British Biotechnology Journal
13(4): 1-17, 2016, Article no.BBJ.25662
ISSN: 2231–2927, NLM ID: 101616695

SCIENCEDOMAIN *international*
www.sciencedomain.org

Current Opportunities and Challenges of Next Generation Sequencing (NGS) of DNA; Determining Health and Diseases

Carlo P. J. M. Brouwer[1,2,3*], Thuy Duong Vu[1], Miaomiao Zhou[1], Gianluigi Cardinali[4], Mick M. Welling[5], Nathalie van de Wiele[1] and Vincent Robert[1,3]

[1]CBS-KNAW Fungal Biodiversity Center, Uppsalalaan 8, Utrecht 3584 CT, The Netherlands.
[2]CBMR Scientific Inc., Suite 161, 2057-111 Street NW, Edmonton, Alberta, Canada.
[3]BioAware Life Sciences Data Management Software, Rue du Henrifontaine 20, B-4280 Hannut, Belgium.
[4]University of Perugia, Department Applied Biology-Microbiology, Borgo 20 Giugno, 74, I-06121 Perugia, Italy.
[5]Departement of Radiology, Leiden University Medical Center, Interventional Molecular Imaging Laboratory, Room C2 -204, Leiden, The Netherlands.

Authors' contributions

This work was carried out in collaboration between all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BBJ/2016/25662
Editor(s):
(1) Chung-Jen Chiang, Department of Medical Laboratory Science and Biotechnology, China Medical University, Taiwan.
Reviewers:
(1) Michael R. Emmert-Buck, Avoneaux Medical Institute, Oxford, USA.
(2) Molobe Ikenna Daniel, International Institute of Risk and Safety Management (IIRSM), Nigeria.
(3) Triveni Krishnan, National Institute of Cholera and Enteric Diseases, Kolkata, India.
Complete Peer review History: http://sciencedomain.org/review-history/14483

Review Article

Received 14th March 2016
Accepted 8th April 2016
Published 6th May 2016

ABSTRACT

Many publications have demonstrated the huge potential of NGS methods in terms of new species discovery, environment monitoring, ecological studies, etc. [24,35,92,97,103]. Undoubtedly, NGS will become one the major tools for species identification and for routine diagnostic use. While read lengths are still quite short for most existing systems ranging between 50 bp and 800 bp, they are likely to improve soon. This will enable easier, faster, and more reliable contig assembly and

[*] Corresponding author: E-mail: c.brouwer@cbs.knaw.nl;

subsequent matching against reference databases. When data generation is no longer a bottleneck, the storage, speed of analysis, and interpretation of DNA sequence data are becoming the major challenges. Also, the integration or the use of data originating from diverse datasets and a variety of data providers are serious issues that need to be addressed. Poor sequence record annotations and species name assignments are known problems that should be instantly addressed and would allow the creation of reference databases used for routine diagnostics based on NGS. Samples with huge amounts of short DNA fragments need to be analyzed and compared against reference databases in an efficient and fast way. Although a number of solutions have been proposed by Industry; offering commercial software, there still remain hurdles to take. One of the challenges that we need to address is data upload from client's computers to central or distributed data storage and analysis services. Another one is the efficient parallelization of analyses using cloud or grid solutions. The reliability and up-time of storage and analyses facilities is another important problem that need to be addressed if one wants to use it for routine diagnostics. Finally, the management, reporting and visualization of the analyses results are among the last issues, but not the least challenging ones. Considering the constant growth of computational power and storage capacity needed by different bioinformatics applications, working with single or a limited number of servers is no longer realistic. Using a cloud environment and grid computing is becoming a must. Even single cloud service provider can be restrictive for bioinformatics applications and working with more than one cloud can make the workflow more robust in the face of failures and always growing capacity needs. In this white paper we review the current state of the art in this field. We discuss the main limitations and challenges that we need to address such as; data upload from client's computers to central or distributed data storage and analysis services; efficient parallelization of analyses using grid solutions; reliability and up-time of storage and analyses facilities for routine diagnostics; management, retrieving and visualization of the analyses results.

Keywords: DNA; next generation sequencing; bioinformatics; clinical approaches, limitations.

1. INTRODUCTION

Bioinformatics is a catalyzer of today's life sciences research. Its development and impact in life sciences is fundamental to understand the scientific progress of the last decades. Biomedical research includes large files of molecular imaging modalities (like MRI, SPECT, PET and opticals), microarray and proteomics data; and nowadays massive sequencing data which became rapidly available from the worldwide web. Bioinformatics fosters the development of computational solutions that facilitate a qualitative and quantitative understanding of life; it supports the interpretation of data coming from life sciences experiments. It is a multidisciplinary area which requires a collaborative effort. Considering that microbial life dominates the world, still many species remain undetectable using conventional culturing assays and make them impossible or difficult to study. By DNA sequencing directly from the environment (metagenomics), culture independent approaches are then possible. More recently, metagenomics present significant new challenges in data analysis. Metagenomic datasets are large collections of sequencing reads from anonymous species within particular environments. Computational metagenomics are extremely time consuming, and there are often many novel sequences in these metagenomes that are not fully utilized yet. Recent technological advances that become available allow faster and cheaper DNA sequencing are now driving biological and medical research. The past years have seen the arrival of high throughput sequencing (HTS) also known as Next Generation Sequencing (NGS) [1-3]. These technologies drastically lowered sequencing costs and increased sequencing throughput [4-6]. They radically changed molecular biology and computational biology, as data generation is no longer the bottleneck. In fact, nowadays a major challenge is the analysis and interpretation of huge amount of available sequencing data [7, 8]. But new knowledge of collected data needs to be discovered as well. The way of thinking has a strong impact on how we deal with statistical, methodological and theoretical studies. New insights need to be discovered and tools need to be developed. Interestingly, most sources created by a variety of companies or research laboratories are open. The use of public or private tools and services provided by multiple institutions that easily aggregate to the federated cloud [7,9-15] or Bio Torrents [16] could be a strategy for scheduling services. To improve the efficiency a cloud provider could execute the job. The executing time is strongly affected by the file transfer and storage services. Unfortunate, the lack of automated, integrated data and management tools [17], poor links to laboratories, clinical and epidemiological data during infectious outbreaks can still inhibit an effective and adequate response to combat diseases. Genomic sequence annotation requires an up to date comprehensive database of DNA sequence information for a given organism, preferentially those including the whole life cycle of mammals, vertebrates, invertebrates and microorganisms. Benefits of databases accepted by the web include questions such as; how to deal with secured data access, increased capacity for data sharing and jurisdictions between different companies (privacy policies), how to increase efficiency and how to speed up the analysis processes.

