**BIG DATA GENOMICS SEQUENCING**

****

**Team Name : Illuminati**

**CMPE 272-02/Fall 2014**

**By**

**Mansi Shah,** [**manasishah88@gmail.com**](mailto:manasishah88@gmail.com)

**Meghana Gogi,** [**meghna1888@gmail.com**](mailto:meghna1888@gmail.com)

**Roshan Srinivas,** [**roshans.4488@gmail.com**](mailto:roshans.4488@gmail.com)

**Rashmi Hadli,** [**hadli.rashmi@gmail.com**](mailto:hadli.rashmi@gmail.com)

**ABSTRACT**

**Big Data For Genomics Sequencing**

**By**

**Mansi Shah**

**Meghana Gogi**

**Roshan Srinivas**

**Rashmi Hadli**

The next generation sequencing methods has changed the perspective of life science research by introducing various low cost sequencing methods. The more-than-one-thousand-fold improvement in throughput and cost has spawned a new era of biomedical research, where sequencing is performed for many individuals and studied over the course of a single project.

Human genomes are almost identical (about 99.9%). But the key to understand the cause of diseases and its treatment is identifying this minute difference among them. The human genome forms 3 billion base pairs. Genome sequencing technology has intensely improved and the number of base pairs generated by modern sequencing techniques has been increasing rapidly. There are numerous sequencing instruments in hospitals and laboratories all over the world which can sequence 15 quadrillion nucleotides per year which equals about 15 petabytes of genetic data. We are dealing with a very huge set of data in genomics. The computation and storage resources needed for this is the biggest challenge for data analyses in a local environment. With the unmatched scale at which genomic sequence data is generated there is a need to properly interpret the high dimensional data sets and adopt high performance, scalable methods of data handling and management. The focus of our project is identifying genetic mutations from whole genomic sequencing data, mainly SNPs (single nucleotide polymorphisms). Genetic variations are identified by aligning all the short reads to a human reference genome and then SNP calls are done. As the growth in throughput of short read sequencing vastly surpasses the improvements in microprocessor speed, there is a very important need to accelerate the commonly used tasks like short read alignment and SNP calling. This can be achieved by large scale parallelization. Our solution is a Hadoop-based parallel pipeline for genomic analysis, for searching SNPs using cloud computing. It uses Bowtie for alignment and SOAPsnp for SNP calling. It runs in a parallel pipeline in cloud using Amazon EC2 making analysis in cloud scalable, robust and cost effective.

Table of Contents

**1. Introduction** 5

**1.1** **What is genome?** 5

**1.2** **What is genome sequencing?** 6

**1.3 Challenges with traditional Genome Sequencing Approach** 8

**1.3.1 High processing cost for sequencing** 8

**1.3.2 High processing time for sequencing** 9

**1.3.3 Accuracy of Reads** 10

**1.3.4 Data storage of genome base pairs.** 10

**1.3.5 Scalability was difficult in genome sequencing** 12

**2.** **Problem Statement**: 12

**2.1 Problem definition** 12

**3.** **Big Data for genomics** 14

**3.1** **Overview** 14

**5. Architecture** 15

**5.1 Hadoop Distributed File System** 16

**5.2 Map Reduce** 17

**5.3 Short Read Alignment and SNP Discovery using BowTie and SOAPsnp** 18

**5.4 CROSSBOW** 19

**5.5 Design** 21

**5.6 Implementation** 23

**6. Analysis of solution based on various Performance Metrics** 24

**6.1 Cost:** 24

**6.2 Time:** 25

**6.3 Accuracy:** 25

**6.4 Storage:** 25

**6.5 Scalability:** 26

**7. Conclusion** 27

**8. References** 27

**Table of Figures**

Figure 1 : Human Genome 6

Figure 2: Genome sequencing. 8

Figure 3: Cost comparison for various periods. 9

Figure 4: Accuracy comparison of Next Generation Methods. 11

Figure 5 : Developing the cost of storing of one byte data versus sequencing base pairs. 12

Figure 6 : k-mer algorithm to reduce the data 12

Figure 7 : The physical cost of sequencing 15

Figure 8: Big Data in genomics 15

Figure 9: HDFC architecture 17

Figure 10:MapReduce Flow 18

Figure 11:Burrows-Wheeler Transformation 19

Figure 12: Design 21

Figure 13:Genome sequencing data analysis steps 22

Figure 14: Sequencing data importing into S3 23

# 

# **1. Introduction**

## **What is genome?**

Human body is made up of about one hundred million of cells and each of these cells has a complete set of instruction which is required for the growth and development of an organism. All these instruction set makes a genome. Genome contains set of chromosomes, which in turn holds gene which is made of chemical DNA.

DNA structure of each animal differs and contains all the information about the species and history of its family. In a human body, each cell with nucleus has more than 3 billion DNA base pairs. In order to store this information and patterns, we use various techniques. DNA sample is made up of four chemicals called adenine (A), guanine (G), thymine (T) and cytosine (C). These A, G, T and C, the first letter of each chemical represents an organism DNA. All leaving organisms will have their own set of genomes. The information present in human genome is divided into 23 volumes called chromosomes. Each chromosome contains genes which represents the genetic information of the organism. All together there are 20,500 number of genes will be present in genome.

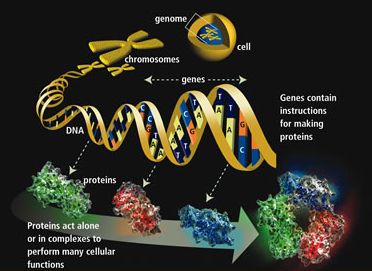


Figure 1 : Human Genome

Human genome is not static, genomes in the cells changes if it is exposed to sunlight, chemicals , x-rays etc, thus the affected cell which is called as tumor cells are responsible for spreading a diseases, which will transfer it to the surrounding tissues and spreads it across the other parts of the body.

Human Genome Project came into picture during 1990 which was responsible in finding the order of As, Cs, Ts and Gs sequences. This project concluded in the year 2003; however the scope of the project was up to sequencing. Research on tools and technologies is still in progress to improve in this field. Here in this report we will be discussing on the tools and technologies that supports genome sequencing.

## **What is genome sequencing?**

In genome sequencing, a large DNA module is broken apart in to fragments and then these fragments are sequenced. The overlapping sequences of these genomes (if any), are again assembled to create a long string of these letters. These long set of strings help the researchers to understand, how these letters work together to build an organism.

