponding codes in genome.

1.2.1. Alterations in Cancer Genome

Global gene encompasses full transcriptone which includes coding messenger RNAs and noncoding micro RNA of complex tumor tissues reflects an array of somatic and epigenomic alterations as discussed in [13]. Also, together with the state of cell differentiation of tumor and admixture of non- cancerous cells will be analyzed and transcriptional profiling defines a unique gene expression for each tumor for successful classification.

1.2.2. Moore's Law and Biological Data

As per Moore's Law the number of transistors that are placed on an IC(Integrated Circuit) board which increases doubling rate exponentially for every 18 months. This phenomenon can be simply put as computers increase their speed in double and decrease their size in for every 18 months. Similarly, nowadays the usage of hard disks and network bandwidth has been increased in huge amount. As a major part of the genome projects, the sequencing of deluge of Biological data has been generated and spurred by the decreased cost. For understanding biological pathways and genomic variation the National Cancer Institute sequenced million of genomes with the practical proof of that the sample of whole genome and the matching normal tissue consumes about 1million TB of data that is uncompressed which is equivalent to 1000 Petabyte or 1 Exabyte.



CODE LETTER	DESCRIPTION
A	Adenine
G	Guanine
T	Thymine
C	Cytosine

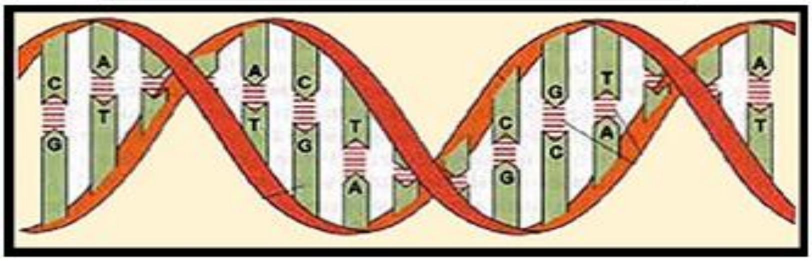


Fig. 1.3. This figure depicts the Platform As Service, which is among one of the category of cloud computing model.

1.2.3. Utility Computing to Big Data

Nowadays, the generation of biological data has outpaced the Moore's Law by its increased storage space and growth. In addition, the biological data sets are more expensive to store, process and analyze. This explosion of data has been proven by International Data Corporation (IDC) that the data in worldwide reaches about 0.8 ZB in 2009 and predicted to reach upto 40 ZB by the year of 2020 [5]. For accessing these data sets for analysis, the users can hire the infrastructure on the basis of "Pay as you Go" which avoids large capital infrastructure and maintenance cost. This hiring of infrastructure concept has been called as Cloud or Utility Computing [6] with virtualization using user friendly web interfaces. But this solution of virtualization alone will not be the end solution. So, the raise of Big Data deviates from traditional structured data and represent in semi-structured data. Additionally, the big data is often referred as "PaaS" Platform as a Service within the cloud environment as shown in Fig. 1.3. Even though many technologies has been utilized for managing the dataset, the bio-informatics based applications are difficult to develop due to the lack in documentation and require many programming library dependencies [12].

2. TECHNOLOGIES FOR GENOME SEQUENCING

All these second generation methods of sequencing [2] involve the processes of amplifying individual DNA molecules on array. These second generation technologies allow to identify the simultaneous alterations and also identify the mutations even in a admixed samples by means of deep coverage. In addition to the identification of mutations, the structural information from other genomic platforms has been offered. Thus enables the global assessment of chromosomal rearrangements in cancer. In future, nearly to all aspects of cancer genome characterization has been sequenced by applying the sequencing based approaches. By deep sequencing the whole genomes in 10% of the sample, the rate has given as increase in sequencing capacity and a linear shrink in costs [10].

