An Estimate of the Similarities in Genome Sequencing Concepts

N Banupriya[1], T Sethukarasi[2]

[1]Assistant professor, CSE,
RMK Engineering College,Kavaraipettai,TN
nbp.cse@rmkec.ac.in
[2] Professor & Head,CSE,
RMK Engineering College, Kavaraipettai,TN
hod.cse@rmkec.ac.in

Abstract. Genome, material of genetic, within all living species consisting of DNA (or) RNA in mRNA. Genome analysis is the process of identifying, measuring or comparing the features of genome lists sequencing DNA, Structural variation, expression of gene, regulatory element annotation and genomic scale of functional element annotation. Genome sequencing is the process of storing the data plays the major role in different field of engineering. It also matters in genome sequence analysis for storing the raw samples, sequenced genome and repetitive data. This paper gives the overview of genome sequencing, its types and storing concepts.

1. Introduction

Gene, fundamental unit of heredity. It is a small piece in genome. Genes are found in the chromosomes and are constructed of DNA. Several genes prefer several Characters. Number of gene in genome is different form one species to other. Gene has coding regions (genes) and noncoding regions (Mitochondrial DNA and Chloroplast DNA). Human genome[1] has 3.2*109 base pairs which is distributed among Twenty two paired chromosomes. Genomics is the detailed investigation of genes, including the interaction between their genes and with environment of an individual.Figure.1 shows a summary guide to genomics [2].

Genome analysis refers Sequencing of DNA, were assembled to create a Chromosome representation originally. Then annotation and analysis of that representation is completed. The process involved in genome analysis are Sequencing(converting amino acids from human into data i.e. sequence),Mapping(converting sequence data into scientific data like 1's and 0's) ,variant calling(comparing the existing with new samples) ,scientific discovery(using scientific data new medicines are discovered).

Genome Sequence is a nucleotide list (A,C,G,T) composed of all the chromosomes of an single or an group of living organisms. It is the process of analyzing DNA from human blood (i.e) extracting DNA and Sequencing. Sequencing of DNA ascertains the Sequence of protein, Sequence of protein ascertains Structure of Protein, and structure of Protein ascertains the function of Protein.

The body structures like inner organs and muscular tissue are made up of Proteins. It controls chemical reactions and carry signals between cells. Protein acts as the muscle example "Heart Muscle". Gene mutation affects the protein region which disrupts the entire body's usual series and be a route to a disorder of structure like cancer.

In human each cell has Twenty three pairs of chromosomes. Each chromosome has double helix which looks in ladder shape. Ladder is built of distinct compound named bases. All together DNA has 6 billion bases (i.e.) 3 billion base pairs and 4 chemical bases in DNA are A,T,G,C. DNA carries the information how the individual resembles in real world. Human have around ~18,000 to ~24,000 genes. DNA Sequencing is the series of sequence having nucleotide bases (A, T, C, and G) in a portion of DNA. Sequencing the DNA short piece is Straight forward sequencing while the entire genome (all of a DNA organisms) is tough task. Table 1.shows the Genome Sequencing Types [3].