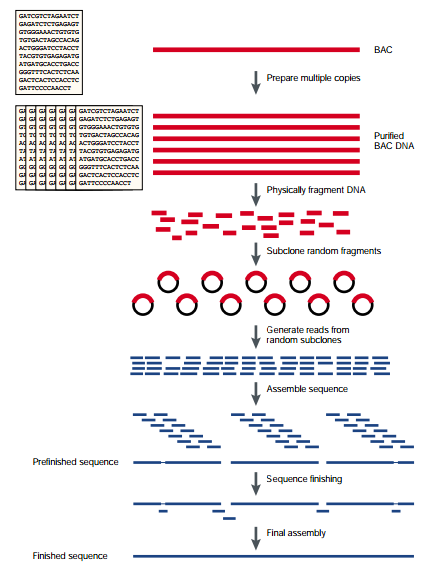


Figure 2: Genome sequencing.

In above figure, we can see how the genome sequencing is done and the final sequence has been developed by overlapping the matching sequences. This final sequence produced is highly accurate to find the correct order of the sequencing. Here we use the “Clone-by-Clone” approach. Firstly, genomes are broken into relatively large chunks, called clones, about 150000 Base Pairs (bp), and using genome mapping techniques each clone belongs can be determined. Next each clone can be broken into even smaller clone and the overlapping sequences will be used to reconstruct the sequence of the clone. Better understanding the human genome will tell us a lot about how life works. It could lead to preventing or curing diseases, because genetics is what getting sick is all about -- our genes are trying to fight off the genes of a virus or bacteria. The next step will be to determine how this battle is played out. Today, researchers know the positions of some genes that control our medical traits. Other genes have been located but their functions are unknown, and still others remain entirely elusive. The point of genome research is to locate the genes and determine just how the four bases are sequenced, and then to learn what the genes actually do which opens the door to all sorts of options. Sequencing is the best way for understanding the genome. It is the shortcut which helps scientists to find the abnormal genes more easily and quickly. The resulted abnormal genes are the clue for the scientists to analyze on the diseases. There are various ways for sequencing the genome.

## **1.3 Challenges with traditional Genome Sequencing Approach**

### **1.3.1 High processing cost for sequencing**

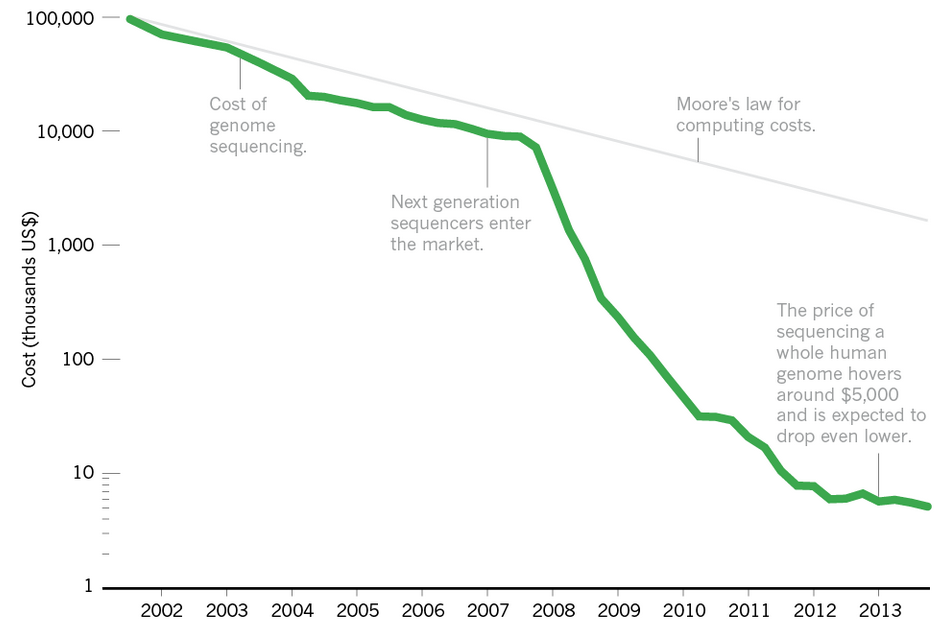
****

Figure 3: Cost comparison for various periods.

In the above figure, it is described how the cost for genome sequencing is reduced drastically. The genome sequencing to find the abnormal genes was invented after the end of the Human Genome Project, the cost for processing 1000 genomes of an individual was very high, during the period of 2006-2007 the next generation sequencers (NGS) entered into market, the principle behind NGS technology is – the bases of the smaller fragments of DNA are sequentially identified from the signals emitted from fragments and re-synthesized from the DNA fragments. This helped in the deduction of the cost in this field. Followed by the Moore’s law during the year 2007, the cost for DNA sequencing got reduced exponentially. As per this law the cost will reduce drastically for every two year. At present we have $1000 genome project, S1000 genome project refers to personalized medicine during which the cost of the full genome sequencing of an individual will be reduced to $1000. Still research is going on $1000 genome project to reduce the cost of sequencing.

### 

### **1.3.2 High processing time for sequencing**

As we all humans have 20500 genes and it should be compared with each other to check the abnormal DNA sequence of letters. Shape of the DNA is called as “Double Helix” with two strands. Each point in ladder is made of chemicals called (A), thymine (T), cytosine (C) and guanine (G). A is matched with T, and C is matched with G, variations in the base pairs is what makes the person unique from each other. Chromosomes are made up of genes which in turn made up of DNA. Each individual will have 23 pairs of chromosomes – 11 pairs are inherited from mom and 11 pairs is inherited from dad. 23rd pair is responsible for sex identification. These chromosomes will be chosen for diseases detection, if chromosome is missing in the person, it leads in medical problems. Unpaired or abnormal sequenced chromosome is responsible for the health of an individual. DNA is covered by the membrane layers, sample of saliva or blood needs to collected and the using some chemicals this membrane will be broken down. This sample will be used to compare with the 3 billion base pairs by using some hygienic machines to make up a person’s genetic code. This part of sequencing is time consuming and this will take couple of months to sequence a person’s genome. Checking the abnormal order of DNA codes is time consuming, which in turn compared with the genetic code of many people’s DNA. Human Genome project is started at 1990, during this time it was thought that the sequencing of genome would take 15 years. But during the year 2003 the sequencer was released, there was few draw backs in it. Although, the estimated cost was $3 billion for human genome project. Later on, the sequencer was released which was handling for $ 2.8 million base pairs per 24 hour period. First the experiment is done on the smallest organism called bacteria. At this rate, 3 years was taken by Sager sequencer to sequence the human genome. In 1990s, Paul Nyren developed Pyrosequencing method which refers to sequencing by synthesis, each time when the nucleotide is incorporated the enzyme reactions will results in light signal that is read by the analyzer. The future optimization is done by 454 life science by using the Polymerase Chain Reaction. In this process, many copies of each reaction is made via PCR called microbeads, are then kept on to special plate, the light reaction is the analyzed using Genome Sequencer FLX analyzer. Genome Sequencer FLX analyzer can sequence 100 billion base pairs per 8 hours, this instrument can run 300 million base pairs in per 24 hours. At this rate, it took 10 days to sequence the human genome. Later on various techniques has be invented and at present the sequencing is taking 24 hour period time and still the research is on the field to reduce the time. During the Next generation sequencing they even target to the initial sequencing and use filers to select the relevant variants after sequencing. But this process hinders the diagnosis and still the time consuming was more in this process. In next few years, experts expect to reduce the time which is consuming for sequences. Nanopores, Nanogaps including National Human Genome Research Institute (NHGRI) groups is working on reducing the time and cost for sequencing.

