International Journal of Biomedical Data Mining

Sharma, Int J Biomed Data Min 2016, 5:2
DOI: 10.4172/2090-4924.1000122

Review Article Open Access

A Review on New Horizons of Bioinformatics in Next Generation Sequencing, Viral and Cancer Genomics

Rahul Kumar Sharma*

School of Bioengineering, SRM University, Tamil Nadu, India

***Corresponding author:** Rahul Kumar Sharma, Department of Bioinformatics, School of Bioengineering, SRM University, Tamil Nadu, India, Tel: 7674005910; E-mail: bif.rahulsharma@gmail.com

Received date: Jul 08, 2016; **Accepted date:** Aug 01, 2016; **Published date:** Aug 07, 2016

Abstract

Genomics and molecular biology has always been a constant source of inspiration and motivational research for worldwide researchers in field of biology and biotechnology. These two fields have always generated a huge amount of data and in order to compile and analyze those, bioinformatics came into action during last decade. Implementation of bioinformatics has a clear intention of doing all these analysis of data in efficient and fast manner in order to cut down the expensive laboratory equipment, chemicals and most precious time. Mostly genomic data is composed of sequencing results at a higher scale and that is why manual curating and handling of these data is quite difficult. Supreme aim of this review is to make awareness about bioinformatics options in cancer genomics and viral genomics apart from next generation sequencing. Next generation sequencing or high throughput sequencing has helped a lot to replace old conventional method of sequencing and with the help of recent advances in technologies. These technologies are very efficient, fast and cheaper also as compare to conventional methods. Recently used methods were also discussed under appropriate headings. Role of bioinformatics is getting bigger and bigger in management and analysis of enormous amount of biological data generated through medical, biotechnological and clinical research globally. But we still need to understand challenges and limitation of bioinformatics as well as reliability.

Keywords: Bioinformatics; Next generation sequencing; Viral genomics; Ebola outbreak; Somatic cancer; Cancer genomics

Introduction

In last decade most of the biological research revolved around molecular biology and genomics related studies. In genomic studies a major portion of study was focused on comparative genomics [1-4] and genome sequencing [5-9]. As soon as human genome project completed during 2003, there was an outburst in number of genome sequencing project was observed. This was because of a belief that all the complication related to human or any other organism is somewhere related to its genome composition and variation among these.

Genome sequencing techniques in early era were limited to Sanger sequencing method and Maxam-Gilbert sequencing method. Sanger sequencing method was anyways also called as chain termination method. And also these methods were too much expensive as well as time consuming. And thus high throughput sequencing methods were taken into consideration [9-14]. Maxam-Gilbert sequencing method was one of the early DNA sequencing methods where any DNA sequence can be determined using synthetic location-specific primers during 1973. Later on there were several modification and Sanger at MRC center, Cambridge, UK and demonstrated a new method for DNA sequencing as DNA sequencing [15-22] with chain-terminating inhibitors in 1977. By using this method later scientists from MRC center also displayed the first complete genome sequencing for Epstein-Barr virus in 1984, which was composed of 172,282 nucleotides. The interesting fact was in this that there was no prior knowledge about genetic profile [23-25] of this virus was known.

Sequencing data and management

After sequencing of small viruses now it was time for go for large sequencing projects. Even sequencing data for small virus was also very big and thus data handling was an issue for conducting these researches. This is where bioinformatics tool [26-30] helped in data management. Integrating these biological data with SQL and other programming languages helped a lot to support compilation and analysis of available data. This data handling was separated in two different departments; one was to create the complete database in forms of rows and columns as in table and the second part was to manage the available data.

Implementation of SQL was a great added advantage for bioinformatics as it was very easy to use and command lines were not very complex as other programming languages. And that was the reason; it became very popular and easy to handle by biological researchers. In management of data also there was a part of querying about select and view mostly and helped in various data mining [31-33].

