Review Article

IJCRR
Section: Healthcare
Sci. Journal
Impact Factor
4.016

BIOINFORMATICS ADVANCES IN GENOMICS – A REVIEW

Raphael John Ogbe[1], Dickson O. Ochalefu[2], Olumide B. Olaniru[3]

[1]Department of Physiology, Pharmacology and Biochemistry, College of Veterinary Medicine, University of Agriculture, P.M.B. 2373, Makurdi, Benue state, Nigeria; [2]Department of Biochemistry, College of Health Sciences, Benue State University, Makurdi, Benue State, Nigeria; [3]Department of Chemical Pathology, Jos University Teaching Hospital, Jos, Plateau State, Nigeria.

ABSTRACT

Summary

Genomics is a discipline in genetics that applies recombinant DNA technology, DNA sequencing methods and bioinformatics to sequence, assemble and analyze the function and structure of genome, the complete set of DNA within a single cell of an organism. Bioinformatics is an inter-disciplinary scientific field that develops methods for storing, retrieving, organizing and analyzing biological data. The advances in bioinformatics have in turn made considerable impact on the development and improvements of genomics technologies such as shot-gun sequencing and high-throughput sequencing methods. The various genomics technologies are used for DNA and genome sequencing, assembly and annotations, which have several applications in medicine, agriculture, pharmaceuticals, biotechnology, research etc. These genomics technologies aided by bioinformatics have contributed to the successful completion of whole organism genome analysis, from prokaryotes to eukaryotes. In fact, the assembly of the human genome is one of the greatest achievements of bioinformatics.

Key Words: Annotation, Assembly, Biotechnology, DNA sequence, Genome analysis

INTRODUCTION

Genomic technologies are generating an extraordinary volume of information never before achieved in the history of biology. With recent advances in technology and the development of ultra high-throughput research, the field of biotechnology began to suffer from data overload. This lead to the emergence and evolution of a broadening field of scientific discipline called bioinformatics, which is at the intersection between biology and computation [38]. Bioinformatics is therefore often considered to be different thing by different people. In its most basic form, bioinformatics might be described as 'the structuring of biological information to enable logical interrogation [35].

Bioinformatics addresses the specific needs in data acquisition, storage, analysis and integration, which research in genomics generates. This relatively new scientific discipline facilitates both the analysis of genomic and post-genomic data, and the integration of information from the various related fields of transcriptomics, proteomics, metabolomics and phenomics. This integration enables the identification of genes and gene products, and can elucidate the functional relationships between genotype and observed phenotype, thereby allowing a system-wide analysis from genome to phenome [35]. Among the current research lines are the following: 1). Gene prediction and modeling of splicing, related to the research on regulation of alternative splicing, and protein synthesis 2). Identification and characterization of genomic regions involved in Gene Regulation, related to the research on Chromatin, Gene Expression, and on the RNA-Proteins Interactions and 3). Molecular Evolution, which includes evolution of the exonic structure of genes and the evolution of splicing. The bioinformatics program also includes a research in microarrays, which is complemented with a new group specifically devoted to Microarray Informatics. Thus, the purpose of this study is to discuss the bioinformatics advances in genomics and highlights some of the applications of bioinformatics advances to biological fields.

GENOMICS

Genomics is a discipline in genetics that applies recombinant DNA technology, DNA sequencing methods, and bioinfor-

Corresponding Author:

Raphael John Ogbe, Department of Physiology, Pharmacology and Biochemistry, College of Veterinary Medicine, University of Agriculture, P.M.B. 2373, Makurdi, Benue state, Nigeria
Tel.; +234 806 530 0976; E-mail: ralphjohn2012@gmail.com

Received: 08.06.2015 **Revised:** 20.07.2015 **Accepted:** 02.09.2015

matics to sequence, assemble and analyze the function and structure of genomes, the complete set of DNA within a single cell of an organism [1, 2]. The field includes efforts to determine the entire DNA sequence of organisms and fine-scale genetic mapping. It also includes studies of intragenomic phenomena such as heterosis, epistasis, pleiotropy and other interactions between loci and alleles within the genome [3].