### **1.3.3 Accuracy of Reads**

The Human genome project was introduced on April 2003, roughly it was decided to finish by June 2000 and by February 2001 but only draft was completed by this time. Finally on April 14, 2003 the final sequencing of human genome got over and the initially accuracy for detecting the mutations was 92%. After this project various others group are able to reach the accuracy of 99.99%. In the below table you can see the comparison of the accuracy on various projects. Still the investigation is going on to reach the 100% accuracy level in various projects.

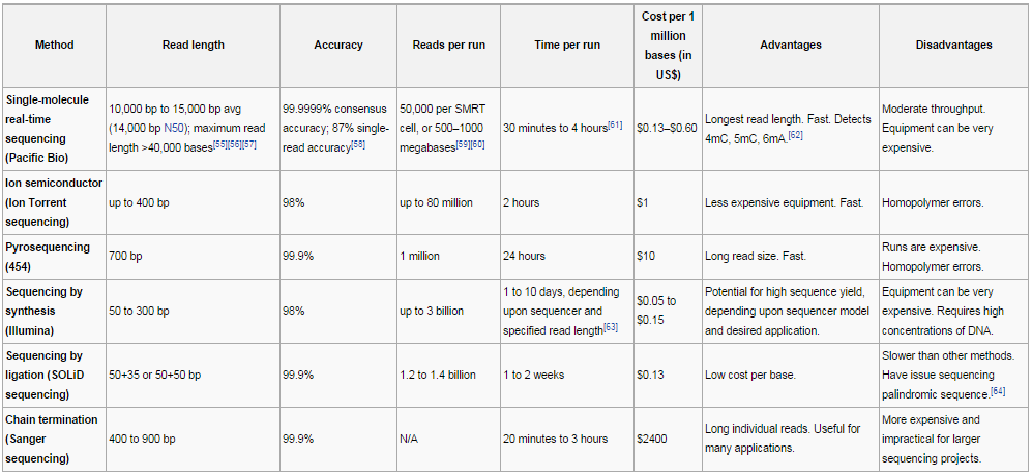


Figure 4: Accuracy comparison of Next Generation Methods.

### 

### **1.3.4 Data storage of genome base pairs.**

Data storage cost occupies the largest proportion of the total cost of analysis and sequencing of DNA. Increase in the DNA sequencing is significantly outstripping the disk storage capacity. Result of this various compressing methods has been discovered to compress and store the DNA sequence data. Sanger sequencing method which is array based is adopted in 2005, followed by new sequencing methods in which the sequencing of data increased sequentially. This results in the 1000 fold drop of the cost but the number of the data of sequencing increased; it includes the 1000 genome project as well. Next is the highly fragmented data during “next generation” project. This was in turn leading the cost of the sequencing to increase drastically. However this data was initially unprocessed and later it was decided to make it processed data in order to reduce the space for storage and the maintenance cost. Even few raw data repository has been invented to keep this compressed data which in turn decide to store in the form of raw image containing sequential data. Later it was decided to store only analysis approach and the informative data instead of all compressed data. Various compressing tools and various reduce algorithms has been invented to manage the data storage. Research is still going on managing the sequencing data storage.

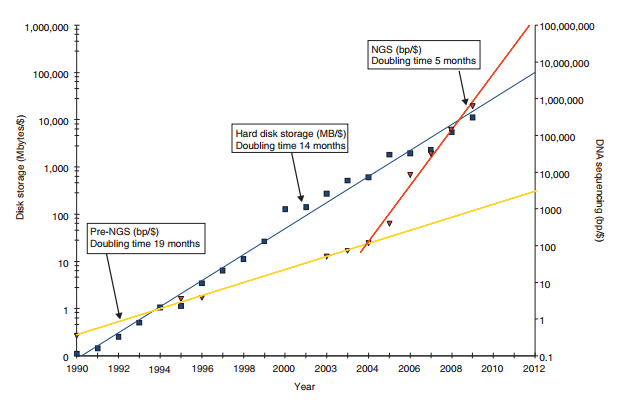


Figure 5 : Developing the cost of storing of one byte data versus sequencing base pairs.

In the above graph, it shows clearly the cost of sequencing of data is reducing compared to sequenced data. Later on various map reduce algorithms has been implemented to reduce the data. For instance, K-mers is also algorithm invented in this to reduce the data storage.

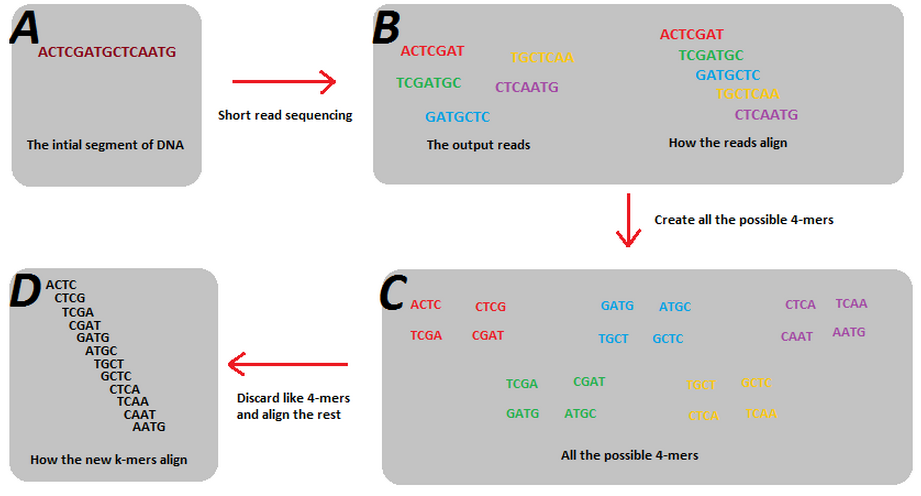


Figure 6 : k-mer algorithm to reduce the data

### **1.3.5 Scalability was difficult in genome sequencing**

Initially Sanger sequencing and fluorescence based technologies have been extremely used for genome sequencing, improvement on the instruments resulted in development of high performance computing and the sequencing of data has been reduced. But Sanger also need the support additional parallel equipments because of high throughput value, this scalability is then reduced by using massively parallel sequencing methods. This can be reduced either by creating the micro-reactors or beads, it will read the millions of sequence reactions in parallel, massively parallel sequencing even provides the smaller reads. Storage and maintenance of the reads will be easy from this, genome reactions is captured in parallel and used to reduce the sequences, and thus the storage space is reduced. Various private companies like Facebook, Google, Microsoft and Twitter are dealing with data sets to provide the high scalable approaches using the latest technologies like Bigdata, Hadoop and more.