In general, data refers to a collection of results and is available in a common format but commonly research institutes or organizations worldwide store data in their own format. This diversity of data sources caused by lack of collaboration includes; lack of universally accepted genetic loci, lack of reference genomes, limitations on identification using DNA sequence analysis, non-negligible proportion of compromised sequences (GenBank [18], EMBL [19], DDBJ [20]); separate names for different sexual stages and differences in taxonomy, sequences with wrong designations, differences between phylogenetically based annotations and BLASTn annotation, reliability of sequences in public databases and important morphological characters by overlapping between species [21-26]. Today there is still not an efficient way to assemble sequence data and the use of a standard set of validated tools available for epidemiological studies. An infection or contamination is a serious threat that is accompanied by great uncertainty and requires in most cases an urgent action by diseases such as bird flu or MERS. There are warning signs, but they are not seen or too late. Today's systems are very complex, they need to be continually updated and connected with the different authorities. A crisis is unavoidable, and could lead to an increase of insecurity. Furthermore most current systems are time consuming; it takes more than 48 hours in case an infection is detected. Spread of the disease is already a fact. By efficient collaboration [27], gaps can be bridged by implementing new technologies and geographic coordination. It is common for researchers to refer to one database in order to pursue the analysis of their data. Exploring biological pathways collected from different databases at the same time, may contain conflicting information and this may affect providing identical information for a certain pathway analysis. Many sequences containing unintentional mistakes as a result of experimental errors, misidentification of the species or by exchange of cultures and have limited taxonomic coverage. Reliability, data collected by evidence or by experimental design and after publication in peer-reviewed journals play a crucial role in decision of interpretation of the analysis and interest of researchers. Raw data produced to produce new data sets are analyzed using standard programs and do not reveal details of the entire collection, sources and processing data are mostly derived from different sources. Researchers usually select data from a single data source and commercially rapid available systems to perform functional analysis using a single data source; a pathway approach that emphasizes mapping and relationship inference based on data acquired from multiple data sources. Fact is that they still think in frameworks of single experiments with small to average amounts of samples. One solution for this is described into the thesis from de Vries, (2013) (http://hdl.handle.net/11245/1.385755). Today, one of the major challenges in research is how to integrate biological data and how to understand the inner working of the cell and coherency defined by complex interaction networks. It's important to identify the evidence of data derived from an earlier non updated existing data source. To be included as a candidate for integration, the data source should be represented by the evidence code. In most data integration algorithms, the user does not contribute thus leading to an integrated data source without any effective utility towards analysis (Luyf et al, 2011). The concept of data handling and analysis still follows the traditional concept of one singular analysis regarding the actual hypothesis, followed by raw data storage for archiving. Most researchers commonly do not think about alleged insignificant steps like fast data upload strategies, efficient storage and interim results or making data accessible for fast query and linkage and affiliation with third party data types and sources. Usually one or few connected steps in the analyze pipeline are investigated and compared. Reflected in the ongoing discussion about upcoming demands for improvement and acceleration of data analysis, several academic institutes discussed the need for improvement of a whole information organization [28,29] moving away from organization based on individual preferences and needs to an established wide data organization.

2. NEED ASSESSMENT

With the growing demand for data analysis capacities, feasible also for non-bioinformaticians, user friendly web-based platforms are needed. Platforms who provide the infrastructure for the whole workflow of rapid data starting from raw data to storage, management, analyzing and interpreting the huge amount of NGS data up to visualization of the results. Rapid developments within the field of high throughput sequencing, driven by frequent updates and price drops on sequencing tools will increase in exponential data output for the next years. A few years ago, this wasn't a big issue, as the amount of data and types of analysis were easily handled by local bioinformatics. Nowadays geneticits perform single experiments routinely that identify up to millions of variant sites in a single individual [30]. A full run on an Illumina HiSeq 2000 sequencer [31] makes it possible to sequence more than 5 human genomes at -30x coverage simultaneously, up to 192 gene expression samples in a single run (that generates approximately 600 GB of large data files). Companies who perform these facilities send data to their customers on Terabyte scale hard disc drives (Fig. 1). Such data volumes are not suitable for uploads to data analysis servers [32] and most publicly available databases are not useful for these large sequencing datasets either. Accelerating processing time and using new scientific workflows for analysis pipeline by offering GRID solutions could be successful. But still several bottlenecks have to be removed before optimal use of internet infrastructures for NGS experiments (Olabarriaga et al, 2010). Overwhelming amounts of data being generated and constant updates make this a challenging field. In general, high performance computer systems [33-34] are needed with clustered processors, high internal bandwidth to fast storage and the proper software to perform the complex multiple-step flow.

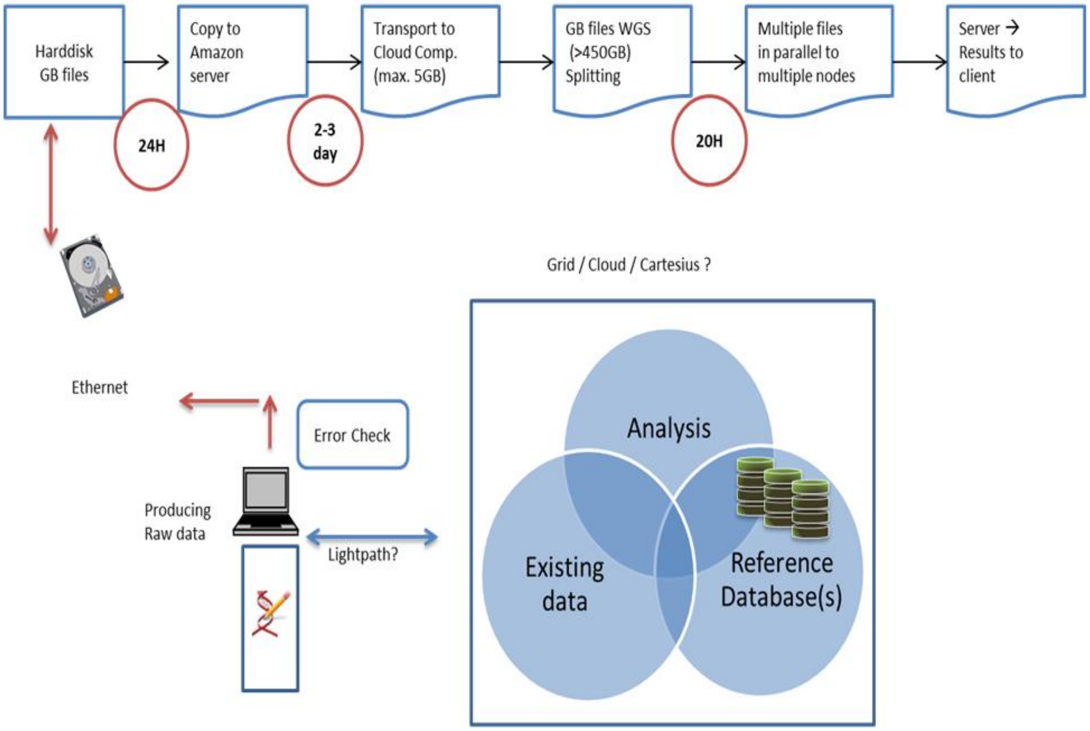

Harddisk GB files
Copy to Amazon server
Transport to Cloud Comp. (max. 5GB)
GB files WGS (>450GB) Splitting
Multiple files in parallel to multiple nodes
Server → Results to client
24H
2-3 day
20H
Grid / Cloud / Cartesius ?
Ethernet
Error Check
Producing Raw data
Lightpath?
Analysis
Existing data
Reference Database(s)