Brief history of genomics

The term "genomics" was coined by Dr. *Tom Roderick,* a geneticist at the Jackson Laboratory (Bar Harbor, Maine), at a meeting held in Maryland on the mapping of the human genome in 1986 [4]. Though the word "genome" (derived from the German word Genom, attributed to Hans Winkler) was in use in English as early as 1926.

Early genes sequencing efforts

Nucleic acid sequencing became a major target of early molecular biologists, following *Rosalind Franklin's* confirmation of the helical structure of DNA around 1941, *James D. Watson* and *Francis Crick'*s publication of DNA structure in 1953 and Frederick Sanger's publication of the amino acid sequence of insulin in 1955 [5]. In 1964, *Robert W. Holley* and colleagues published the first nucleic acid sequence ever determined, the ribonucleotide sequence of alanine tRNA [6,7]. Extending this work further, *Marshall Nirenberg* and *Philip Leder* revealed the triplet nature of the genetic code and were able to determine the sequences of 54 out of 64 codons in their experiments [8].

In 1972, *Walter Fiers* and his team at the Laboratory of Molecular Biology, Ghent, Belgium, were the first to determine the sequence of a gene: the gene for Bacteriophage *MS2* coat protein [9].

BIOINFORMATICS

As a result of recent advances in technology and the development of ultra high-throughput sequencing research techniques, the field of biotechnology started to experience data overload. This lead to the development of an ever-broadening field of science known as bioinformatics, in which biology and information technology converge. This is an interdisciplinary scientific field that develops methods for storing, retrieving, organizing, and analyzing biological data. A major activity in bioinformatics is to develop software tools to generate useful biological information or knowledge. Therefore, simply defined, bioinformatics uses computers to better understand biology, by working with basic biological data e.g. DNA bases. That means it works on a small scale paying attention to details unlike computational biology which builds large scale general theoretical models of biological systems.

Brief history of bioinformatics

Paulien Hogeweg was the one that coined the term "Bioinformatics" in 1970, to refer to the study of information processes in biotic systems. Bioinformatics is an interdisciplinary field in a broader field of biotechnology.

Computers became essential in molecular biology when protein sequences became available. After the sequence of insulin was determined in the early 1950s by Frederick Sanger, comparing multiple sequences manually turned out to be impracticable, hence the need for computers. Margret Oakley Dayhoff, recognized as the "mother and father of bioinformatics", was a pioneer in the field and compiled one of the first protein sequence databases, initially published as a book, and also pioneered methods of sequence alignment and molecular evolution [10].

DNA sequencing technology developed

Fier's group then expanded on their *MS2* coat protein work, determining the complete nucleotide sequence of bacteriophage *MS2*-RNA (whose genome encodes for just four genes with 3569 base pairs) and Simian virus 40, in 1976 and 1978 respectively [11, 12]. *Frederick Sanger* and *Walter Gilbert* shared the 1980 Nobel prize in chemistry, for independently developing methods for the sequencing of DNA. Sanger and his colleagues played a key role in the development of DNA sequencing techniques, which enabled the establishment of comprehensive genome sequencing projects [3]. In 1975, he and *Alan Coulson* published a sequencing procedure using DNA polymerase with radio-labeled nucleotides, which he called the "Plus and Minus technique" [13, 14]. In 1977, his group was able to sequence most of the 5,386 nucleotides of the bacteriophage φX174, completing the first fully-sequenced DNA-based genome [15].

BIOINFORMATICS CONTRIBUTIONS TO ADVANCES IN GENOMICS

The primary goal of bioinformatics is to increase the understanding of biological processes. Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them, creating and viewing 3-Dimentional models of protein structures. Over the past few decades, rapid developments in genomics, other molecular research technologies and developments in information technology have combined to produce a tremendous volume of information related to molecular biology, leading to the advancement of bioinformatics. These bioinformatics advances have in turn lead to greater developments of genomics.