2.1 Hadoop and Genomics

The emerging Big Data Technologies are not only transforming the technical domain in all fields but also it is saving many of lives. Thus, the concept of big data analytics have being bringing up in most of biological data sets. The issues involved in the present biological data sets are not in the part of storing but in the part of analyzing the data set. The predecessors in Map Reduce projects have been applied in biotechnology space which lead to Genome Analysis Tool Kit as discussed in [4]. Followed by this, the antecessor of the Map Reduce Projects developed the noval technology named Hadoop which is discovered as revolutionary methods in order to manage the big data. By using the substantial sorting tool Hadoop, the samples of genome and genotyping has been ordered the alignments by the implementation of SoapSNP. These primary adopted projects have evolved only for the person who is technically persuasive. Personalized medicine trial for analyzing and controlling pediatric cancer had been initially approved by first FD. Then unambiguously in PaaS space, the collaboration of NextBio has been announced by Intel@ in order to optimize HDFS (Hadoop Distributed File System), Hadoop, the other technology HBase (Distributed Database) for analyzing the genome data. In this epoch of genome research the analysis of human and bacterial genomes had been included, also it includes study of metabolic pathways of both the categories such as normal and also disease affected states of an organism [3]. The health care organization note the trade-offs in terms of cost, scalability. On completion of this Concept Statement the methodology is approved. On the completion of the entire module of first step, the process is proceeded to the next step "Proposal Development Stage" which triggers out more questions based on the concept statement. The following questions have been framed What problem to be addressed?, Why it is important to the health care provider?,

How much interested by the health care provider or not? Why the Big data Analytics has been approached. In *Stage 2* of the process the project team should provide the background information about the problem domains. Next, the process proceeded to the *Stage 3* where the methodology has been inspected and implemented. To help the big data analytic process, the concept statement of the stage 1 has been broken into a series of propositions. Also, from these different propositions, the dependent and independent variables has been identified. The important aspect of this stage is

Table 1. Table consisting of the stages and the module description involved in the process of analyzing the health care data using map reduce.

TRAVERSAL TYPES	**DESCRIPTION**
TraverseReads	Read with associated reference bases given to analysis walkr
TraverseLoci	Single base locus in genome, associated read, reference ordered data, reference base given to analyzer.
TraverseLocusWindows	Read, reference ordered data, reference bases for whole interval of genome are given to walker.
TraverseDuplicates	List of duplicate reads, unique read at reference locus given to the analyzer.

evaluation of the platform/ tool which is explained in Fig. 1.3. After the evaluation by various data sources, the various big data analytics techniques has been implemented to the data and various insights are gained upon various multiple iterations. From these insights the decision can be done on the data as in Table 1. In *Stage 4,* testing of models and their findings has been done, validated and presented to the stack holders. At each stage, the feedback loops has been incorporated to minimize the risk of the failure. The final section of the stage describes about several big data analytics applications in the health care [11].

3. IMPLEMENTATION OF GENOME ANALYSIS TOOL KIT

The Platform-Independent programming language Java 1.6 framework is the base of the GATK development environment. Standard sequence Alignment/Map (SAM) format has being used by the core system to represent reads using a production-quality SAM library. The major functions of the SAM Java development kit are (i) handles parsing the sequencer reads, (ii) provides ways to query for reads the span specific genomic regions. The binary alignment/map (BAM), which is a binary alignment version of SAM format, is compressed and indexed, also used by the GATK for producing better performance with the following reasons such as due to its smaller size and ability to be indexed for search. The major activities of the core system are (i) can accommodate reads from any sequencing platform, (ii) conversion to BAM format, (iii) sorting on read coordinate order. This core system has been extensively tested on Illumina , on the Applied Bio systems SOLiD System [2] test has been conducted , also to test the 454 Life Sciences and also the Complete Genomics[8]. The BAM files with alignments emitted from next- generation sequence aligners which had been tested with huge number os BAMs that re aligned by using a variety of publicly available alignment tools, these BAM files are accepted and supported by the core system. Also, it supports common emerging SNP formats like [GLF, VCF], GELI text format, public database formats like Hap Map, dbSNP variation databases which are discussed in the forums of broad institute.

Table 2. Table denoting the different types of the traversal types defined in the GATK.

STAGE No.	**MODULE**	**DESCRIPTION**
1	CONCEPT STATEMENT	*Need for Big Data Analytics based on V's established*
2	PROJECT PROPOSAL DEVELOPMENT	*QuestionsFramed.*
3	IMPLEMENTATION	*Propositions,variable selection done,collection of data,ETL and data transformation selection of tool/ platform, developing conceptual model,different analytic techniques, association,clustering, classification, insights for decision making*
4	TESTING AND VALIDATION	*Deployment, Testing, Evaluation*

3.1 Architecture of GATK

The major designing of this tool is based on the functional programing paradigm of Map Reduce [7] the concept in Hadoop.the analysis tools have been constructed in order to easily parallelize and distribute processing of underlying framework.

1. Larger data sets have been divided into distinct independent fragments which are given into the region of map function.

2. Later the results from the map function is fed into reduce function which joins the results of map function back to Final product. Chromati immune precipitation (ChIP) exploits this reduce function for integrating the heights of pileups read across Loci for the detection of the transcriptional regulation[8].