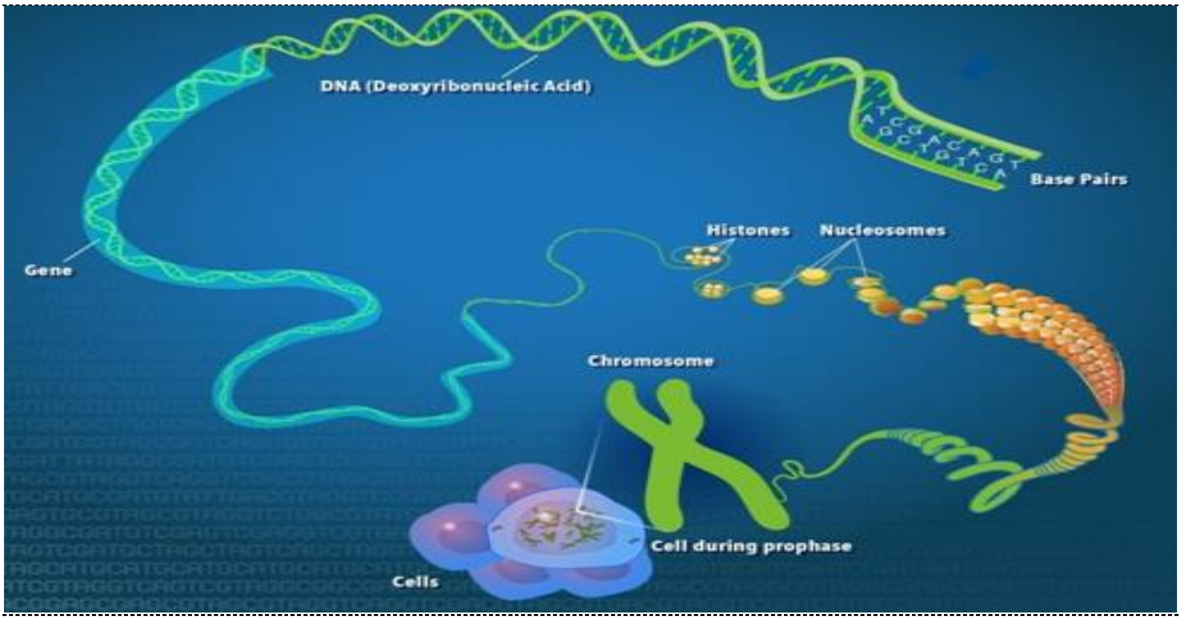

Figure 1. A Summary guide to Genomics [2]

There is a concept of New Generation Sequencing (NGS), a tool used for data management system, in production concepts and in analysis of downstream. It is a subset of genes steps in which cancer mutations focus on a limited genes number, whereas the WGS is focused on protein coding regions (~2% of the genome) and does not require subset of genes.

Table 1. Types of Genome Sequencing

Types of Genome Sequencing	Explanation
1.Whole Genome Sequencing [WGS]	Focus on the sequencing of the entire DNA in an organism's genome. There are about 6 Application of Whole Genomic Sequencing [4]. They are (i) Justbornand diseases of Pediatric, (ii) Drug trails and pharmacogenomics (iii) Regulatory variation and eQTLs (iv) Very Rare Tumor Types (v) Clan Genomics(vi) Large Cohorts with Extensive Phenotyping.
(a)De-novo	Sequencing start from the beginning. An organism genome is sequenced and assembly is done without referral genome.
(b) Resequencing	An organism genome is sequenced and assembly is done using referral genome.
2.Tagetted Genome sequencing	Is pointed to a specific region of interest within genome.

(a)Exome pull down	Sequencing only the gene portion in genome.

2. Literature Survey

[5] In 1995, first genome sequence is analysed for cellular life form named parasitic bacterium Haemophilus influenza. Later in 1999, 20 complete genomes of bacteria is analysed. Genome analysis is the trend in genetics. Genome analysis is computational and theoretical methods.[6] A comparative study in genome analysis of 2 flowers Arabidopsis thaliana and capsella rubella ,belong to diploid species of Brassicaceae family is made.Comparision of orthogonal genes in two flowers revealed similarity of exon-intron structures and identities of sequence are appreoximately 89% or more. Exons are the sequence that contain the code for protein(polypeptide). Introns are the sequence of non-coding in mRNA. It is essential for sequencing because plant genomes are much more dynamic when compared to animal genome.

[7] In figure2.Intron/Exon Splicing is shown.A small fraction of DNA is Transcipted into mRNA and not all mRNA is translated to protein. Intron regions are removed or spliced out of pre-mRNA(precursor mRNA) which contains only exons. This is process carried over in computational gene identification. In this way Eukaryotic genome that encode for proteins picked out.