Fig. 1. Whole genome sequence analyse

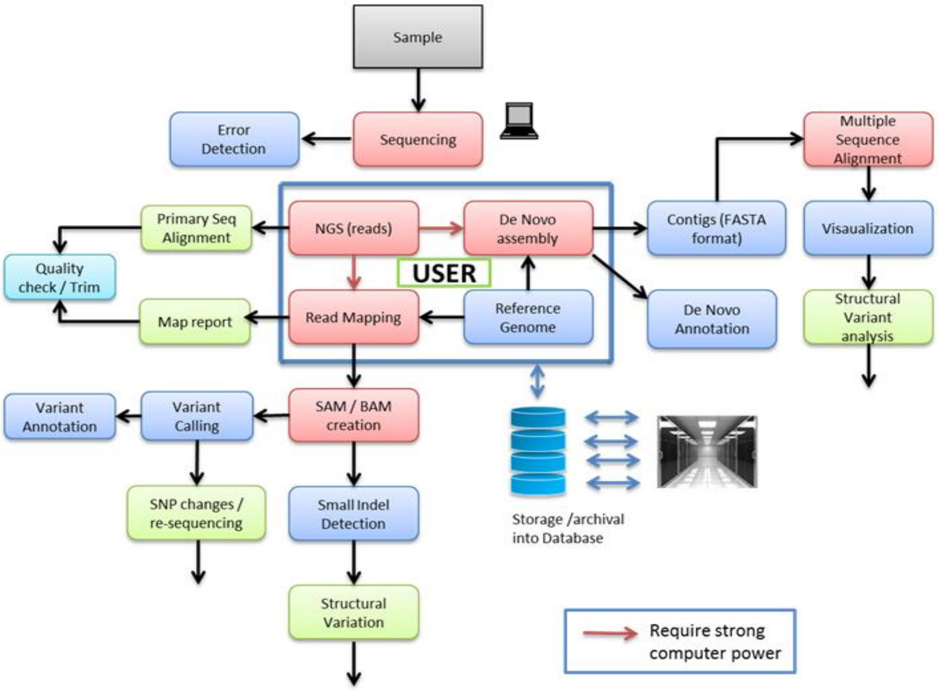
Schematic overview of the WGS pipeline. Red line indicate process and time of a single run of large data files send on terabyte scale hard disc drives. Blue line indicates control and distributes large volumes of data storage and bioinformatic programs in cloud environment. High performance of computer systems with high internal bandwidth for fast storage and complex multistep processing of data; efficient parallelization of analyses using cloud or grid solutions

2.1 Alignment and Software

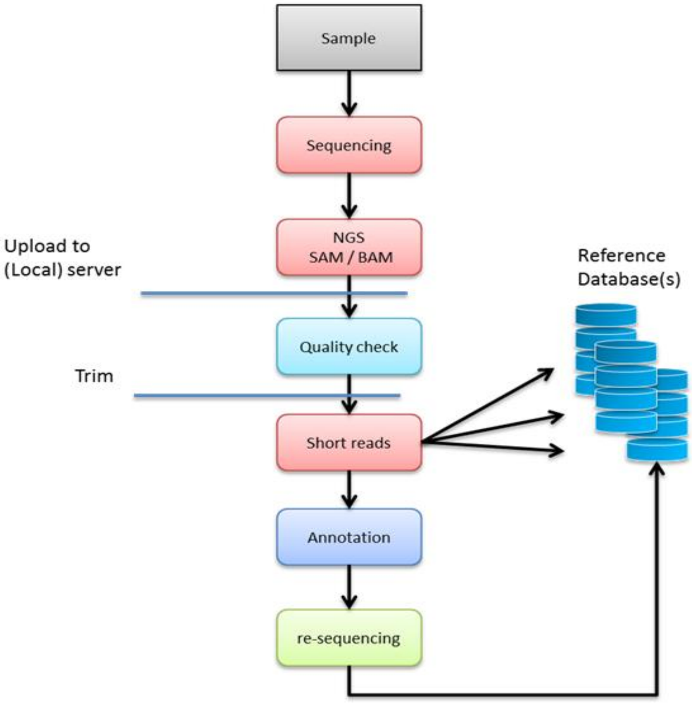
Most important for biological application is accurate alignment or assembly of raw short reads to a reference genome for which a variety of algorithms and software has been developed. Short lengths of the reads generated by NGS could limit the analysis [35-38] as algorithms of short reads alignment normally uses a two-step procedure; mapping and extending. In the first step of mapping, seed (short fragment of sequence with fixed length) reads are compared with the reference genome based on exact matches but the sensitivity of this approach is rather low. The difference between alignments algorithms mainly occur in the first step of mapping seeds. The next extension step, the rest portion of a read is fully mapped to the reference genome. This approach extension is the key to balancing alignment accuracy and speed. When the reference genome has many repeats; the number of hits (successful mappings) of a seed is very large, making the alignment very slow as it is time consuming [39]. Commercial software are under development (CLC Genomics Workbench) which allow insertions and deletions in the alignment by using spaced seeds that only treat substitutions as mismatches. With the increasing volume of NGS data, expected to double every two years, resequencing of reference strains, and functioning data like RNA-seq. data will be available from different databases (see below). To fully utilize the diverse types of NGS data for medicine the urgent need for assembly programs that can integrate different types of reads into similar assembly process are necessary (loci sequencing / DNA barcoding). Development of new assembly algorithms that utilize parallel computation on large scale processor clusters in all steps will be a challenge. The need for assemble programs that can efficiently assemble genomes based on multiple data sets with very diverse read lengths from different NGS databases are another issues that has been raised [35,40]. Furthermore, these advances have revealed a new problem in DNA sequencing; functional classification and interpretation of newly discovered genetic variations in genetic research and clinical applications of genomic technologies.