Impact of NGS technology in virology

Viruses are the most abundant and the smallest organisms on this planet, which are comparatively simple to sequence. Although available data offers an opportunity to study viral diversity and taxonomic hierarchy at various levels, it also challenges for systematic and structured organization of data and its downstream processing as well. Extensive computational analyses using a number of algorithms and programs have opened exciting opportunities for virus discovery and diagnostics, in which bioinformatics played a vital role or key player. Molecular analysis of viruses using data generated by NGS has revolutionized complete idea of virology. The main idea of bioinformatics was to analyze sequence, structure and function relationships, but eventually also resulted in the development of new areas of research such as phyloinformatics and immunoinformatics, which translates raw data into information about evolutionary history and interaction of protein bodies [34-38].

Bioinformatics methods for viral genomics

Bioinformatics approaches help to estimate and analyze population diversity by studying genetic recombination, mutation, selection and, thereby, assist in correlation of genotype to phenotype. There are plenty of methods available, among which some of them are discussed below [39-43].

Quasispecies reconstruction: Quasispecies reconstruction is calculation of number of viral variants and their frequency. Every viral variant in a quasispecies is considered as a haplotype. Several tools can be implemented for this process, which include Short Read Assembly into Haplotypes, Quasispecies Reconstruction algorithm and QuasiRecomb.

Population genetics studies: Genetic structure of a population refers to the number of distinct subpopulations, identified using a characteristic set of allele frequencies. A population analysis can be performed using the model based STRUCTURE program using available genomic data. The program can infer the genetic structure in haploid, diploid and polyploid species as per requirement [44,45]. Simulation studies in population genetics play an important role in helping to better understand the impact of various evolutionary and demographic scenarios on sequence variation and sequence patterns, and they also permit investigators to better assess and design analytical methods in the study of disease-associated genetic factors. To facilitate these studies, it is imperative to develop simulators with the capability to accurately generate complex genomic data under various genetic models. Currently, a number of efficient simulation software packages for large-scale genomic data are available, and new simulation programs with more sophisticated capabilities and features continue to emerge. There are three basic simulation frameworks termed as coalescent, forward, and resampling.

Linkage equilibrium: Linkage equilibrium is actually the statistical independence of alleles at all loci and indicates evidence of free recombination. Thus, linkage disequilibrium is a measure of the correlation between the occurrences of nucleotides at different location of a complete genome. The extent to which recombination occurs can be estimated by specialized programs such as Linkage Analysis and DNA Sequence Polymorphism.

Pressure analysis: The selection pressure can be classified as pervasive and episodic. Various statistical methods for analysis of pervasive and episodic selection are available at the Datamonkey web-server of Hypothesis testing using Phylogenies software package.

Phylogenetic analysis for viruses: Whole genome-based phylogenetic trees are widely used for various viruses owing to their small genome sizes and conservation of genomic structure. Phylogenomics is getting popularity to monitor epidemiology and disease surveillance, in particular. This field when analysed in the context of spatio-temporal data helps to understand the disease spread and progression during sudden outbreaks. The program such as Bayesian Evolutionary Analysis by Sampling Trees (BEAST) is exclusively designed for phylogeography studies and is used widely to study spatio-temporal dynamics of viruses at population scale. BEAST software provides a Bayesian Markov chain Monte Carlo (MCMC) framework for parameter estimation and hypothesis testing of evolutionary models from molecular sequence data. It brings together a large number of evolutionary models into a single coherent framework for evolutionary inference.

NGS has become extremely useful and an integral part of virus research and opened up new horizons in studying viral evolution. Recent utilization of this technique was for the characterization of the Ebola virus infection in West Africa during 2014.

Current Challenges of Next Generation Sequencing (NGS)

Now we have seen enough number of applications of bioinformatics as well as NGS in our ongoing and future researches. Most important challenge of NGS and bioinformatics is to implement these results into real medicine research or can say clinical translation of these results. As we can see in a study by Sandeep Pingle in illumine blog about cancer genomics, to detect direct somatic cancer genome there can be 3 major approaches.

1. Whole genome sequencing,
2. Whole exome sequencing,
3. RNA sequencing (Transcriptome)

This somatic cancer genome alteration may be nucleotide substitution, copy number variation, insertion or deletion or it may be a chromosomal rearrangement. These kinds of studies can reveal not only about a clear pathogenesis of article, it may also lead to identification of certain important biomarkers for future target of drug development [46,47].