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theories to solve formal and practical problems arising from the management and analysis of biological data. In the field of genetics and genomics, it aids in sequencing and annotating genomes and their observed mutations. It plays a role in the textual mining of biological literature and the development of biological and gene ontologies to organize and query biological data. As whole genome sequences became available, again with the pioneering work of *Frederick Sanger* [15], bioinformatics was re-discovered to be useful in the creation of databases such as GenBank in 1982. It is now known to play a role in the analysis of gene and protein expression and regulation. Bioinformatics tools aid in the comparison of genetic and genomic data, and more generally in the understanding of evolutionary aspects of molecular biology. At a more integrative level, it helps to analyze and catalogue the biological pathways and networks that are important part of systems biology. In structural biology, it aids in the simulation and modeling of DNA, RNA and protein structures, as well as molecular interactions.

Sequence analysis

Since the bacteriophage, φX174, was sequenced in 1977 [15], the DNA sequences of thousands of organisms have been decoded and stored in databases. This sequence information is analyzed to determine genes that encode polypeptides (proteins), RNA genes, regulatory sequences, structural motifs and repetitive sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species i.e. the use of molecular systematics to construct phylogenetic trees.

Phylogenetic Tree

A phylogenetic tree, also known as a phylogeny, is a diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor. Thus, a phylogenic tree or evolutionary tree is a branching diagram or “Tree” showing the inferred evolutionary relationships among various biological species or other entities, based upon similarities and differences in their physical or genetic characteristics. Phylogenies are useful for organizing knowledge of biological diversity, for structuring classifications, and for providing insight into events that occurred during evolution. Furthermore, because these trees show descent from a common ancestor, and because the strongest evidence for evolution comes in the form of common ancestry, one must understand phylogenies in order to fully appreciate the overwhelming evidence supporting the theory of evolution [37]. The importance of phylogenetic trees is that they provide efficient structure for organizing knowledge of biodiversity and allow one to develop an accurate, non-progressive concept of the totality of evolutionary history of different organisms, species or genes. The trees are useful in the field of bioinformatics, systematics and comparative phylogenetics.

Blast Analysis

BLAST – An acronym for Basic Local Alignment Search Tool, is an algorithm for comparing primary biological sequence information, such as the amino acid sequences of different proteins or the nucleotides of DNA sequences. It has long ago became impracticable to analyze DNA sequences manually, due to the growing amount of biological data. Therefore, computer programs such as BLAST are used daily to search for sequences from more than 260,000 organisms, containing over 190 billion nucleotides [16]. These programs can compensate for mutation (exchanged, deleted or inserted bases) in the DNA sequence, to identify sequences that are related but not identical. BLAST is actually a family of programs that can be used for several purposes. These include identifying species, locating domains, establishing phylogeny, DNA mapping and comparison. For example, with the use of BLAST, it is possible for scientists to correctly identify a species or find homologous species, which can be useful when a scientist is working with a DNA sequence from an unknown species. The BLAST program was designed by Altschul and his colleagues at the National Institute of Health, USA [36]. BLAST is one of the most widely used bioinformatics programs because it addresses a fundamental problem and the heuristic algorithm it uses is much faster than calculating an optimal alignment.

Genome sequencing approaches

As sequencing technology continues to improve, however, a new generation of effective, fast turnaround bench-top sequencers have become available to the academic research laboratories [25, 26]. On the whole, genome sequencing approaches fall into two broad categories, the Shotgun and High-throughput (also known as Next-generation) sequencing [3].

Shotgun sequencing technique

Shotgun sequencing is a sequencing method designed for analysis of DNA sequences longer than 1000 base pairs, up to and including entire chromosomes [34]. It is named by analogy with the rapidly expanding, quasi-random firing pattern of a shotgun. Shotgun sequencing is a random sampling process, requiring over-sampling to ensure a given nucleotide is represented in the reconstructed sequence by computer softwares.

High-throughput sequencing method

The demand for low-cost sequencing has driven the development of high-throughput sequencing (or Next generation sequencing) technologies that perform sequences in parallel, producing thousands or millions of DNA sequences at once [27]. High-throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond the standard dye-terminator methods. In ultra-high-throughput sequencing, as many as 500,000 sequencing-by-synthesis operations may be run in parallel [28].