3.2 Read Based and Locus Based Traversals

The collection of common data representations are provided by GATK which are named as “traversals” as mentioned in Table2. The locus based traversal is the most commonly used traversal which reads every single base position in genome by covering the regions completely without any missing of the region. It involves all associated genomic data such as genomic location, variation data, associated interval and specific locus in genome as shown in Fig. 4.1. These references are passed to the walkers map function. Computation over sequencer read pile up such as the depth of coverage computations, genotyping concerned with variant analysis. In the analyzer the read based traversal type data as shown in Fig. 4.2 are passed to the walker function only once. In addition to the sequencer read, the reference bases that the read overlaps are also overlapped in the walker function. Also, the other process such as sharding, interval processing, merging input files, parallelization, data collection and processing as discussed in [1].

4. EXPERIMENTAL RESULTS

4.1 Bayesian Estimation

Bayesian estimation plays an important role in predicting genotypes. This enhanced framework would serve both as an initiator for more advanced tool for statistical interference and shared memory highlighter and also triggers out the maximum distributed capabilities of parallelization of GATK The computation of posterior probability from (1) of each genotype by reading current locus and expected heterozygosity of the data samples of genome. This computation done by Bayesian is given by

$$p(G \mid D) = \frac{p(G)\,p(D \mid G)}{p(D)} \qquad \textbf{(5.1)}$$

The term p(G) is the prior probability estimation influenced by homozygous reference. p(D) representing constant overall genotypes where D represents read base pileup data and G representing the given genotype. In order to cover the target locus the constant b and the prior probability which had been introduced and defined as

$$p(b \mid G) = p(b \mid \{A_1, A_2\}) = \frac{1}{2}\,p(b \mid A_1) + \frac{1}{2}\,p(b \mid A_2) \; \textbf{(5.2)}$$

Where the genotype G={A_a ,A_2} is decomposed into two of its alleles(alternative form of dame genes) as given in Eq.(5.2). The probability of getting the base given an alternative form of gene is given as

$$p(b \mid A) = \begin{cases} \frac{e}{3} : b \neq A \\ 1 - e : b = A \end{cases} \qquad \textbf{(5.3)}$$

Where e is the epsilon term, the reversed phred scaled quality score at the base. The genotype with largest prior probability are emitted to disk if the set threshold exceeded by the log-odds as defined in Eq.(5.3).. This naïve genotyper has the performance level of identifying 315,202 variants that were upon the given chromosome with 81.70% of dbsnp a concordance of 89.76% as discussed in [1]. The comparison of single genotyping efforts with the concordance level values for individual of 86.4 percentage and 99.6 percentageas shown in Fig. 5.2and the comparison of the distributed data and the processor count is shown in Fig 3.3 in comparison with the elapsed time to determine the performance of the genotype

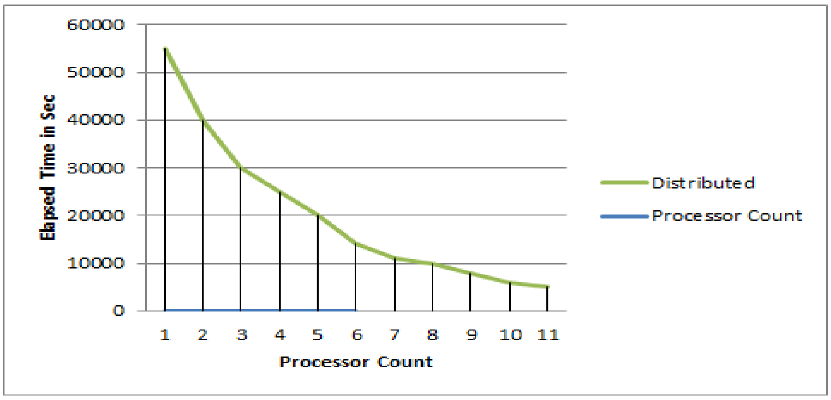


Fig. 5.1. Figure depicting the graph taken for Genome 1000 project sample NA12878s chromosome genotype using both distributed and parallelization processor count.