2.2 Databases and Storage

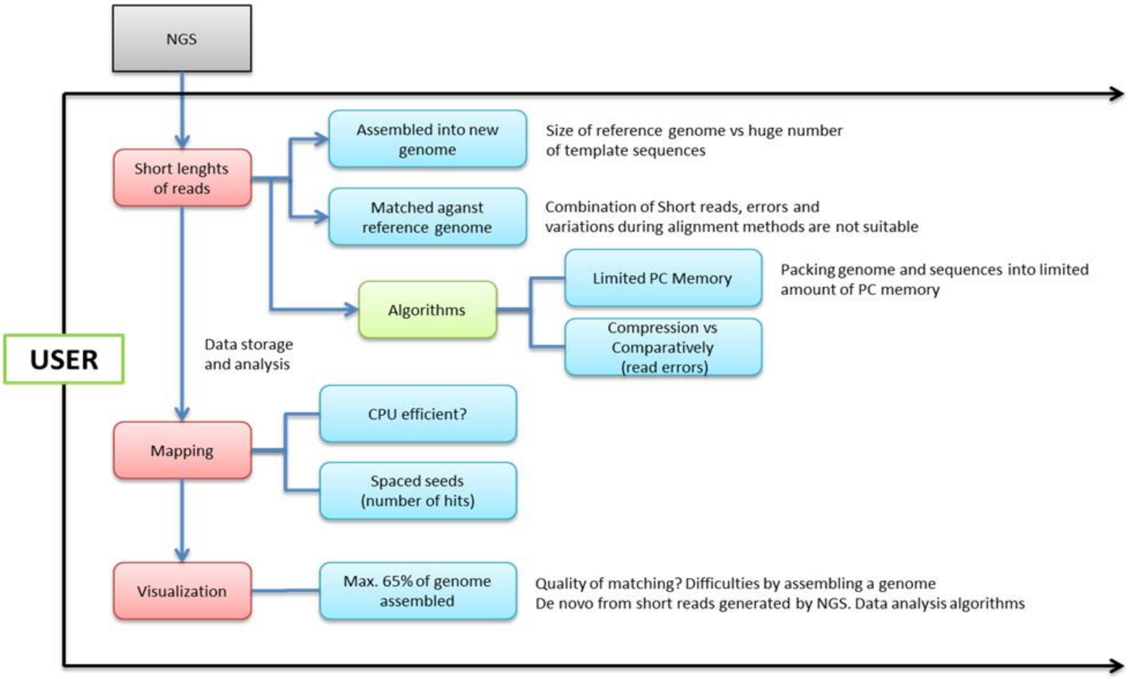
Genome databases such as the UCSC Genome browser [41] (http://genome.ucsc.edu/), Esembl [42] (http://www.ensemble.org), GenBank (NCBI) [18,43] (http://www.ncbi.nlm.nih.gov/), EMBL [19,44] (http://www.ebi.ac.uk/), DDBJ [20,45] (http://www.ddbj.nig.ac.jp/), The Genomes Online Database (GOLD) [46] (http://www.genomesonline. org/) and Mycobank [47,48] (http://www.mycobank.org/) accessible via web browsers are useful in the search for annotation information for DNA sequences. However, despite their capacities, the main limitation by using those databases lies in the limited amount of data that can be accessed at a given time. Furthermore, most free genome browsers do not support multiple genomes simultaneously, do not capture spatial or temporal information and cannot be customized. Researchers have to rely on commercially available software solutions. The new generation of DNA sequencing can generate large amounts of DNA sequence data. Platforms as Roche/454 (2004) [35,49], HiSeq (Illumina,2006) [31,35], ABI/SOLiD (2008) [35,50], Ion Torrent-Proton II [51], Helicos tSMS Sequencing [52], PACBIO RS II [53] and Oxford Nanopore minion [54] are capable of generating Giga base pairs of sequences in one single run and projects such as the 1000 Genomes project (http://www,1000 genomes.org) and the human Microbiome project (http://commonfund.nih.gov/hmp) are examples of projects generating on terabyte-scale amounts of DNA sequence data. Unfortunately, differences in methodologies between used NGS platforms result often in differences in the way data is represented. The way sequences are measured is not related to a formatting problem. The Illumina method differs from the Roche/454 and the SOLiD is even more complicated (as it contains base sequence translation errors). It means that separate pipelines have to be built for each platform. Here, the need for development of free available software that can cope with, and combine data from the different platforms comes into view. Furthermore, the large amount of produced data can only be handled by powerful computational infrastructures and architecture, sophisticated algorithms, efficient programs, and well-designed workflows (Figs. 2a, 2b, 2c). In recent years, different technologies and service providers have emerged to challenge the paradigm of cloud computing [7,9-15,55-57]. Users have transparent access to a wide variety of distributed infrastructures and systems. In this environment, computing and data storage necessities are accomplished in different and unanticipated ways to give the user the illusion that the amount of resources is unrestricted. In a collaborative environment, cloud computing is an interesting option to control and distribute processing of large volumes of data produced in genome sequencing projects and stored in public databases that are freely available. Applications special for bioinformatics are continuously being developed to combat the constant growing computational and storage power.



(2a)



(2b)


NGS
USER
Short lenghts of reads
Assembled into new genome
Size of reference genome vs huge number of template sequences
Matched aganst reference genome
Combination of Short reads, errors and variations during alignment methods are not suitable
Algorithms
Limited PC Memory
Packing genome and sequences into limited amount of PC memory
Compression vs Comparatively (read errors)
Data storage and analysis
Mapping
CPU efficient?
Spaced seeds (number of hits)
Visualization
Max. 65% of genome assembled
Quality of matching? Difficulties by assembling a genome De novo from short reads generated by NGS. Data analysis algorithms

(2c)

Fig. 2. Methodologies use for Next Generation Sequencing and Metagenomics. 2a. Workflow of NGS; computational demands end bottlenecks. Blue square indicates process of data handled by powerful computational infrastructure and power. 2b. Workflow of Metagenomic pipeline, bottlenecks are between upload server and trimming data. 2c. Next generation sequence pipeline; software limitations

Working with one single cloud service provider can be restrictive while working with more than one cloud can make the workflow more robust in the face of failures and unanticipated needs [13,58-61]. One of the technologies to execute bioinformatics programs in cloud is the Apache Hadoop framework [62], in which the MapReduce [61,63] model and its distributed file system (HDFS) [64] are used as infrastructure to distribute large scale data processing and storage facilities. Since they are independent from one another parallelization of MapReduce does not require communication among simultaneously processed tasks. Still main issues are remaining; costs are comparatively higher to perform computations in the cloud [10]. Wireless sensor networks can be broadly applied in various areas such as medical care and environmental monitoring. Today, we can rapidly and affordably sequence a single human genome. We have sensors that can remotely track virtually any physiologic metric parameter, from vital signs to glucose to intraocular pressure. We can add a lab-on-a-chip [65] to a smartphone to assay almost any routine chemistry and digitize pills to ensure adherence. Or use a smartphone app to conduct all the components of do-it yourself (DIY) physical examinations [66]. This is superimposed and convergent with a remarkable digital infrastructure that includes ever-increasing bandwidth, pervasive connectivity, cloud- and supercomputing, enormous social networks, and those small mobile devices that we cannot put down. Unfortunately, a great deal of energy will be wasted if data which include time and space correlation is transmitted [25,67,68]. Data compression algorithm for those networks could be a solution [69-74].