Sequencing Complete Genome

A technique called Shotgun Sequencing (which was used by the Institute for Genomic Research for instance, to sequence the first bacterial genome i.e. *Haemophilus influenza* genome) in 1995 [17], does not produce entire chromosomes, instead it generates the sequences of many thousands of small DNA fragments. The ends of these fragments overlap and when align properly by a genome assembly program, can be used to reconstruct the complete genome. The following year, i.e. in 1996, a consortium of researchers from laboratories across North America, Europe and Japan, announced the completion of the first complete genome sequence of a eukaryote, *Saccharomyces cerevisiae* (12.1 Mb), and since then genomes have continued to be sequenced at an exponentially growing pace [18]. As at October 2011, the complete genome sequences are available for: 2,719 viruses, 1,115 archaea and bacteria, and 36 eukaryotes, out of which about half are fungi [19, 20].

Genome analysis

After an organism has been selected, genome projects involve three components: the sequencing of DNA, the assembly of that sequence to create a representation of the original chromosome, followed by the annotation and analysis of that representation [3]. Overview of a genome analysis project is that: First, the genome must be selected, which involves several factors such as cost and relevance. Second, the sequence is generated and assembled at a given sequencing centre. Third, the genome sequence is annotated at several levels: DNA, protein, gene pathways, or comparatively.

Shotgun sequencing yields sequence data quickly, but the task of assembling the fragments can be quite complicated for larger genomes. For a genome as large as the human genome, it may take many days of CPU time on large-memory, multiprocessor computers to assemble the fragments, and the resulting assembly will usually contain numerous gaps that have to be filled later. The core of comparative genome analysis is the establishment of the correspondence between genes (orthology analysis) or other genomic features in different organisms. It is these inter-genomic maps that make it possible to trace the evolutionary processes responsible for the divergence of two genomes. A multitude of evolutionary events acting at various organizational levels shape genome evolution. At the lowest level, point mutations affect individual nucleotides. At a higher level, large chromosomal segments undergo mutations; duplication, lateral transfer, inversion, transposition, deletion and insertion.

Human Genome Analysis

The rationale for the human genome project is to acquire fundamental information concerning our genetic make-up which will advance our basic scientific understanding of human genetics and the role of various genes in health and in disease. A rough draft of human genome was provided by the Human Genome Project in 2001, while by 2003, the project was completed which sequenced the entire genome of one specific person. By 2007, this sequence project was declared “finished”, with less than one error in 20,000 bases and all chromosomes assembled [22]. Since the years after completion of the human genome project, the genomes of many other individuals have been sequenced, partly under the auspices of the 1000 Genomes Project, which announced the sequencing of 1,092 genomes in October 2012 [23]. The completion of this project was made possible by the development of dramatically more efficient sequencing technologies and required the commitment of significant bioinformatics resources from a large international collaboration [24]. The continued analysis of human genomic data has medical benefits but with profound political and social repercussions for human societies. It is expected that future approach to patient’s treatment will require the human genome.

Sequence assembly

Sequence assembly refers to aligning and merging fragments of a much longer DNA sequence in order to reconstruct the original sequence. This is needed because the common currently used DNA sequencing technology can not read whole genomes as a continuous sequence but only reads small pieces of between 20 and 1000 bases, depending on the technology used. Typically, the short fragments called *reads*, are formed from shotgun sequencing genomic DNA or gene transcripts. Multiple, fragmented sequence *reads* must be assembled together on the basis of their overlapping areas [3]. Assembly of the human genome is one of the greatest achievements of bioinformatics.

Sequence assembly approaches

Assembly can be broadly categorized into two approaches: de novo assembly, for genomes which are not similar to anyone sequenced in the past, and comparative assembly, which uses the existing sequence of a closely related organism as a reference during assembly [29].

Finishing - Finished genomes are defined as having a single continuous sequence, with no ambiguities representing each replicon [30].