# **Problem Statement**:

To build fast, scalable and a cost effective system that provides a highly paralleled batch processing capability for sequencing genome for multiple users.

## **2.1 Problem definition**

There are many different sequencing methods: the basic methods like Maxam-Gilbert

Sequencing, Chain-termination methods, and advanced methods and de novo sequencing likeShotgun sequencing and Bridge PCR. The current state of the art method in genome sequencing is different ‘Next generation sequencing’ methods.

The Next generation sequencing technologies, parallelize the sequencing process, producing thousands or millions of sequences concurrently. High-throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods. In ultra-high-throughput sequencing as many as 500,000 sequencing-by-synthesis operations may be run in parallel.

There are many NGS Technologies like Single-molecule real-time sequencing (Pacific Bio), Ion semiconductor (Ion Torrent sequencing), Pyro sequencing (454) Sequencing by synthesis (Illumina), Sequencing by ligation (SOLiD sequencing), Chain termination (Sanger sequencing).

The rapid evolution in the Next-Generation Sequencing has led to the complete transformation of the genomic research. The Size and the number of datasets have dramatically increased. Therefore, modern biology now presents new challenges in terms of data management, query and analysis. The human genome is about 3 billion DNA nucleotides (characters), about the same as the English portion of the Wikipedia. Storing or searching one genome by itself is not too difficult, and standard tools are quite efficient for searching it on a single computer. However, because of the limitations of DNA sequencing technology, we cannot simply read an entire genome end-to-end. Instead the machine reports a very large number of tiny fragments called read, each 25-500 letters long, collected from random locations in the genome. Then, much like how raindrops will eventually cover the whole sidewalk, we can sequence an entire genome by sequencing many billions of reads, with 20-fold to 30-fold oversampling to ensure each nucleotide is seen. Presently, this process generates about 100GB of compressed data (read sequences and associated quality scores) for one human genome.

Once collected, we can map the billions of reads to the reference human genome using sequence alignment algorithms, and then scan the alignments to find differences between the newly sequenced genome and the reference genome. Again, the problem of mapping and scanning 100GB of data isn’t too onerous; especially for large sequencing centers with large compute grids, and recent studies of individual sequenced genomes have been able to do the analysis in about 1000 CPU hours of computation.

However, the technical improvements that decreased sequencing costs to a greater extent. Small to midsize research groups are producing more sequencing data due to its lower costs. This trend is likely to continue as the newer generation technologies keep driving the costs down. The rapid increase in the volume of data has enabled quick adoption of Whole-Genome Sequencing to enhance research and development, which lead to the need for computational methods and more scalable tools. Clearly we need very efficient and scalable methods especially, as sequencing moves from large sequencing centers, to smaller research centers, and perhaps eventually to hospitals and clinical labs. For a WGS consisting of 50 subjects, 20 terabytes of disk space are required to store the raw reads alone. The data storage and CPU resources needed pose a huge practical challenge for data analyses in a local environment.

The physical costs of sequencing are dropping:

2007 -$1M

2009 - $100k

2010- $5k-10k

2012- $1k (estimated)

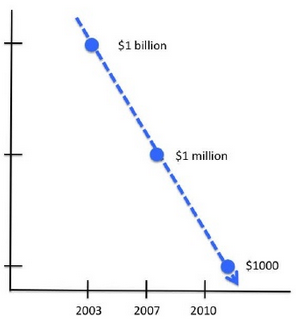


Figure 7 : The physical cost of sequencing

Lower cost drive an increase in sequencing, activity and data produced. Computation for sequence assembly and analysis is becoming a bottleneck.

# **Big Data for genomics**

## **Overview**

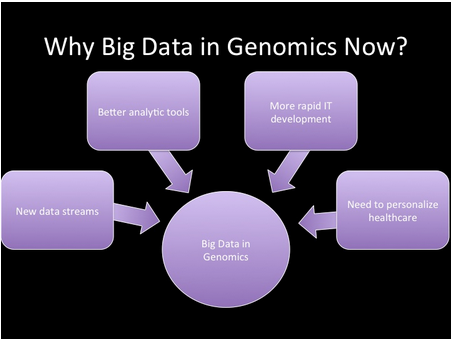


Figure 8: Big Data in genomics

Big data being new generation of technologies and architectures, it is designed to extract value from large volumes of variety of data. Most of the Big data surge is unstructured data and is typically not suited for traditional databases to analyse it.As genomics has very large, semi-structured, file-based data and is modelled on post-process streaming data access and I/O patterns that can be parallelized, it is ideally suited for Hadoop The Hadoop system allows us to perform very custom analysis that you wouldn’t find in a traditional business intelligence tool or that would work in a SQL\* relational type of structure.

The Hadoop Map Reduce framework runs on the compute cluster using the data stored on the HDFS (Hadoop Distributed File System). MapReduce jobs aim to provide a key/value based processing ability in a highly parallelized fashion. Since the data is distributed over the cluster, a MapReduce job can be split up to run many parallel processes over the data stored on the cluster. The map parts of MapReduce only run on the data they can see – that is the data blocks on the particular machine its running on. The Reduce brings together the output from the Maps. Our goal is to build a system that provides a highly paralleled batch processing capability which scales well with the multiple users. This opens the very real possibility of personalised medicine; the ability to treat individuals on a case by case basis, tailored to their speciﬁc genomic blueprint.

# **5. Architecture**

The application uses BowTie for short read alignment, SOAPsnp for SNP calling, SNP merger for consolidating SNP and is built using Hadoop MapReduce framework.

## **5.1 Hadoop Distributed File System**

Hadoop Distributed File System (HDFS) is a Java based distributed file system that is designed to span large clusters of commodity hardware. HDFS consists of a namenode along with metadata and several datanodes. The client’s request is sent to namenode which looks up the requested data block in the metadata. Name node looks up the information in metadata and sends the addresses of data nodes and block information to client.