In [8], River buffalo (Bubalusbubalis) DNA is taken for sequencing. Auto assembler software is used to align before sequencing.[9] Whole genome analysis is made in Campylobacter jejuni. It is causes of food poisoning in Europe and in united states.18 strains from diverse sources are taken to analyze by comparing genomichybridization DNA to a DNA microarray. Result in this paper gives the way to geneticfuture typing schemes and microarray related studies in epidemiological field. [10]Some respiratory tract diseases are affecting human being individual. Many literary genre represented respiratory viruses including human metapneumovirus, SARS, (Corona virus) SARS-CoV, Human Corona Virus NL63, were discovered in past 10 years. In this paper the genome of corona is sequenced and thenanalysedby pneumoniapatients. The basic characterization and complete sequence of genome are analysed.

[11]Whole genome[WG] analysis of marine Bacteroidetes is worked out in this paper. Short gun sequencing technique is also applied to the analysis forGramellaforsetii KT0803 genome. This work gives the comparative study of Bacteriodetes survival by attaching to the organism and to see the predicted hydrolytic activities differentiate once planktonic (the collection of different organism in marine which are unable to swim against current) genomes of the Bacteroidetes are available.
Genome Analysis of N.Eutropha C91 using whole genome analysis with short gun libraries is done. (Nitrosomonaseutropha) N.Eutropha C91 is found in municipal and industrial wastes which elevates ammonia concentration with high tolerance. This paper mainly explains the adaptation of N.Eutropha in N.Environment.

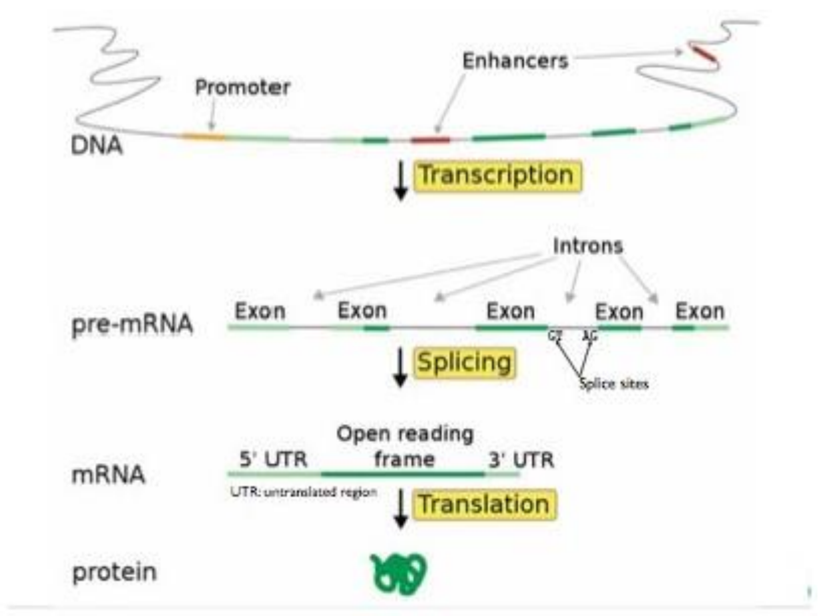

Figure 2. Intron/Exon [7]

[12]Sager Sequencing is the traditional methods with long reads from short reads or pair of short reads is sequenced. In next few years the data is sequenced in large amount and requires detailed change of stored data and how query of users needed raw information. Some of tools used for short read is discussed in this paper like Illumina GAII or AB SoLiD, BLAST. ZOOM is a new sequence comparison tools for second generation short read sequence. GenBank is a large repository used for genome sequencing. NCBI is the used for the GenBank DNA Sequence. The “1000 genome project” is having the samples from 2008 to 2015 with large human variation and genotype data. It has the generic variants with frequencies of atleast 0-1% in the populations studied and reduced in cost of sequencing. For SNP, the database used is dbSNP (repository) and support the human and bovine HapMap projects (genomic structure of cattle).Some of the browsers revised here are Ensembl, Generic Genome Browser, and UCSC Genome Browser. In [13], the investigation of NGS techniques is well discussed and its strategies enable its user by characterizing thefull variation spectrum of human sequence of DNA.[14]Galaxy Tool introduced in 2010, is high end and user interface ,hides the details of computation and memory storage management. They are also distributed as public which provides genomic analysis tool, genomic compression and genomic data functions or package which can be installed in individual research laboratories.