Genome Annotation

Another aspect of bioinformatics in sequence analysis is annotation. This involves computational gene finding, to search for protein-coding genes, RNA genes, and other functional sequences within a genome. Not all of the nucleotides within a genome are part of genes. Within the genomes of higher organisms, for example, large parts of the DNA do not serve any obvious purpose. This so-called "junk DNA" may however, contain unrecognized functional elements.

Bioinformatics helps to bridge the gap between genome and proteome projects-for example, in the use of DNA sequences for protein identification. In the context of genomics, annotation is the process of marking the genes and other biological features in a DNA sequence. The first genome annotation software system was designed in 1995 by *Owen White*, who was part of the team that sequenced and analyzed the first genome of a free-living organism to be decoded. i.e. the bacterium, *Haemophilus influenza*. White built a software system to find the genes (fragments of genomic sequence that encode proteins), the transfer RNAs and to make initial assignments of function to those genes.

The DNA sequence assembly alone is of little value without additional analysis [3]. Genome annotation is the process of attaching biological information to sequences. Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA such as the GeneMark program, which was developed and used to find protein-coding genes in *Haemophilus influenza,* are constantly changing and improving.

Steps in Genome Annotation

Genome annotation consists of three main steps [21]:

1. Identifying portions of the genome that do not code for proteins
2. Identifying elements on the genome, a process called gene prediction, and
3. Attaching biological information to these elements.

OTHER APPLICATIONS OF BIOINFORMATICS

Computational evolutionary biology

Evolutionary biology is the study of the origin and descent of species, as well as their change over time. Bioinformatics has assisted evolutionary biologists by enabling researchers to:

- Trace the evolution of a large number of organisms by measuring changes in their DNA, rather than through physical taxonomy or physiological observations alone.
- More recently compare entire genomes, which permits the study of more complex evolutionary events, such as gene duplication, horizontal gene transfer, and the prediction of factors important in bacterial speciation.
- Build complex computational models of populations to predict the outcome of the system over time.
- Track and share information on an increasingly large number of species and organisms.

Genomic medicine

Next-generation genomic technologies allow clinicians and biomedical researchers to drastically increase the amount of genomic data collected on large study populations [31]. When combined with new bioinformatics approaches that integrate many kinds of data with genomic data in disease research, they allow researchers to better understand the genetic bases of diseases and drug responses, which can lead to personalized medical care.

Understanding the genetic factors of diseases

With the advent of next-generation sequencing technology, we are beginning to obtain enough sequence data to map the genes of complex diseases such as infertility, breast cancer, or Alzheimer's disease. Genome-wide association studies are essential to pinpoint the mutations for such complex diseases [33]. This is made possible with the advancement of bioinformatics and genomics.

Analysis of mutations in cancer

In cancer, the genomes of affected cells are arranged in complex or even unpredictable ways. Massive sequencing efforts are used to identify previously unknown point mutations in a variety of genes in cancers. Bioinformaticians continue to produce specialized automated systems to manage the volumes of sequence data produced. They create new algorithms and softwares to compare the sequencing results to the growing collection of human genome sequences and germline polymorphisms, which is one of the causes of cancer. New physical detection technologies are being employed, such as oligonucleotide micro-arrays, to identify chromosomal gains and losses (called comparative genomic hybridization), and single-nucleotide polymorphism arrays, to detect known point mutations.

Analysis of gene expression

The expression of genes can be determined by measuring mRNA levels with multiple techniques including Microarrays, Expressed cDNA sequence tag (EST) sequencing, Serial Analysis of Gene Expression (SAGE) tag sequencing, Massively Parallel Signature Sequencing (MPSS), RNA-seq, also known as "Whole Transcriptome Shotgun Sequencing" (WTSS) or various applications of multiplexed *in-situ* hybridization. All of these techniques are extremely noise-prone and subject to bias in the biological measurements. Thus, a major research area in computational biology involves developing statistical tools to separate signal from noise, in high-throughput gene expression studies. Such studies are often used to determine the genes implicated in a disorder; so that one might compare microarray data from cancerous epithelial cells to data from non-cancerous cells, to determine the transcripts that are up-regulated and down-regulated in a particular population of cancer cells.