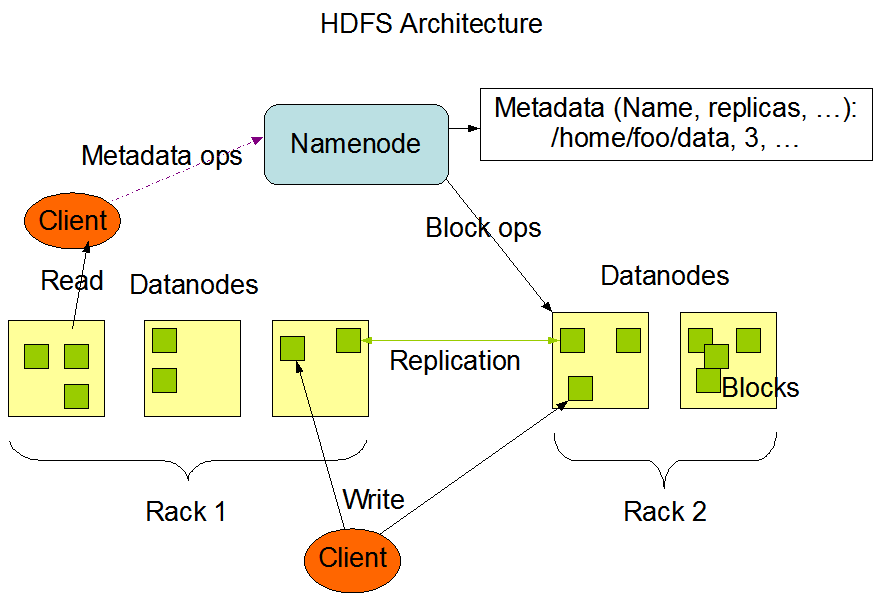


Figure 9: HDFC architecture

HDFS has a master/slave architecture. Internally the HDFS File is split into one or more data blocks and distribute the block amongst the nodes in the cluster. The block size being 64Mb or 128MB by default. An HDFS cluster has single Name Node (Master) and multiple instances of DataNode. An HDFS file system namespace operations like opening, closing, and renaming files

And directories are executed by the NameNode. The Metadata information like the current location of the block on the DataNodes, the list of files belonging to each block, the state of the file and the information about the access control are all managed by the NameNode.Its an arbitrary and repository for the metadata information. The design of the system is such that the user data never flows through the NameNode.

Usually, there is one DataNode running on each node in the cluster that manage storage attached to the nodes. The responsibility of the DataNode is to serve read and write requests from the HDFS file System’s clients. Upon the instruction from the Name Node, the Data Node also perform block replica creation, replication, and deletion.

Name Node and Data Node are the pieces of the software. Since HDFS is built using Java Language, Name Node and Data Node software can be run on any machine that supports Java. HDFS can be deployed on a wide range of machines as Java is highly portable language.

Features of HDFS:

* Highly fault tolerant
* Low Cost Hardware
* High Throughput
* Scalable
* Portable File­System
* Its suitable for the application where large Data­sets are involved
* The application requires streaming access to their data sets

## **5.2 Map Reduce**

MapReduce consists of two separate tasks, the Map job and the Reduce job. The input for MapReduce comes from files which are loaded to a processing cluster in HDFS. The input files are distributed evenly across all nodes. The mapping task will be running on all nodes within the cluster and all tasks are equivalent which makes each mapper equally capable of performing the mapping job. The map function reads the data and processes it to key value pairs. When all mappers complete their job, MapReduce will start allocating the values to the reducers according to the keys. All the reducer tasks will also be on the same nodes as mappers. The reducer’s job is to shuffle and sort keys and run reduce function on them. Finally the reduce function outputs key value pairs which can be stored in distributed file system and can be reused by any other MapReduce job. In HDFS, all data are scattered and MapReduce tasks run closer to the data. MapReduce programming model is ideal when processing of large data sets is involved in a distributed file system using parallel algorithm.

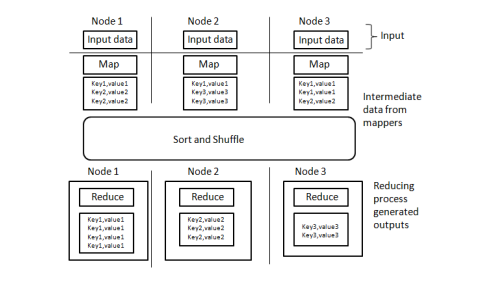


Figure 10:MapReduce Flow

## **5.3 Short Read Alignment and SNP Discovery using BowTie and SOAPsnp**

Next generation sequence generators generate lots of short sequences of DNA. The ability to analyse and derive useful information with those short sequences depends on the ability to align those sequences to a reference sequence. Alignment is very hard because the reference genome is very large and looking for exact matches results in a very large number of possibilities which cannot be solved by brute force algorithms. But there are many algorithms which can quicken the process. Dynamic programming used in modern computer science can be used for this problem. But even this is not efficient for large data sets. Indexing is required to make it feasible.

Bowtie algorithm aligns the many short DNA reads in a memory efficient and ultrafast manner by adding Burrow-Wheeler index which keeps the size of the index around 2.2 GB for human genome ensuring the memory taken is very small. In Burrows-Wheeler transform a series of characters are transformed into runs of similar characters. This is an efficient compression technique as it becomes easier to compress a series which has runs of repeated characters. The process is reversible without having to save any additional data.

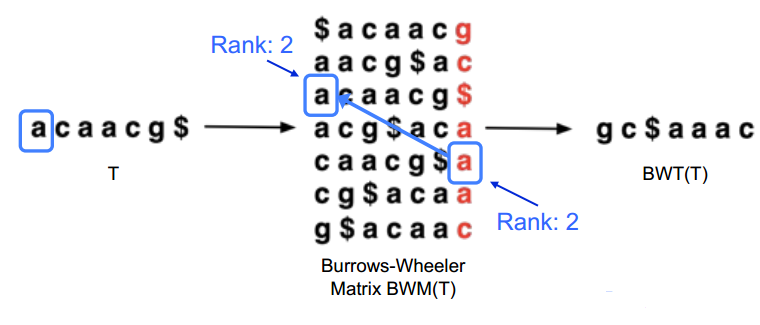


Figure 11:Burrows-Wheeler Transformation

Steps of algorithm:

1. Trim off the bases which are of very low quality from ends of sequences

2. Execute DFS (depth-first-search) of the suffix tree represented by Burrow Wheeler Transform (BWT)

3. Finally report all the best ‘n’ alignments

1. Best alignments are one having the fewest mismatches against edit distance
2. Some reads might have millions of mapping positions which are equally good.
3. If some of the reads are paired, mapping must be found which satisfies both.

Single Nucleotide Polymorphisms (SNP) discovery is the process of measurement of genetic variations of SNPs between different members of species. SNPs are one of the most common kinds of genetic variation. It’s a very important technique and has many uses as variations in DNA sequence of humans can determine how humans develop different diseases and how they react to vaccines, pathogens and drugs. This helps in providing personalized healthcare solutions.