[15]GATK is an essential supporting structure designed for NGS uses the philosophy of Map Reduce in functional programming. This can be the part of beneficiary in improving the management data engine. GATK should support the additional data access object pattern to enable reference of local guided assemblyimplementation, CNV, and structure of generalvariation algorithm in future. Denovo genome assembly remains challenge due to short read length, data missing, errors in sequence, characterized by repetition regions and this is known as Local reference guided assembly.CNV known as slight difference in condition and it occurs due to the duplicate of genes varies from one individual organism to another. Inversion is the chromosomal rearrangement in which segment of chromosomes is reversed end to end. Cytogenetic techniques are used to detect inversion. Inversion is also inferred by genetic analysis. General structure variation algorithms also referred as structural variation or CNV. This largely made impact in functions of encoded genesin genome andmade responsible for disease diverse in human. This paper evaluates ~70 SV algorithms of detection uses various multiple simulation and data sets of WGS.

[16] gives information about how to use GATK and BWA to map exactly sequencing of genomefrom one to another data reference and induce text file format in which gene sequence variation is stored, that can be used in downstream analyses. Data of NGSprocessingpreliminary steps were analysed using GATK and methodology involving in discovery of variant using GATK.

By using GFF(general feature format) all genetic data is stored and by using (variant call format)VCF the variation need to be stored along with referential genome.[17]Human genome have more than 3 billion nucleotides and about 23,000 genes to 23,510 genes .Every Gene have protein-coding region(exons) and it has 1,80,000 exons collectively known as exome. Here we concentrate in WES (whole Exome Sequence) and in cardiovascular problems.DNA Sequence Variants (DSVs) in exome is identified using WES. Data of WES is used in Clinical analysis, needs deep understandingof medicinal genetics and clinical medicine.Table 2.Shows the copious of Human Genome DNA sequence variants.

Table 2.copious ofHuman Genome DNA sequence variants.[17]

Nucleotides	3.2 x 10^9(base pairs)
Protein-coding genes	~24,000
Number of exons	120,000 to 181,000
Size of exome	30 x 10^6(base pairs)
DSVs	4 x 10^6
Single nucleotide polymorphisms (SNPs)	3.5 x 10^6
De novo variants	25-31
Variants associated with inherited diseases	70-105

[18] Gene Expression is challenge in Computational biology. Genetic Neural Network used to predict genome-wide expression of gene. Natural Language Processing, recurrent neural network, Bi recurrent neural network is compared with GNN. It uses nodes of a cell capturing the province and dynamics non-linear, exist in gene networks. These two key note factors concentrated in this paper. [19]NGS used for whole genome at a low cost. Assemblies of Denovo genome keepremains challenge in short read length, repetitive regions, missing of data, sequencing error and polymorphisms till now. In this paper, reference guided assembly approach is used. Normal Denovo and reference guided Denovo assembly approach is roughly calculated for diverse in character of genomes of plants.

WGS is used for the diversity in genetic of two species of Bdellovibrio, isolated from soil. This species is one type of gram negative bacteria which is present in fresh water, river side etc and nontoxic to human. Mainly in this work, the predatory features of this species and genetic characteristics were analyzed which can contribute as an application to biocontrol agent. ANI/AAI is Matrix basedgenomeand utilized asmatrix distance calculator , used to find similarity features and their ecological living. The Annotation server namely Rapid Annotation using Subsystem Technology is used for predicting some gene to enhance predation in Bdellovibrio spp.

[20] Hereditary disease ALS, a disease causing variants that have been identified using Dutch cohort project MinE dataset is used to which contains healthy individuals.CNN is used to predictgenotype data's ALS prevalence. Deep learning in genotype-phenotype association analysis is the initiative step made in Deep neural network (DNN).

3. Generation of DNA Sequencing

In First generation, DNA 3D structure in 1953 is analyzed by Watson [21]. Optics data produced by Franklin contributed for both DNA copies and encoding proteins in nucleic acids. Dideoxychain termination method or Sanger Sequencing for long read were also introduced this period.