During this study one of the challenges may be quality and quantity of samples available. It can be overcome with increasing sequencing depth, which can ultimately increase low sample purity and increase ploidy.

Most of the data obtained with state-of-the-art sequencers is in the form of short reads [48]. Hence, analysis and interpretation of these data encounters several challenges, including those associated with base calling, sequence alignment and assembly, and variant calling. These challenges have led to the development of innovative computational tools and bioinformatics approaches to facilitate data analysis and clinical translation [49,50].

NGS and its role in personalized medicine

The potential of next-generation sequencing (NGS) to revolutionize personalized medicine and to peer into our genetic studies are very high. While recent technological advances in NGS have propelled our knowledge and understanding of genomics forward, several technical challenges still remain in order to gain that next level of understanding and clinical utility. These challenges need to be discovered and resolved with maximum available possibilities.

Current improvements and highlights

Almost 600 bioinformatics tools were developed during this period of 2012-14 to address these challenges and they are being used for data analysis and data interpretation. Some of these tools can detect quality of short reads, as for example Fast QC and htSeqTools. A tool called as Mutect can be used for sequence alignment with low allele fractions. The tool called MuSiC is a mutational analysis pipeline, which can also help in establishing correlation between mutation, genes and Pathways. As this tool uses sequencing data in addition to clinical information, it has ability to differentiate between passenger mutation as well as driver mutation.

Future perspectives

In current scenario, Multiple bioinformatics tools are being used by cancer researchers, among that everyone have specific requirements because cancer genome data:

1. Needs to be analyzed in association with normal matched genome
2. Contains highly rearranged genomes, and
3. Have enormous heterogeneity

But still there is hope of development of a single interface tool or software, which can be utilized to detect all the anomalies in sequence data from somatic cancer cells.

Conclusion

With support of these data and current developments in field of bioinformatics tools, we may hope for a better tool associated with cancer genomic studies which can be used for both clinical information as well as next generation sequence data.