SOAPsnp is a member of the SOAP (Short Oligonucleotide Analysis Package). It’s a sequencing utility which can be used to assemble consensus sequence for a newly sequenced individual based genome. This is done by alignment of raw sequencing reads on the known reference. The SNPs can then be discovered on the consensus sequence by comparing with the known reference. SOAPsnp uses Bayes’ theorem which is a reverse probability model for calling consensus genotype by considering alignment, quality of data, experimental errors.

## **5.4 CROSSBOW**

Crossbow is a tool that employs Bowtie and SOAPsnp .Also, it is highly scalable, automatic and portable cloud computing to find the SNPs from short read data.

Crossbow can be run on

1. In the Cloud
2. On a Hadoop-Cluster
3. On a single Computer

Crossbow can be run in four different ways.

|  |  |  |  |
| --- | --- | --- | --- |
| **Crossbow Web-Interface** | **Amazon Elastic MapReduce** | **On Hadoop-Cluster** | **On single Computer** |
| In this case, the [Crossbow](http://bowtie-bio.sf.net/crossbow) and the user interface are installed on EC2 web servers | In this case, the Crossbow is hosted by Amazon | In this case, the Crossbow is hosted on [Hadoop](http://hadoop.apache.org/) cluster,supporting tools: [Bowtie](http://bowtie-bio.sf.net/), [SOAPsnp](http://soap.genomics.org.cn/soapsnp.html). | In this case, the Crossbow and all supporting tools ([Bowtie](http://bowtie-bio.sf.net/), [SOAPsnp](http://soap.genomics.org.cn/soapsnp.html), and must be installed on the computer running Crossbow. |
| The computers running the Crossbow computation are rented from Amazon, and the user must have required amazon accounts and must pay for the resources used. | The computers running the Crossbow computation are rented from Amazon. | Crossbow is compatible with latest Hadoop version(2.5.2) | The user specifies the maximum number of CPUs that Crossbow should use at a time. This mode does not require [Java](http://java.sun.com/) or [Hadoop](http://hadoop.apache.org/). |

**Table 1: Crossbow execution in different environments**

However, crossbow has a number of limitations when applied to multiple subjects for genome sequencing. The data storage and CPU resources that are required for genome sequencing data analyses are very large for many core facilities and individual laboratories to provide. To help meet these challenges, we designed the architecture that can assist in the automation of genome sequencing data analyses.

## **5.5 Design**

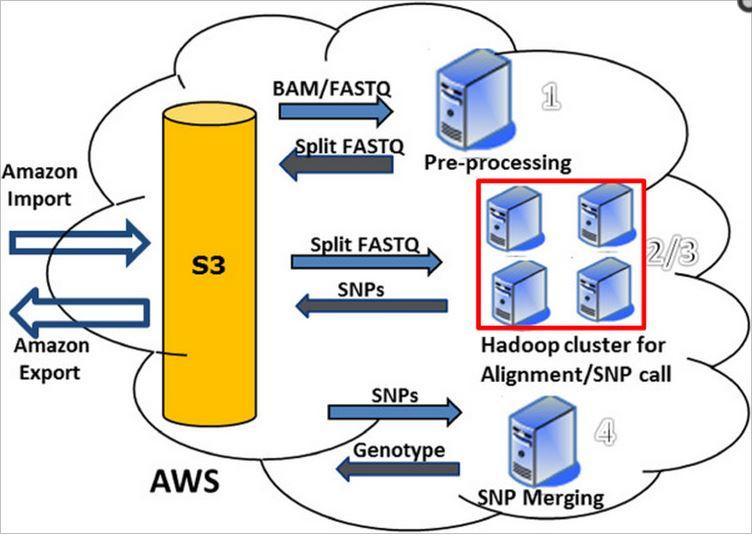


Figure 12: Design

The Steps to perform analyses are:

1. Import
2. Execute
3. Export

Alignment and SNP call are performed in a cluster with multiple nodes.

Output from next generation sequencers which are raw reads in the form of BAM files are preprocessed and are sent to short read aligner. Short read alignment and SNP  calling  process form a  series  of  Map and  Reduce  jobs. Short read alignment is the  Map  job  and SNP calling is the Reduce job and it is implemented using Hadoop MapReduce as a Hadoop cluster.

This entire sequence of operations for mapping each individual read to a reference genome can be run for multiple instances in different clusters in a parallel pipeline as reads, read alignment and SNP calling are independent of one another. SNP integration consolidates all the different SNP calls. In local environment this setup has many challenges because of limited computational and storage resources and for parallel execution. Cloud computing is necessary for such large scale analysis on large data sets.

For cloud services, Amazon Elastic Compute Cloud (EC2), Amazon Simple Storage Service (S3) are used. Amazon Web Services was chosen as the cloud provider because to implement parallel pipeline, different Hadoop clusters can be hosted on multiple instances of EC2. And by using Amazon S3, data storage, inputs, intermediate results, and outputs are all centralized .Computation and data security is guaranteed because of redundancy and multiple copies of same data. All in all, AWS provides virtually limitless computational and storage resources offering a computational environment which is best suited for our application.

**The four main steps in Genome Sequencing data analysis:**

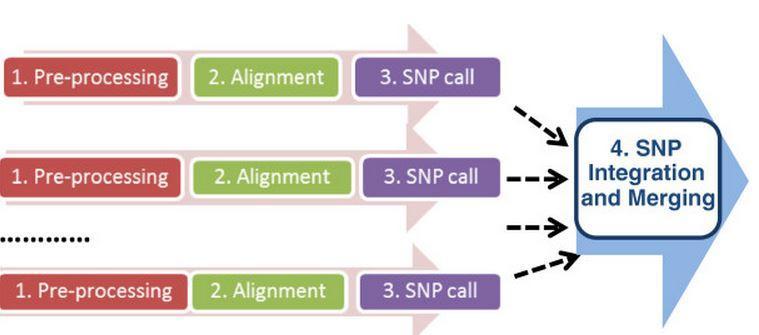


Figure 13:Genome sequencing data analysis steps

Step 1: Pre-Processing

Step 2: Alignment

Step 3: SNP Call

Step 4: SNP Integration and Merging

Steps 1, 2, 3 are all executed in parallel in the cloud by launching multiple instances or clusters.