In Second generation, Illumina sequencing platforms is used for DNA sequences. Large Scale Dideoxy sequencing was under process to prove in the market. On this busy time, Sequence by synthesis technique was introduced which is the combination of Sangers Dideoxy and Pyrosequencing method. Each nucleotide is washed through the system in turn over the template DNA of fixed to solid phase.NGS is evolved in this generation. First High throughput machine called GS20 ,later extended and named as 454GSFLX was used for genome sequencing .Solexa sequencing is one of the massively Parallel sequencing techniques, later acquired by Illumina.

In Third generation some of the sequencing methods, SMRT, and Simple scalar are in practice. SMRT (Single Molecule real-time). PacBio Machines is used and Nanopore Sequencing established before second generation.ONT (oxford Nanopore Technologies), the first company offering nanopore sequences. GridION Mk1 and Flongle Flow Cells are Nanopore platforms. Table 3.gives the information about the estimate of tools, Storage/Repository used for genome sequencing.

Table 3. Estimate of tools, Storage/Repository used for genome sequencing.

	Tools	Description	Year
1	AutoAssembler [8]	Significant sequence identity (78.95%) between buffalo sequence (Bubalusbubalis) and Cattle. [8]	2001
2	Whole genome DNA[9] microarrays	Genome has ORF,refersfor examining the diversity of genetic betweendifferent 18 isolates of C. jejuni.	2003
3	Short Gun Sequencing [11]	Genome Analysis of Marine Bacteroidetes	2006
4	Whole Genome analysis(Short gun Libraries)	Genome analysis of NitrosomonaseutrophaC91	2007
5	Illumina GAII or AB Solid,ZOOM[12]	Tools used for Short Read	2009
6	Galaxy[14]	It is a framework acts as simple interfaces to powerful tools provided	2010

7	GATK[15]	It’s a platform designed for DNA sequences of next generation analysis	2010
8	QUAST[19]	quality assessment tool	-
9	ANI/AAIMatrix	Genome-based distance matrix calculator	2018
10	RAST	Rapid Annotation using Subsystem Technology server	2018
11	Basic local alignment search tool.[21]	It finds similarity regions between biological sequences.	1990
12	GenBank	NIH genetic sequence database	2012
13	The DNA Databank of Japan (DDBJ) ,	Rice Annotation Project Database	2006
14	European Molecular Biological Laboratory (EMBL)	Maintained in collaboration with partners DNA Data Bank of Japan and GenBank includes whole genome sequencing project data	2005
15	dbSNP [22]	Database used for SNP	2000
16	UHTS[23]	ultra-high-throughput sequencing	2013
17	BAC(bacterial artificial chromosome)	Tools for genome sequencing	2000
18	COGs	Phylogenetic Classification for theproteins encoded with complete Bacteria genomes, archaea genomes, and eukaryotes genomes.	1999

4. Conclusion

This paper gives a brief idea regarding what is genome sequencing, concepts and its types used in various field like cattle, marine, flowers, environments etc, tools used for sequencing and storage concepts are also discussed. The genome sequencing is performed using CNN, DNN, ANN and GNN in 2018. Furthermore research can be deep into ANN for genome sequencing. Some of the tools Illumina, BLAST, GATK, Galaxy used for sequencing genomes and NGS are also used for Whole Genome sequencing. GenBank, dbSNP repositories etc can also be used for annotation and comparing with existing samples.

5. References

1. Arthur M.Lesk,”Introduction to Genomics”,Secondedition,Oxford University press 2012, ISBN 978–0–19–956435–4.
2. https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics

3. www.yourgenome.org/facts/types-of-genome-sequencing
4. https://www.genengnews.com/insights/6-applications-for-whole-genome-sequencing/
5. Why genome analysis? TIG April 1999, volume 15, No. 4.
6. AdileAcarKon, Mathias Robberg, Marcus Koch, Renate Schmidt," Comparative Genome analysis reveals extensive conservation of genome organization for Arabidopsis thaliana and Capsella rubella", accepted 25 April 2000.
7. Prof.ManolisKellis, "ComputationalBiology: Genomes, Networks,Evolution",January6, 2016.
8. N. Navani, P. K. Jain, S. Gupta, B. S. Sisodia and S. Kumar, "A set of cattle microsatellite DNA markers for genome analysis of riverine buffalo (Bubalusbubalis)", 2002 *International Society for Animal Genetics*, Animal Genetics, 33, 149-154.
9. B.M. Pearson, C. Pin, J. Wright, K. I'Anson, T. Humphrey, J.M. Wells, "Comparative genome analysis of Campylobacter jejuni using whole genome DNA microarrays", 2003 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.doi:10.1016/S0014-5793 (03)01164-5.
10. doi:10.1128/JVI.79.2.884–895.2005.
11. Margarete Bauer, Michael Kube, Hanno Teeling,ElkeAllers, Chris A. Würdemann,ChristianQuast, HeinerKuhl,FlorianKnaust, Dagmar Woebken, Kerstin Bischof,MarcMussmann, Jomuna V. Choudhuri, Folker Meyer, Michael Richter,ThierryLombardot,Richard Reinhardt, Rudolf I. Amann and Frank Oliver Glöckner,"Whole genome analysis of the marine Bacteroidetes 'Gramellaforsetii' reveals adaptations to degradation of polymeric organic matter", *Journal compilation © 2006*, Environmental Microbiology, 8, 2201–2213.
12. Jacqueline Batley and David Edwards , "Genome sequence data: management, storage, and visualization",BioTechniques46:333-336 (April 2009 Special Issue) doi 10.2144/000113134.
13. doi:10.1093/bib/bbq016.
14. Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor,GuruprasadAnanda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor, "Galaxy: AWeb-Based Genome Analysis Tool for Experimentalists", Current Protocols in Molecular Biology 19.10.1-19.10.21, January 2010, DOI: 10.1002/0471142727.mb1910s89.
15. Aaron McKenna,Matthew Hanna, Eric Banks, AndreySivachenko, Kristian Cibulskis, Andrew Kernytsky, KiranGarimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo,"The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data", 2010 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/10; www.genome.org, http://www.genome.org/cgi/doi/10.1101/gr.107524.110.
16. DOI: 10.1002/0471250953.bi1110s43.
17. AJ Marian, M.D,"SEQUENCING YOUR GENOME: WHAT DOES IT MEAN?", houstonmethodist.org/debakey-journal, 2014.
18. AmeenEetemadi and IliasTagkopoulos,"Genetic Neural Networks: an artificial neural network architecture for capturing gene expression relationships", The Author(s) 2018. Published by Oxford University Press. All rights reserved, Bioinformatics, 35(13), 2019, 2226–2234 doi: 10.1093/bioinformatics/bty945.
19. Heidi E. L. Lischer and Kentaro K. Shimizu,"Reference-guided de novo assembly approach improves genome reconstruction for related species", Lischer and Shimizu BMC Bioinformatics (2017) 18:474, DOI 10.1186/s12859-017-1911-6.
20. Bojian Yin, MarleenBalvert, Rick A. A. van der Spek, Bas E. Dutilh, Sander Bohte, Jan Veldink and Alexander Schonhuth,"Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype", The Author(s) 2019. Published by Oxford University Press., Bioinformatics, 35, 2019, i538–i547 doi: 10.1093/bioinformatics/btz369.

21. Stephen F. AltschuP, Warren Gish ~, Webb Miller 2 Eugene W. Myers 3 and David J. Lipman ~,”Basic Local Alignment Search Tool”, Received 26 February 1990; accepted 15 May 1990J. Mol. Biol. (1990) 215, 403-410.
22. Smigielski, E.M., K. Sirotkin, M. Ward, and S.T. Sherry. 2000.” dbSNP: a database of single nucleotide polymorphisms.”
23. C. Bertelli and G. Greub” Rapid bacterial genome sequencing: methods and applications in clinical” Clinical Microbiology and Infection, Volume 19 Number 9, September 2013. 10.1111/1469-0691.12217.

Acknowledgments

We thank Our Management for fullest support.Thanks to my department HOD to guide me and suggest some Bio informatics materials to complete this paper. I would like to continue this paper for Funding projects.