3. PROCESSOR UNITS

As the sequence technology continues growths and improved tools become available on the market, sequencers are increasingly producing larger quantities of data. This big data make computational analysis with contemporary tools more challenging. Unfortunately, calculation speed has been frequently found insufficient; especially for analyzing large data obtained from

NGS [75]. Faster search tools, such as BLAST, have sufficient search sensitivity for Metagenomic analysis, but are only valuable if you know how to deal with the data. Here, the use of Bioinformatic tools is very important. Development of new highly efficient homology search algorithms suitable for central processing units (CPUs; serial processors) is warranted [76]. To get shorter processing times and executing virtually any algorithm or software, hardware acceleration applying a specialized hardware for a given problem instead of CPU could be a concept [77]. Specific accelerators for those applications have a custom architecture that fits the needs of a certain family of algorithms but are unfit for other general tasks. They use parallel architectures which allow them to exploit the parallelism available in the given application by performing independent operations simultaneously. Hardware accelerators such as graphics processing unit (GPUs) [76] are utilized in many scientific applications when the time consuming operations makes it impractical or even impossible to use ordinary CPUs. The use in Bioinformatics is not an exception; it includes many problems and algorithms which are computationally expensive due to the large amount of NGS data to be processed or the complex operations that are involved [32]. CPU programs have also several limitations, for example: when it is impossible to access the host during CPU execution. Calculations of results have to be stored to memory on a CPU, and size of memory on a CPU can be a limiting factor. Storage of results often failed because of the shortage of CPU memory and data analysis represents a serious bottleneck in the NGS platform [78,79].

4. CLINICAL ASPECTS

It has been often quoted that drugs are one of the problems for the emergence of resistant organisms. The problem of this antimicrobial resistance is expected to increase disproportionately and controlling infections is becoming difficult because of the rapid spread of those microorganisms. Recent infectious outbreaks, for example with *E. coli* in Germany, Multi Resistant *Staphylococcus aureus* (MRSA), and multi resistant fungi in the USA are responsible for an annual death rate yearly over 20.000 people in the US only. Those microorganisms have been modifying themselves to evade the action of available drugs or causing rapid spread before they can be prevented. Due to the increase in surgical interventions, transplantations, stem cell therapies, and oncological treatments those illicit immunocompromised side effects, is often accompanied by opportunistic infections. Fast and precise identification of those microorganisms in medical samples is necessary to apply the optimal medical treatment as soon as possible. The biomedical informatics provided a proper interdisciplinary context to integrate data and knowledge when processing available information, aiming effective decision making support in clinics [80]. From the patient’s side, usually a sample is taken and cultures are set up. This approach is very highly time consuming and error proning, since many medical microorganisms are difficult to detect for biopsy and hard to culture. The sample usually contains a mixture of human cell material, bacteria, parasites, fungi and often in a limited volume. Thus not too many different cultivation media can be used. The information for identification using morphology is slowly available and can be very limited. The consequences: no immediate individualized treatment possible, patients are infectious over a long time and suffer from the damage the microorganism produced during prolonged untreated infection. The complex nature of those patients condition have been difficult to assess the resultant increase in mortality, length of hospital stay (LOS), and costs attributable to the infection (US costs approximately $ 20 billion / yearly). Given that delay in appropriate antimicrobial therapy is associated with increased mortality, improved use of early empirical, pre-emptive, and prophylactic therapies should help to reduce invasive infectious associated mortality. Lack of specific clinical findings and slow insensitive diagnostic testing complicates the early recognition and treatment of Invasive infections and become a persistent public health problem. The incidence and mortality rates associated with infectious diseases have remained unchanged for more than a decade despite major advances in the field of antimicrobial therapy.

4.1 Next Generation Sequencing in Clinical Settings

Next generation sequencing platforms have accelerated clinical and research genomics because they provided an inexpensive and scalable way to interrogate genetic differences, gene expression, and other epigenetic and epitranscriptomic variations of DNA and RNA [36]. High-throughput method for proteomics and metabolomics are now being added as features for patients to examine during routine medical visits. Metagenomics and metabolomics are the newly emerging field of research in the high-throughput identification and quantification of the small molecule metabolites in the metabolome [81-83]. Defined as the complete collection of all small molecules (<1,500-Da) metabolites (peptides, amino acids, sugars, bases, lipids, etc.) are found in an organism, organ or specific cell [84]. Because of its unique focus on small molecules and interactions, metabolomics are widespread found in a variety of clinically important areas, including infectious diseases [85], clinical toxicity [86], diabetes [87], osteoarthritis [88], genetic disease diagnosis [89,90], trauma and surgical interventions and organ/stemcell transplant monitoring [91]. In the context of a modern laboratory for analysing infectious samples and biopts, NGS offers a solution compared to current alternatives [92]. Taxonomic characterizing of the microbial communities offers a new challenge into the Metagenomic studies. DNA barcoding and sequencing of specific loci are major steps to address this issue. Directed to the future, molecular epidemiology performed for outbreak and surveillance of infections for microbes that are difficult to grow can be a routine use of Whole Genome Sequencing (WGS) in the diagnostic and public health microbiology [93]. Implication of these diagnostic tools is still pending, owing to the time-consuming processing of the results. Adequate typing and diagnosing of the pathogens involved in infection disease will definitely result in dedicated and personalized antimicrobial therapies. The use of microbial analysis [94] is still under development, new data in this field of genomics, proteomics or transcriptomics [95] are readily available and easily analyzed through free online databases. Data can help to define gene models, correcting errors in genome annotation. Furthermore, metabolomics have strong emphasis on chemicals and analytical chemistry techniques such as mass spectrometry (MS), nuclear magnetic resonance spectroscopy (NMR), and chromatography. The software used by metabolomics, particularly as it relates to metabolite identification, is often different than the software used in genomics, proteomics, and or transcriptomics. The field of metabolomics is not only concerned with the identification and quantification of metabolites, but it is also concerned with relating metabolite data to genes, proteins, pathways, physiology, and phenotypes. As a result, metabolomics requires that whatever chemical information it generates must be linked to both biochemical causes and physiological consequences. This means that computational approaches to metabolomics must combine two completely different disciplines: bioinformatics and chemical informatics. Currently, the genetic research of NGS sequencing technologies is revolutionizing and revived methods in drug design. Furthermore, the last years have witnessed the emergence of different computational tools aimed at understanding and modeling this process at a molecular level [96-99]. Although still rudimentary, these methods are shaping a coherent approach to help in the design of molecules with high affinity and specificity, both in lead discovery and in lead optimization. For Cancer, integrating multiple genome wide data, increase predictive performance of clinical decision support models [100]. However, for identification of microbial species those experiments are expensive and time consuming. If we can provide biomarker genes that show unique expression patterns during infection, this approach opens new insights. Another aspect of genome annotation is gene prediction, identification of uncharacterized genomic sequences; important for understanding alterations and evolution in biological functions of the sequences of these genomes in human health and medicine [101-102]. But the current increase in amount of available data emphasizes the need for a methodological integration framework.