Synthetic biology and bioengineering

The growth of genomic knowledge has enabled increasingly sophisticated applications of synthetic biology. In 2010, researchers at the J. Craig Venter Institute announced the creation of a partially synthetic species of bacterium, *Mycoplasma laboratorium,* derived from the genome of *Mycoplasma genitalium* [32]. More of such synthetic bioengineering are expected in the near future.

CONCLUSION

Recent advances in genomic technologies have led to an explosion of data and a rapid growth in bioinformatics within biotechnology and the broader biomedical sciences. The advances in bioinformatics have in turn greatly advanced the development of the field of genomics. Biotechnology demands intelligent searching and filtering of numerous, complex data, to address specific issues, cutting across specialist research fields outside the knowledge of any one person, hence the importance of bioinformatics. Applications of bioinformatics include prediction of protein structure from genome analysis, medicine and health care, genomics and metabolomics research, agriculture, pharmaceuticals, biotechnology, etc. Though bioinformatics is rapidly expanding its applications alongside the rise of the new "-omic" technologies such as genomics, transcriptomics, metabolomics, and post-genomic technologies such as proteomics, its focus and strengths remain in the analysis of DNA sequences and genomes of living organisms, with their ever increasing applications. Despite the achievements in genomics research brought about by advances in bioinformatics, there are still challenges in applications of specific methods. It is therefore recommended that more methods should to be developed to tackle the problems encountered in the applications of bioinformatics in genomics research.

ACKNOWLEDGEMENTS

The authors would like to appreciate the Professors and Senior Lecturers of the Department of Biochemistry, Faculty of Medical Sciences, University of Jos, Jos, Plateau state, Nigeria; for their guidance and contributions to make this work successful. They are also grateful to the Scholars and Publishers of all scientific articles cited in this review manuscript.

REFERENCES

1. National Human Genome Research Institute. A brief Guide to Genomics, 2010. www.genome.gov. Retrieved 3/12/2011.
2. San Francisco. Concepts of Genetics, 10th ed. (Pearson Education, 2012). ISBN9780321724120.
3. J. Pevsner. Bioinformatics and Functional genomics, 2nd ed. (Hoboken, N. J.: Willey Blackwell, 2009). ISBN97804700851.
4. S. P Yadav. The wholeness in suffix – "omics", -"omes" and the word "om". Journal of Biomolecular techniques 2007; **18**(5): 277.
5. R.A. Ankeny. Sequencing the genome from nematode to human: changing methods, changing science. Endeavour 2003; **27**(2):87-92.
6. R.W. Holley, G. A. Everett, J. T. Madison, A. Zamir. Nucleotide sequences in the yeast alanine transfer Ribonucleic acid. J Biol Chem 1965; **240**(5):2122-28.
7. R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. I. I. Merrill, J. R. Penswick, A. Zamir. Structure of a Ribonucleic acid. Science 1965; **147**(3664):1462-65.
8. M. Nirenberg, P. Leder, M. Bernfield, R. Brimacombe, J. Trupin, F. Rottman, C. O' Neal. RNA code words and protein synthesis, VII: On the general nature of the RNA code. Proc Natl Acad Sci USA 1965; **53**(5): 1161-68.
9. W. Min jou, G. Haegeman, M. Ysebaert, W. Fiers. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature 1972; **237**(5350):82-88.
10. R. V. Eck, M.O. Dayhoff. Evolution of the structure of Ferredoxin based on the living Relics of primitive amino acid sequences. Science 1966; **152**(3720):363-66.
11. W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min jou, I. Molemans, A. Racymaekers, A. Vanden Berghe, G. Volckaert, M. Ysebaert. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 1976; **260**(5551):500-07.
12. W. Fiers, R. Contreras, G. Haegeman, R. Rogiers, A. Vande voorde, H. J. Van Heuverswyn, Van Heneweghe, G. Volckaert, M. Ysebaert. Complete nucleotide sequence of SV40 DNA. Nature 1978; **273**(5658):113-20.
13. R. H. Tamarin, Principles of genetics, 7th ed. (London: McGraw Hill, 2004).
14. F. Sanger, Nobel Lecture: Determination of nucleotide sequences in DNA, 1980. www.nobelprize.org. Retrieved 18/10/2010.