References

1. Irizarry KJL, Punt N, Bryden R, Bertone J, Drechsler Y (2016) Leveraging Naked Mole Rat (Heterocephalus glaber) Comparative Genomics to Identify Canine Genes Modulating Susceptibility to Tumorigenesis and Cancer Phenotypes. J Veterinar Sci Techno 7: 322.
2. Nava GM, Hernandez YE (2016) Comparative Genomics of Salmonella Could Reveal Key Features of Adaptation. J Data Mining Genomics & Proteomics 7: e121.
3. Ali A, Soares SC, Barbosa E, Santos AR, Barh D, et al. (2013) Microbial Comparative Genomics: An Overview of Tools and Insights Into The Genus Corynebacterium. J Bacteriol Parasitol 4: 167.
4. Kumar M, Balajia PV (2014) Diversity, Abundance and Distribution of O-linked Glycosylation Pathway Enzymes in Prokaryotes-A Comparative Genomics Study. J Glycomics Lipidomics 4: 117.
5. Ebbesen M, Sundby A, Pedersen FS, Andersen S (2015) A Philosophical Analysis of Informed Consent for Whole Genome Sequencing in Biobank Research by use of Beauchamp and Childress' Four Principles of Biomedical Ethics. J Clin Res Bioeth 6: 244.
6. Ahmed I (2015) Chloroplast Genome Sequencing: Some Reflections. Next Generat Sequenc & Applic 2: 119.
7. Dhillon V, Li X (2015) Single-Cell Genome Sequencing for Viral-Host Interactions. J Comput Sci Syst Biol 8: 160-165.
8. Blum HE, Oexle K (2014) Clinical Interpretation and Implications of Whole Genome Sequencing. Next Generat Sequenc & Applic 1: 105.
9. Gryganskyi AP, Muszewska A (2014) Whole Genome Sequencing and the Zygomycota. Fungal Genom Biol 4: e116.
10. Ebomoyi EW (2011) Establishing Genome Sequencing Centers, the Thematic Units in the Developing Nations and the Potential Medical, Public Health and Economic Implications. J Drug Metab Toxicol 2: 108.
11. Hens K (2011) Whole Genome Sequencing of Children's DNA for Research: Points to Consider. J Clinic Res Bioeth 2: 106e.
12. Santos RCV, de Almeida Vaucher R, Alves SH (2012) Current Trends in Sporotrichosis. Fungal Genom Biol 2: e105.
13. Sheehan D (2013) Next-Generation Genome Sequencing Makes Non-Model Organisms Increasingly Accessible for Proteomic Studies: Some Implications for Ecotoxicology. J Proteomics Bioinform 6: e21.
14. Alharbi KK, Khan IA, Tejaswini YRSN, Devi YA (2014) The Role of Genome Sequencing in the Identification of Novel Therapeutic Targets. J Glycomics Lipidomics 4: 112.
15. Ansorge WJ (2016) Next Generation DNA Sequencing (II): Techniques, Applications. Next Generat Sequenc & Applic S1: 005.
16. Marzooq Ammar AL (2015) Discovery of Novel DNA Variants in Jordanians Population by Re- Genotyping Affymetrix DMET Arrays Data Using DNA Sequencing. Mol Biol 4: 126.
17. Cai ZX, Tang XD, Gao HL, Tang C, Nandakumar V, et al. (2014) APC, FBXW7, KRAS, PIK3CA, and TP53 Gene Mutations in Human Colorectal Cancer Tumors Frequently Detected by Next-Generation DNA Sequencing. J Mol Genet Med 8: 145.
18. Sharma S, Madan M (2014) Detection of Mutations in rpob Gene of Clinically Isolated M. tuberculosis by DNA Sequencing. J Mycobac Dis 4: 156.
19. Sun Q, Xu X, Zhang Qq, Wang Hy, Liu Y (2013) Diagnostic Direct DNA Sequencing and Systemic Treatment with Voriconazole inScedosporium apiospermum Keratitis? A Case Report. J Clin Exp Ophthalmol 4: 299.
20. Krstic PS (2012) Challenges in Third-Generation DNA Sequencing. J Nanomed Nanotechol 3: e116.
21. Lu C, Yu P (2012) Biological and Solid-State Nanopores for DNA Sequencing. Biochem Pharmacol (Los Angel) 1: e109.
22. Miller NA, Kingsmore SF, Farmer AD, Langley RJ, Mudge J, et al. (2008) Management of High-Throughput DNA Sequencing Projects: Alpheus. J Comput Sci Syst Biol 1: 132-148.
23. Demirhan O, Tanriverdi N, Suleymanova D, Cetinel N (2015) Cytogenetic Profiles of 1213 Children with Down Syndrome in South Region of Turkey. J Mol Genet Med 9: 157.
24. Orlando RA (2014) Combining Genetic Profiles and Energy Expenditure Measurements to Guide Weight Loss Management Programs. J Biomol Res Ther 3: e135.
25. Hong Y, Chalkia D, Ko KD, Bhardwaj G, Chang GS, et al. (2009) Phylogenetic Profiles Reveal Structural and Functional Determinants of Lipid-binding. J Proteomics Bioinform 2: 139-149.
26. Su J, Huang D, Yan H, Liu H, Zhang Y (2012) Advances in Bioinformatics Tools for High-Throughput Sequencing Data of DNA Methylation. Hereditary Genet 1: 107.
27. Hashemi M, Behrangi N, Borna H, Akbarzadeh A, et al. (2012) Evaluating New Targets of Natural Anticancer Molecules through Bioinformatics Tools. J Proteomics Bioinform 5: 050-053.
28. Nanda T, Tripathy K, Ashwin P (2011) Integration of Bioinformatics Tools for Proteomics Research. J Comput Sci Syst Biol S13.
29. Mehmood MA, Sehar U, Ahmad N (2014) Use of Bioinformatics Tools in Different Spheres of Life Sciences. J Data Mining Genomics Proteomics 5: 158.
30. Najafi M (2012) Bioinformatics Tools as Power Hypothetical Predictors. Biochem Physiol 1: e113.
31. Banerjee AK (2015) Computation in Analyzing Inflammation: A General Perspective. Interdiscip J Microinflammation 2: 130.
32. Murty USN, Amit KB, Neelima A (2009) An In Silico Approach to Cluster CAM Kinase Protein Sequences. J Proteomics Bioinform 2: 097-107.
33. Amit KB, Neelima A, Varakantham P, Murty USN (2008) Exploring the Interplay of Sequence and Structural Features in Determiming the Flexibility of AGC Kinase Protein Family : A Bioinformatics Approach. J Proteomics Bioinform 1: 077-089.
34. Pukhovskaya NM, Vysochina NP, Bakhmetyeva SV, Zdanovskaya NI, Belozerova NB, et al. (2016) Detection of the Insect-Specific Flavivirus Chaoyang in Mosquitoes in the Jewish Autonomous Region of the Far East of Russia. J Neuroinfect Dis 7: 205.
35. Abubakar AG, Ozumba PJ, Winter J, Buttner P, Abimiku A (2015) Current Trends in the detection of Acute HIV Infection among Blood Donors: Reliability of Pooled Nucleic Acid Amplification Technology and the Need for Population Specific Algorithms: A Systematic Review. J Antivir Antiretrovir 7: 089-103.