5. CONCLUSIONS AND FUTURE CONSIDERATIONS

First, fast connecting with other databases helps in healthcare discussion making as independent evidence [103]. Use of tools available by other databases; SAM-Tools (finding significant genes) [104], Genome expression (Omnibus DB) [105], Burrows-Wheeler (reads >220bp) [106], Phylogenic approaches (reliability, avoiding increase presence of polluting sequences) and ITS region pipelines [107] results in > 96% identification of species. Now NGS has fundamentally altered the genomic research, development costs will drop down and the technology will bring extreme potential for fast and accurate molecular bacterial typing for clinical microbiology. One problem that rises; global available databases use different methods of data access. While some databases allow data to be downloaded via web access, others provided flexible access to their databases only through their commercial ingredients (API). A lot of intervention by scientists is required to download the required information from databases with no public API and creates challenges for software developers to obtain information from such available databases. For databases whose API is public accessible, there's no guarantee that all such API would use the same programming languages. This causes developers to incorporate clumsy wrappers within their applications to adhere to the API of the databases. Software solutions from Galaxy combined with program MEGAN5 (Huson et al.) could be solving this problem. All databases release their pathway information via some non-standard graphical format. Such a graphical representation is useful for visual manual analysis. However, it is inconvenient for large-scale computational analysis and provided incompatible data formats. Databases such as GenBank and EMBL are mentioned as reference sources for publicly available sequence information of patients. Unfortunately, both databases only include granted patents, which limit the searches. Various patent offices do not provide freely available and comprehensive feeds of their sequence data or sequences are only available as TIFF or PDF image files which require the use of optical character recognition software for extracting the sequence data. Besides that, neither organization is clear in the way they get data from patent offices and represent their RNA sequences converted to DNA. In terms of sequence comparison, it represents a loss of information in the dataset. This can be an issue in addressing the IP landscape around a sequence, especially in the area of small RNA sequences.

5.1 Limitations

Secondly, epidemiology as study of genetic factors determining health and diseases at population level can be seen as science dealing with etiology, control and distribution of diseases in individuals with relationship between them and the population to a certain extent [93,103,108,109]. New tablet devices combined with high bandwidth network (4G), significantly enhance speed and improved access to rich data visualization delivered by location intelligence functionality. Epidemiological advantages such as infection spread are visible, and medically relevant microorganism differences between clinical significance and invasive species can be achieved [110-113] due new development of innovative techniques. The incidence and mortality rates associated with infectious diseases have remained unchanged for more than a decade despite major advances in the field of antimicrobial therapy. The excess length of hospital stay (LOS) and additional interventions associated with invasive infections carries with it significant extra hospital costs, to the extent that annual expenditures for those infections have been estimated at more than ~$1 billion in the United States alone (2012). Epidemiological studies have recently identified species that may vary geographically in frequency of isolation (Yapar 2014, Caggiano 2015). It is also apparent that no class of antimicrobial agent is immune to the development of resistance. While national and international surveillance is important to recognize trends in epidemiology it is, however, of utmost importance to gain knowledge about the local epidemiology as this information should guide the empiric therapy of patients.

The use of mobile devices continues to grow unabated, and is expecting to have a forecast by 2017 of more than $ 20 billion. Recent research indicates its potential worth in the patient support and education. Unfortunately, despite the explosion of Apps in other industries, medical industries have generally been slow to exploit the possibilities they represent. With the possibilities of mobile Apps, direct links to different databases can be archived. This opportunity helps to reduce the wide spread of antimicrobial multi-resistance of cultures. Furthermore, epidemiological visualization can be used for governmental discussions for treatment of populations / groups confronted with infection diseases and emerging outbreaks. New data provided from unexplored localities and hosts helps in accumulating knowledge about the microbial life form and discover patterns of resistance, which may later form the basis of further questions on the complex life cycle of enigmatic microbial species. Knowledge of areas of increased incidence may improve diagnostic or prevention measures in patients at risk for endemic diseases, including those receiving immunosuppressive medications or with new environmental exposures and may affect diagnostic or prevention measures for patients at risk. The use of databases opens a broad definition of questions about capture standards and protection of privacy while accessing invaluable information.

Most important, only a few updated databases are available for clinical medical bacteria or fungi. The BioloMICS database [48,114] which include Mycobank is an open and continuously updating database. User actions and samples are traceable which allow the database to grow. The setup of this system can as well be used for identification and rapid treatment of infections caused by other microorganisms such as bacterial and /or parasites. The use of databases opens a broad definition of questions about capture standards and protection of privacy while accessing invaluable information. With the direct link to the online database (Qpr-Apps, patient samples, photo's (fungi / yeasts), LIMS tools as DNA extracts, PCR sequencing data and Nanopore technologies (GridION, minION) [54] can be sent for analyzing. Results visible on tablet and /or mobile phone can then directly be used in the laboratory or clinic. The software can act as a gene structure prediction in higher organisms. Data used to plot species presence on MAPS, exploring what signal indicating the environmental preferences in the growing online databases. There's a need to establish current distributions in the face of changing environmental conditions. A large amount of undetected data, currently available fungal ITS sequences (1% of estimated 1.5 Million fungal species) and morphological characters overlapping between species; fungi / plants / invertebrates can be used for standardization and reliability. Size and restriction analyses by PCR; amplified ITS region DNA, can be used as rapid and reliable method to identify clinically significant yeasts including new or merging pathogenic species. To provide a better biological inference, microarray experiments could provide information. It's possible to superimpose microarray data to RNA-seq. by Meta comparison methods. Unfortunately such techniques run into issues if the data source used is not consistent (low level) or comprehensive. Incompatibility of the different data sources renders this option extremely challenging [115]. The existing IT infrastructure built for microarray data is not suitable for NGS. The problem between microarray data and NGS data analysis is mainly a platform problem; to simply use Roche 454-software to handle Illumina data or data from the SOLiD system (representing nucleotides) could de done only with commercial software. Limitations may exceed into; incompatible methods of data access; incompatible data formats; incompatible molecular representations; and incompatible pathway names. Incompatibilities between different databases makes cross- data accesses another option to investigate. Introducing a standard of communication between programs will also help future expansion by integrating more bioinformatics tools and will provide a development environment for open source projects. Additional genomic information, such as alternative splicing and expression data derived from EST, SAGE and microarray experiments, can be integrated into the system. Improvement of new chips, like Field programmable gate arrays (FPGAs), could decrease the time for processing. Those silicon chips are providing extremely fast results for certain operations up to 11 times faster than CPUs. Alignments from BLAST processed on a 4U chassis-sized FPGA cluster would have an equivalent computational power of ''over 2,000 dual-core processors". But there is still a bottleneck for large FPGA applications; memory interface affects the performance of the FPGA and high bandwidth is necessary for greater speedups [116].