15. F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, M. Smith. Nucleotide sequence of bacteriophage phiX174 DNA. Nature 1997; **265**(5596):687-95.
16. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, D. L. Wheeler. GenBank Nucleic Acids Research 2008; **36** (Database issue):D25 – 30.
17. R. D. Heischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. I. Tom, B. A. Dougherty et al. Whole-genome random sequencing and assembly of *Haemophilus influenza Rd*. Science 1995; **269**(5223):496-512.
18. A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, E. Galibert, J. D. Hoheised, C. Jacq et al. Life with 6000 genes. Science 1996; **274**(5287):546-63.
19. National Council of Biotechnology Information. Complete genomes: Viruses, 2011. Retrieved 18-11-2011.
20. Z. Entre. Genome Project, Genome Project Statistics, 2011. www.genomeproject.com. Retrieved 18-11-2011.
21. L. Stein. Genome annotation: From Sequence to Biology. Nature Reviews Genetics 2001; **2**(7):493-503.
22. V. Mc Elheny. Drawing the map of life: Inside the Human Genome Project (New York NY: Basic Books, 2010). ISBN 9780465043330.
23. G. A. Mc Vean, D. M. Abecasis, R. M. Auton, G. A. R. Brooks, D. R. Depristo, A. Durbin, A. G. Handsaker, et al. An integrated map of genetic variation from 1092 human genomes. Nature 2012; **491**(7422):56-65.
24. R. Nielsen. Genomics: In search of rare human variants. Nature 2010; **467**(7319):1050-51.
25. B. Monya. Benchtop sequencers Ship off (Blog). Nature News Blog, 2012. Retrieved 22-12-2012.
26. M. Quail, M. E. Smith, P. Coupland, T. D. Otto, et al. A tale of three next generation sequencing platforms: Comparison of Ion torrent, Pacific Biosciences and Illumina Miseq sequencers. BMC Genomics 2012; **13**: 341.
27. N. Hall. Advanced sequencing technologies and their wider impact in microbiology. Journal of Experimental Biology 2007; **210**(9):1518-25.
28. J. R. Ten Bosch, W. W. Grody. Keeping up with the next generation. The Journal of Molecular Diagnotics 2008, **10**(6):484-92.
29. M. Pop. Genome assembly reborn: Recent computational challenges. Briefings on Bioinformatics 2009; **10**(4):354-66.
30. P. S. G. Chain, D. V. Grafham, R. S. Fulton, M. G. Fitzgerald, J. Hostetler, D. Muzny, J. Ali, et al. Genome project standards in a new era of sequencing. Science 2009; **326**(5950): 236-37.
31. F. W. Gregory, A. E. Guttmacher, K. L. Hudson. Genomic Medicine: Genomics, Health Care and Society. The New England Journal of Medicine 2011; **365**(11):1033-41.
32. M. Baker. Synthetic genomes: The next step for the synthetic genome. Nature 2011; **473**(7347):403-05.
33. E. Londin, P. Yadav, S. Surrey, L. J. Kricka, P. Fortina. Use of Linkage Analysis, Genome-wide Association Studies and Next-Generation Sequencing in the Identification of Disease-causing Mutations. Pharmacogenomics Methods in Molecular Biology 2013; **1015**:127-46.
34. R. Staden. A strategy of DNA sequencing employing computer programs. Nucleic Acids Research 1979; **9**(7):2601-10.
35. D. Edwards, J. Batley. Plant bioinformatics: From genome to phenome. Trends in Biotechnology 2004; **22**(5):232-37.
36. S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman. Basic Local Alignment Search Tool. Journal of Molecular Biology 1990; **215**(3):403-10.
37. D. Baum. Reading a phylogenetic tree: The meaning of monophyletic groups. Nature Education 2008; **1**(1): 190.
38. Wikipedia. Bioinformatics, 2014. www.wikipedia.com