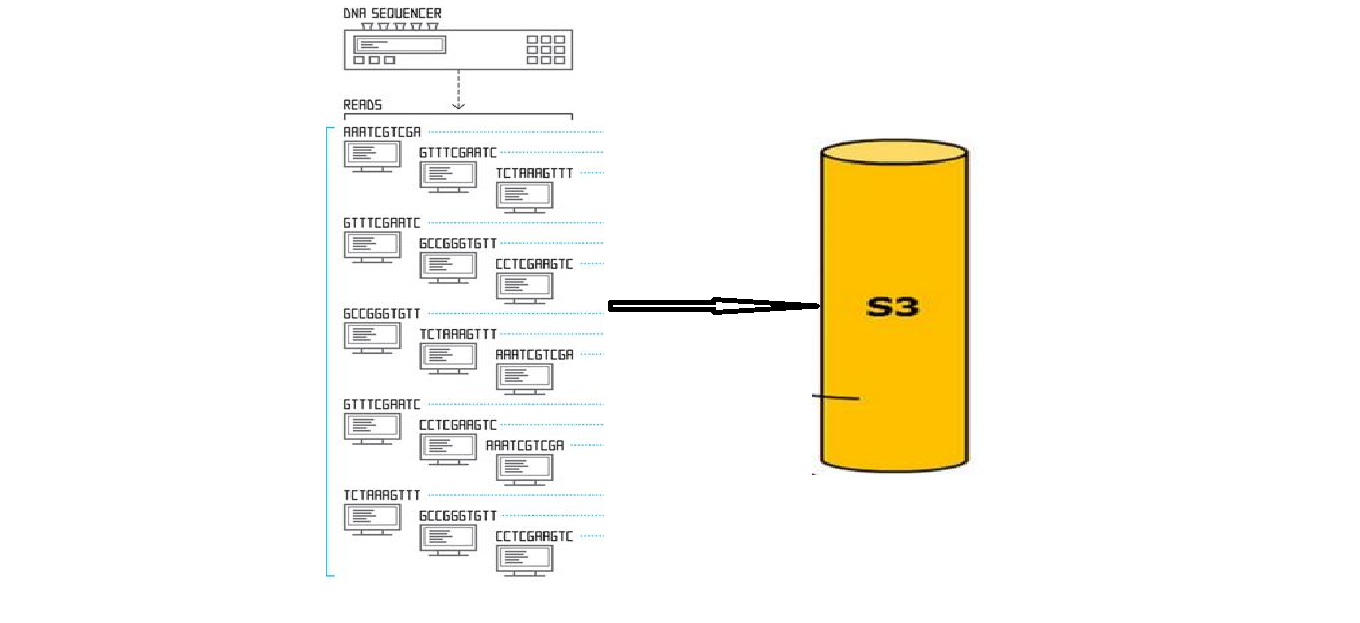


Figure 14: Sequencing data importing into S3

Initially the whole genome sequencing data from sequencers are loaded into S3 by doing Amazon import operation.

In the Preprocessing step, raw reads from all the BAM files are extracted and input file are prepared for the short read alignment and further steps.

Short read alignment is performed by the Map job using BowTie algorithm. Once the alignment is completed the sort job sorts it along the forward reference genome strand.

SNP calling is done by the Reduce job using SOAPsnp. SNP calls from different workflows are finally consolidated by SNP integration block.

## **5.6 Implementation**

The main challenges for genome sequencing analysis include

* Data storage
* CPU resources.

A single sample has 1~2 billion of short sequence reads and requires 1~ 2TB data storage space. The workflow is optimized for practical usage and runs in a semi-automatic fashion. The key building blocks in the workflow are shown in the diagram above of which, Steps #1, #2 and #3 are extremely resource-demanding.

**Step: 1 Pre-Processing Step (BAM->Fastq)**

This step extracts raw reads from BAM files and prepares input files for further steps. Alternatively, it can also receive paired-end FASTQ files (uncompressed or compressed with GZIP) as input and processes those accordingly in preparation for Steps #2/#3. Its main functionalities are:

* Automatically launch an Amazon EC2 instance, request additional resources (such as storage) and configure instance, install required software on instance.
* Fetch data from S3, and perform the computational tasks
* Split big Fastq files into small pieces, and prepare manifest file
* Automatically log running metrics such as data storage and computational time, log progress, including successes, errors and failures and tear down all the requested resources

**Steps 2 and 3: Reads mapping/Alignment and SNP calls:**

In fact, the mapping of each individual read onto a human reference genome is independent of all other reads, and therefore, read mapping (as well as SNP call) can be done in parallel. However, the popular alignment algorithms, such as Bowtie, BWA, SOAP, etc., are not implemented for a cloud environment. Fortunately, Crossbow which is a Hadoop-based is implemented on EC2 Instance and is thus perfect for step 2 and 3.

**Step: 4 SNP Integration/merging:**

The default output from SOAPsnp has its own specific format and an additional filtering is needed to get rid of low-quality SNPs. For large-scale genome sequencing, samples might come from different groups (such as disease vs. normal, responders vs. non-responders, etc.), and those SNPs for the same group need to be grouped together. The SNP call results for all samples and to generate a simple integrated text file needs are consolidated.

# **6. Analysis of solution based on various Performance Metrics**

## **6.1 Cost:**

Genome sequencing analysis is a big data problem, as it requires large number of computing resources to deduce important and useful results from large datasets. Single genome analyses can take significant amount of time i.e. from several days to years on a single core processor. And purchasing these resources is not cost effective when its requirement is periodic. This solution addresses the challenge of limited access to computing resources by offering a cloud computing solution that allows genome analysis pipelines to execute efficiently and in a cost effective manner using Amazon’s elastic cloud infrastructure. The primary advantage of these cloud services is the on demand access to computing resources and relinquishing it when not in use. The estimated cost for analyzing and sequencing one sample can be approximated as around $100 on the basis of a cost comparison chart in [Table 2].

## **6.2 Time:**

As genome sequencing exceeds improvements in computer speed, there is a grave need to accelerate tasks sequencing, alignment and SNP calling. A conventional human re-sequencing project generates billions of short reads (100 GB of compressed data) needed to be uploaded to the Hadoop cluster for further processing, requiring more than 100 days of cluster time. The time needed to transfer this large dataset over Internet is significant and is very well addressed by allocating multiple EC2 instances of many nodes and having each node transfer subset of data from the source to S3 in parallel. Here are characteristic features of four pillars of the solution enabling better performance and addressing the time and speed related lacunae in an efficient manner by taking 3-5 hours for genome sequencing of one subject. [Table 2]

* Amazon Cloud Services – Efficient cloud storage
* Hadoop Framework – Distributed & parallel processing using (40 node 300-400 core cluster)
* Bowtie Algorithm – Ultrafast alignment algorithm using quality-directed best first search.
* SOAP Snp – Faster SNP calling and detection

## **6.3 Accuracy:**

Accuracy is one of the most important factors in genome sequencing. However traditional methods are devoid of accuracy and the ability to describe the context in which sequence variations co-occur in a cost effective manner. This challenge is contended by a technique offered by the solution called as SOAP Snp. It is a re-sequencing utility that can assemble consensus sequence for a newly sequenced subject’s genome on the basis of alignment present in the raw sequencing reads in the reference genome. Thereby identifying SNP’s on the consensus sequence through genome reference comparison. It is single cohesive integrated technique focusing on array of factors like data quality, alignment, tracking periodic errors and it supports SOAP aligner alignment format. Its accuracy is close to 99% as stated by many genotyping sites.