Storage mirroring provides an important element of data protection, and databases should be regularly being backed up. While databases grow in size, new optimized techniques are required to constrain both speed and the processing and calculation time. Failure caused by human errors; mistakes by update, cleaning or development of critical tables is an issue that has to addressed. As a suggestion, these issues can be encountered by a WIKI-based database structure. Restore and recovery can take a long time and transactions made after the time of error can be lost [117]. To protect against data loss, many IT organizations are investing in secondary data centers with standby databases, which can be synchronized with the changes being made in the production environment. The traditional method of synchronization requires expensive, remotely mirrored storage solutions. The storage replicates every write performed on the production system to the standby system. This means that expensive, high-bandwidth networks are required between datacenters, incurring additional cost and limiting the distances that can exist between the datacenters.

To better understand pathways and genes related to diseases new insight in this big data is necessary. Enabling diverse backgrounds together in a network can help by making decisions more patient's specific, leading to improved treatment. Medical doctors and researchers should without special computational training and Bioinformatic background use statistical techniques. Data integration (Bayesian) and machine learning algorithms to understand this huge amount of data is often too specialized and is only understandable to those who are working in this field.

ACKNOWLEDGEMENTS

The authors would like to thank Michel de Vries for useful comments on the manuscript.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet. 2012;13(9):667-72.
   DOI: 10.1038/nrg3305
2. Sanmiguel P. Next-generation sequencing and potential applications in fungal genomics. Methods Mol Biol. 2011;722: 51-60.
   DOI: 10.1007/978-1-61779-040-9_4
3. Su Z, Ning B, Fang H, Hong H, Perkins R, Tong W, Shi L. Next-generation sequencing and its applications in molecular diagnostics. Expert Rev Mol Diagn. 2011;11(3):333-43.
   DOI: 10.1586/erm.11.3
4. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: Higher than you think! Genome Biol. 2011;12(8):125.
   DOI: 10.1186/gb-2011-12-8-125
5. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform; 2013.
   DOI: 10.1093/bib/bbs086
6. Angiuoli SV, White JR, Matalka M, White O, Fricke WF. Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. PLoSOne. 2011;6(10):e26624.
   DOI: 10.1371/journal.pone.0026624
   Epub 2011 Oct 19.
7. Dai L, Gao X, Guo Y, Xiao J, Zhang Z. Bioinformatics clouds for big data manipulation. Biol Direct. 2012;7:43. discussion 43.
   DOI: 10.1186/1745-6150-7-43
8. Desai N, Antonopoulos D, Gilbert JA, Glass EM, Meyer F. From genomics to metagenomics. Curr Opin Biotechnol. 2012;23(1):72-6.
   DOI: 10.1016/j.copbio.2011.12.017.
   Epub 2012 Jan 5.
9. Zhao G, Bu D, Liu C, Li J, Yang J, Liu Z, Zhao Y, Chen R. CloudLCA: Finding the lowest common ancestor in metagenome analysis using cloud computing. Protein Cell. 2012;3(2):148-52.
   DOI: 10.1007/s13238-012-2015-8.
   Epub 2012 Mar 17.
10. Wilkening J, et al. Using clouds for metagenomics: A case study. In Cluster Computing and Workshops, 2009. CLUSTER ’09. IEEE International Conference. 2009;1–6.
11. Nguyen T, Shi W, Ruden D. CloudAligner: A fast and full-featured MapReduce based tool for sequence mapping. BMC Res Notes. 2011;4:171.
    DOI: 10.1186/1756-0500-4-171
12. Fusaro VA, Patil P, Gafni E, Wall DP, Tonellato PJ. Biomedical cloud computing with Amazon Web Services. PLoS Comput Biol. 2011;7(8):e1002147.
    DOI: 10.1371/journal.pcbi.1002147.
    Epub 2011 Aug 25.
13. Bittman TJ. The evolution of the cloud computing market; 2008.
    Available:http://blogs.gartner.com/thomas_bittman/2008/11/03/ the-evolution-of-the-cloud-computing-market
14. Buyya R, Ranjan R, Calheiros RN. Intercloud: Utility-oriented federation of cloud computing environments for scaling of application services, proceedings of the towards a hybrid federated cloud platform to efficiently execute bioinformatics workflows. 10th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP 2010), Springer. 2010;21–23.
15. Celesti A, Tusa F, Villari M, Puliafito A. How to enhance cloud architectures to enable cross-federation, 3rd IEEE International Conference on Cloud Computing (IEEE Cloud 2010), Miami, Florida, USA. 2010;337–345.
16. Langille MGI, Eisen JA. Bio torrents: A file sharing service for scientific data. Plos ONE. 2010;5:4.
17. Zhang Z, Vladimir B. Bajic, Jun Yu, Kei-Hoi Cheung, Jeffrey P. Townsend. Data integration in bioinformatics: Current efforts and challenges, bioinformatics - trends and methodologies, mahmood A. Mahdavi (Ed.); 2011.
    ISBN: 978-953-307-282-1, InTech,

Available:http://www.intechopen.com/books/bioinformatics-trends-and-methodologies/data-integration-in-bioinformatics-current-efforts-and-challenges

18. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2013;41(Database issue):D36-42. DOI: 10.1093/nar/gks1195. Epub 2012 Nov 27.

19. Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bate Ks, Browne P, van den Broek A, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Gamble J, Garcia Die F, Harte N, Kulikova T, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Sobhany S, Stoehr P, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R. The EMBL nucleotide sequence database. Nucleic Acids Res. 2005;33(Database issue): D29–D33. Available: http://www.ebi.ac.uk/embl/

20. Miyazaki S, Sugawara H, Gojobori T, Tateno Y. DNA data bank of Japan (DDBJ) in XML. Nucleic Acids Res. 2003; 31(1):13–16. Available: http://www.ddbj.nig.ac.jp

21. Alkan C, Sajjadian S, Evan E. Eichler. Limitations of next-generation genome sequence assembly. Nat Methods. 2011; 8(1):61-65. DOI: 10.1038/nmeth.1527

22. Nicholas R, Anderson E, Sally Lee, J. Scott Brockenbrough, Mark E. Minie, Sherrilynne Fuller, James Brinkley, Peter Tarczy-Hornoch. Issues in biomedical research data management and analysis: Needs and barriers. J Am Med Inform Assoc. 2007;14(4):478–488.