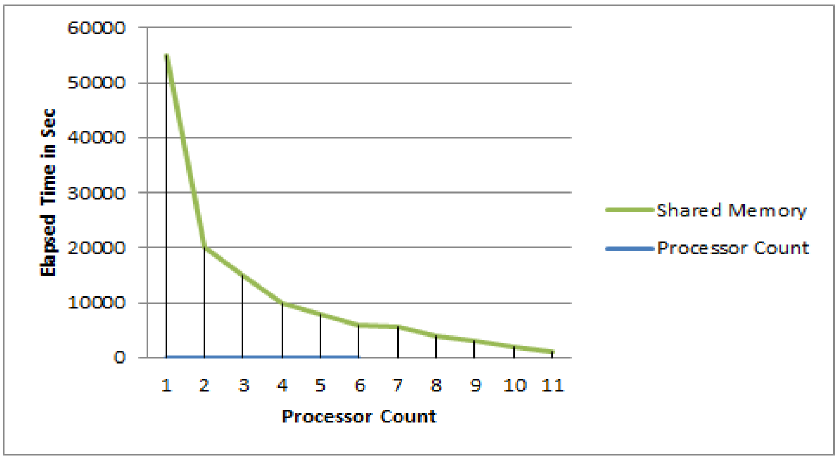


Fig.5.2. Figure depicting the graph taken for Genome 1000 project sample NA12878s chromosome genotype using both shared memory and parallelization processor count.

5. CONCLUSIONS AND CHALLENGES

As an emerging technology for analyzing the larger data generated, the Big Data Analytics has been used and in that the Map Reducer plays an important role. In the medical field, this analyzer has a vital role in sequencing the Genome and determines whether it's a normal genome or the cancer affected genome. In order to access the massive next generation sequencing data from the logic, the GATK's Map reduce architecture is used to separate the complex infrastructure. This proposal discusses in detail about how to use the big data analytics in genome sequencing and how to include the simple Bayesian genotyper in the combination of GATK. Also, it discusses about how to sequence the genome data set and ho to determine its features using parallelization. The biggest challenge that has been contributed to the future researcher is to analyze the genome for the normal and the cancer affected genome by using the map reducer and GATK as an emerging tool for cancer genome identification.

6. REFERENCES

[1] AaronMcKenna1, MatthewHanna1, Eric Banks1, Andrey Sivachenko1KristianCibulskis1, AndrewKernytsky1, KiranGarimella1, David Altshuler1,2,Stacey Gabriel1, Mark Daly1,2 and Mark A. DePristo1,3. Genome Res. 2010. 20:1297-1303,(2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data.

[2] Biosciences and Illumina MiSeq sequencers. BMC Genomics 2012;13:341.

[3] "Cloudera Chief Scientist Jeff Hammerbacher Teams with Mount Sinai School ofMedicine to Solve Medical Challenges Using Big Data." <http://www.marketwire.com/press-release/Cloudera-Chief-Scientist-Jeff- EMC Sitting In Sweet Spot Of $70 Billion Big DataIndustry.<http://www.forbes.com/sites/greatspeculations/2011/11/18/emc-sitting-in-sweetspot-of-70-billion-big-data-industry/>.

[4] Dai L, Gao X, Guo Y, Xiao J, Zhang Z. Bioinformatics clouds for big datamanipulation. Biology Direct 2012,7:43.

[5] Gantz J, Reinsel, D. (2012) The Digital Universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. In: IDC iView: IDC Analyze the, Future.

[6] Healthcare Cloud Computing (Clinical, EMR, SaaS, Private, Public, Hybrid) Market – Global Trends, Challenges, Opportunities & Forecasts (2012–2017).<http://www.reportlinker.com/p0924631-summary/Healthcare-Cloud-Computing-Clinical-EMR-SaaS-Private-Public-Hybrid-Market-Global-Trends-Challenges-Opportunities-Forecasts-.html.

[7] How Hadoop Makes Short Work of Big Data. <http://www.forbes.com/sites/netapp/2012/09/24/hadoop-big-data/>.Taylor RC. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinform 2010;11(Suppl. 12):S1.

[8] Lynda Chin,1,2,3 William C. Hahn,1,2 Gad Getz,2 and Matthew Meyerson1,2, (2015) "Making sense of cancer genomic data", Cold Spring Harbor Laboratory Press.

[9] Obama Administration Unveils ''Big Data'' Initiative: Announces $200 Million In New R&DInvestments.<http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf>.

[10] Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific.

[11] Raghupathi W, (2010) Data Mining in Health Care. In Healthcare Informatics:Improving Efficiency and Productivity. Edited by Kudyba S. Taylor & Francis,:211–223.

[12] Shachak A, Shuval K, Fine S. (2007) Barriers and enablers to the acceptance of bioinformatics tools: a qualitative study. J Med Libr Assoc;95:454–8.

[13] Yeh RF, Lim LP, Burge CB. (2001) Computational

inference of homologous genestructures in the human genome. Genome Res11:803–16.