## **6.4 Storage:**

Sequencing and analyzing a genome of one subject involves petabytes of data considering the intricate DNA structure and the protein sequences. It is next to impossible to handle this data using traditional storage techniques taking the amount of cost involved in consideration. This issue is very well catered by the solution using Amazon simple storage service (Amazon S3). It is a highly scalable object storage, secure, durable, easy to use, customized, simple user interface to store and retrieve large amount of data from anywhere and everywhere. There are no cost overheads involved, which make it even more accessible to the end users. In order to address genome sequencing data it has multipart upload functionality where large data objects can be stored in smaller data chunks in parallel.

## **6.5 Scalability:**

Until now sequencing one genome sample was a mammoth task in itself. So making this scalable to address genomes of many subjects requires high amount of parallelization, constant monitoring and logging of various instances and consolidation of these multiple instances in a cohesive manner so that the integrated result can be used for further analysis. Here is how scalability is achieved using the solution to create a massive effect in genome sequencing where approximately 1000 samples can be analyzed in 2-4 weeks

* Launching multiple instances of EC2 in Amazon in parallel to trim raw input sequence data files for better load balancing
* Managing large scale whole genome sequencing analysis by using a logging and monitoring module that keeps a constant track of application’s execution, progress, status and errors, common hardware and network failures. And records runtime metrics like processing times, transfer times and file sizes.
* Integration module that aggregates all SNP’s from multiple samples and combines into a standard format so that it can be utilized further for detailed analysis of genome for a specific disease pertaining to individual subject.

|  |  |  |  |
| --- | --- | --- | --- |
| EC2 Nodes | 1 master, 10 Worker nodes | 1 master, 20 worker nodes | 1 master, 40 worker nodes |
| Worker CPU cores | 100 | 100-200 | 300-400 |
| Approximate Time | 6-8hrs | 4-6hrs | 3-5hrs |
| Approximate Cost | $40-60 | $60-80 | $80-200 |
| Approximate cluster setup time | 20-30min | 20-30min | 30min-1hour |

**Table 2. Cost comparison Chart for sequencing**

# **7. Conclusion**

This solution offers motivation for implementation of large-scale and cost effective whole genome sequencing analysis in the cloud by leveraging Amazon cloud services, Hadoop, efficient Bowtie and SOAP Snp algorithms for alignment and SNP detection. As of now the solution to genome sequencing has been restricted to research laboratories and far from the clinical reach. Thus we believe this solution can serve as a basis for taking computational biology to a new commercialized level by utilizing it to conduct a wave of citywide genome analysis to detect and cure diseases by inventing personalized medicines.

# **8. References**

1. Koboldt D, (2009,November 23), Crossbow: NGS Informatics in the Cloud, Retrieved from *Mass Genomics* blog post, <http://massgenomics.org/2009/11/crossbow-ngs-informatics-in-the-cloud.html>
2. Zhao S, Prenger K, Smith L, Messina T, Fan H, Jaeger E and Stephens S,(2013, June 27), Rainbow: a tool for large-scale whole genome sequencing data analysis using cloud computing, *BMC Genomics*, <http://www.biomedcentral.com/1471-2164/14/425>, doi:10.1186/1471-2164-14-425
3. Nekrutenko N, Taylo J, 2012, September), Next-generation sequencing data interpretation: enhancing reproducibility and accessibility, *Nature Reviews Genetics,* Vol 13, pp 667-672, <http://www.nature.com/nrg/journal/v13/n9/full/nrg3305.html>, doi: 10.1038/nrg3305
4. Amazon,(2012, December), Rainbow Semi – Automated Workflow, *Amazon AWS*, <http://s3.amazonaws.com/jnj_rainbow/code/RainbowManual_Dec12.pdf>
5. Langmead B, Schatz M, Lin M, Pop M, Salzberg S, (2009, November 20), Searching for SNPs with cloud computing, *Genome Biology,* 10(R134), <http://genomebiology.com/2009/10/11/r134> doi:10.1186/gb-2009-10-11-r134
6. Schatz M, Langemead B, Lin J, Pop M, Salzberg S, (2009, November 27) Whole genome Sequencing analysis in Clouds, *Schatzlab Edu Publications*, <http://schatzlab.cshl.edu/publications/posters/SC09-Crossbow.pdf>
7. [Bisciglia](http://blog.cloudera.com/blog/author/christophe/" \o "Posts by Christophe Bisciglia) C, (2009, October 15), Analysing Human Genome with Apache Hadoop, Retrieved from *Cloudera Blog Post,* <http://blog.cloudera.com/blog/2009/10/analyzing-human-genomes-with-hadoop/>
8. Kumari P, Kumar S, (2014, May), Analyse Human Genome Using Big Data, *International Journal Of Science and Research*, 3(5), <http://www.ijsr.net/archive/v3i5/MDIwMTMyMTgy.pdf>
9. Snyder M, Du J, Gerstein M, (2010), Personal genome sequencing: Current approaches and challenges, *Genes and Development Open Access*, Cold Spring Harbor Laboratory Press, <http://www.genesdev.org/cgi/doi/10.1101/gad.1864110>.
10. Peters B, Kermani B,Sparks A, Alferov O, Hong P, Jiang Y,Dahl Fredrik, Tang T, (2012, July 11), Accurate whole- genome sequencing and halotyping from 10 to 20 human cells. *Nature,* 190-195, <http://www.nature.com/nature/journal/v487/n7406/full/nature11236.html>, doi:10.1038/nature11236
11. Koboldt D, (2014,July 10), New Challenges of Next-Gen Sequencing, Retrieved from *Mass Genomics* blog post, <http://massgenomics.org/2014/07/new-ngs-challenges.html>
12. Koboldt D, (2014,November 7), Brace yourself for Large- Scale Whole Genome Sequencing, Retrieved from *Mass Genomics* blog post, <http://massgenomics.org/2014/11/brace-yourself-for-large-scale-whole-genome-sequencing.html>
13. Wetterstrand KA. (2014, October 31), DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: [www.genome.gov/sequencingcosts](http://www.genome.gov/sequencingcosts)
14. Green E, (2001, August), Strategies for the systematic Sequencing of complex Genomes, Retrieved from *Nature Reviews Genetics*, Macmillan Magazines, Vol 2, 573-583, <http://www.cs.tufts.edu/comp/167/green-01-natrevgen.pdf>
15. Genome News Network, (2003, January 15), Genome Sequencing, <http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp2_1.shtml>
16. S Michael, (2009, June 18), High Throughput Sequencing with MapReduce, Retrieved from *JCVI Informatics Seminar*