36. Pukhovskaya NM, Belozerova NB, Bakhmetyeva Sv, Zdanovskaya NI, Ivanov LI and Morozova OV (2015) Isolation of the Tick-Borne Encephalitis Virus from Mosquito in Khabarovsk Region of the Far East of Russia. J Neuroinfect Dis S2: e001.
37. McCullough KC, Milona P, Démoulins T, Englezou P, Ruggli N (2015) Dendritic Cell Targets for Self-Replicating RNA Vaccines. J Blood Lymph 5: 132.
38. Borgoyakova MB, Karpenko LI, Ilyichev AA (2015) Isolation And Studying of Specificity of Bacteriophages Binding To Murine Lung Adenocarcinoma. Biol Med S2: 002.
39. Balakrishnan A, Shahir P (2012) Early IgM Antibody Response in Chandipura Virus Infection: T cell- Independent Activation of B-cells. J Clin Cell Immunol 3: 123.
40. Rusnati M (2012) The Impact of Surface Plasmon Resonance in Virology. J Bioeng Biomed Sci 2: e108.
41. Boretska M, Bellenberg S, Moshynets O, Pokholenko I, Sand W (2013) Change of Extracellular Polymeric Substances Composition of Thiobacillus thioparus in Presence of Sulfur and Steel. J Microb Biochem Technol 5: 068-073.
42. Okpokoro E, Osawe S, Datong P, Yakubu A, Ukpong M, et al. (2013) Preparing for HIV Vaccine Trials in Nigeria: Building the Capacity of the Community and National Coordinating, Regulatory and Ethical Bodies. J AIDS Clin Res 4: 260.
43. Stone CB, Mahony JB (2014) Molecular Detection of Bacterial and Viral Pathogens–Where Do We Go From Here?. Clin Microbiol 3: 175.
44. Zhang Y (2011) Population Genetics for 15 STR loci of Liaoning Han in Northeastern China. J Forensic Res 2: 123.
45. Dogra D, Shrivastava P, Chaudhary R, Gupta U, Jain T (2015) Population Genetics for Autosomal STR Loci in Sikh Population of Central India. Hereditary Genet 4: 142.
46. Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865-2871.
47. Ranuncolo SM (2016) Towards the Dreamed Biomarkers? J Mol Biomark Diagn S2: e002.
48. Fenghai D (2016) The Use of Molecular and Imaging Biomarkers in Lung Cancer Risk Prediction. J Biom Biostat 7: 299.
49. Du R, Mercante D, An L, Fang Z (2014) A Statistical Approach to Correcting Cross-Annotations in a Metagenomic Functional Profile Generated by Short Reads. J Biomet Biostat 5: 208.
50. Jakobsen JC, Tamborrino M, Winkel P, Haase N, Perner A, et al. (2015) Count Data Analysis in Randomised Clinical Trials. J Biomet Biostat 6: 227.