23. Keegan KP, Trimble WL, Wilkening J, Wilke A, Harrison T, D'Souza M, Meyer F. A platform-independent method for detecting errors in metagenomic sequencing data: DRISEE. PLoS Comput Biol. 2012;8(6):e1002541. DOI: 10.1371/journal.pcbi.1002541. Epub 2012 Jun 7.

24. de Magalhães JP, Finch CE, Janssens G. Next-generation sequencing in aging research: Emerging applications, problems, pitfalls and possible solutions. Ageing Res Rev. 2010;9(3):315-23. DOI: 10.1016/j.arr.2009.10.006. Epub 2009 Nov 10.

25. Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, Pachter L. Identification and correction of systematic error in high-throughput sequence data. BMC Bioinformatics. 2011;12:451. DOI: 10.1186/1471-2105-12-451

26. Thessen AE, Patterson DJ. Data issues in the life sciences. Zookeys. 2011;150:15-51. DOI: 10.3897/zookeys.150.1766 Epub 2011 Nov 28.

27. Schmitt CP, Burchinal M. Data management practices for collaborative research. Frontiers in Psychiatry, Review Article. 2011;2. DOI: 10.3389/fpsyt.2011.00047

28. Anderson NR, Sally Lee E, Scott Brockenbrough J, Minie ME, Fuller S, Brinkley J, Tarczy-Hornoch P. Issues in biomedical research data management and analysis: Needs and barriers. J Am Med Inform Assoc. 2007;14(4):478–488.

29. Big data: The next frontier for innovation, competition, and productivity. Report June 2011. McKinsey Global Institute. Available: www.mckinsey.com/mgi

30. Shetty AC, Athri P, Mondal K, Horner VL, Meltz Steinberg K, Patel V, Caspary T, Cutler DJ, Zwick ME. SeqAnt: A web service to rapidly identify and annotate DNA sequence variations. BMC Bioinformatics. 2010;11:471.

31. Illumina. Illumina life sciences; 2012. Available: http: //www. illumina. com

32. Camerlengo T, Ozer HG, Onti-Srinivasan R, Yan P, Huang T, Parvin J, Huang K. From sequencer to supercomputer: An automatic pipeline for managing and processing next generation sequencing data. AMIA Summits Transl Sci Proc. 2012;2012:1-10. Epub 2012 Mar 19.

33. Middleton AM. HPCC systems: Introduction to HPCC (High-Performance Computer Cluster). LexisNexis, White Paper; 2011.

34. Niemenmaa M, Kallio A, Schumacher A, Klemelä P, Korpelainen E, Heljanko K. Hadoop-BAM: Directly manipulating next generation sequencing data in the cloud. Bioinformatics. 2012;28(6):876-7. DOI: 10.1093/bioinformatics/bts054. Epub 2012 Feb 2.

35. Zhang J, Chiodin R, Badr A, Zhangd G. The impact of next-generation sequencing on genomics. J Genet Genomics. 2011; 38(3):95–109.

36. Wooley JC, Ye Y. Metagenomics: Facts and artefacts, and computational challenges. J Comput Sci Technol. 2009; 25(1):1-81.
DOI: 10.1007/s11390-010-9306-4

37. Scholz MB, Lo CC, Chain PSG. Next generation sequencing and bioinformatics bottlenecks: The current state of Metagenomic data analysis. Current Opinion in Biotechnology. 2012;23:9-15.

38. Woollard PM, Mehta NAL, Vamathevan JJ, Van Horn S, Bonde BK, Dow DJ. The application of next-generation sequencing technology to drug discovery and development. Drug Discovery Today. 2011;16:11-12.

39. Hong HX, Zhang WQ, Shen J, et al. Critical role of bioinformatics in translating huge amounts of next-generation sequencing data into personalized medicine. Sci China Life Sci. 2013;56: 110–118.
DOI: 10.1007/s11427-013-4439-7

40. Thomas T, Gilbert J and Meyer F. Metagenomics – a guide from sampling to data analysis. Microbial Informatics and Experimentation. 2012;2:3.
DOI: 10.1186/2042-5783-2-3

41. Dreszer TR, Karolchik D, Zweig AS, et al. The UCSC genome browser database: extensions and updates 2011. Nucleic Acids Res. 2012; 40:D918-23.

42. Flicek P, Amode MR, Barrell D, et al. Ensemble 2011. Nucleic Acids Res 2011;39:D800-6.
PMC3013672.

43. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW. Genbank. Nucleic Acids Res. 2012; 40(D1):D48–D53.

44. Cochrane G, Akhtar R, Aldebert P. Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. Nucleic Acids Res. 2008;36(Database issue):D5–D12.

45. Kodama Y, Mashima J, Kaminuma E, Gojobori T, Ogasawara O, Takagi T, Okubo K, Nakamura Y. The DNA data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments. Nucleic Acids Res. 2012;40(D1):D38–D42.

46. Pagani I, Liolios K, Jansson J, Chen IA, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. The genomes OnLine Database (GOLD) v.4: Status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 2012;40(D1):D571–D579.

47. Crous PW, Gams W, Stalpers JA, Robert V, Stegehuis G. Mycobank: An online initiative to launch mycology into the 21st century. Studies in Mycology. 2004;50:19-22.
Available: http://www.mycobank.org/

48. Robert V, Szoke S, Jabas J, Vu D, Chouchen O, Blom E, Cardinali G. BioloMICS software, biological data management, identification, classification and statistics. The Open Applied Informatics Journal. 2011;5:87-98454.

49. 454 life sciences. Roche diagnostics corporation.
Available: http://www.454.com/

50. ABI / SOLID system, Life Technologies.
Available: www.appliedbiosystems.com

51. Ion Torrent proton. Sequencing; 2012.
Available: http://iontorrent.com

52. Helicos tSMS Sequencing.
Available: http://www.helicosbio.com

53. Available:http://www.pacificbiosciences.com/

54. Oxford Nanopore MinION.
Available: http://www.nanoporetech.com/. Oxford Nanopore Technologies 2012 (GridION, MinION) PhysOrg.com.

55. Amazon. Amazon Elastic Compute Cloud (Amazon EC2); 2012.
Available: http://aws.amazon.com/ec2/

56. Redkar T. Windows azure platform, 2th edn, Apress, Berkeley, CA, USA. 2011;1.

57. Vaquero LM, Rodero-Merino L, Caceres J, Lindner M. A break in the clouds: Towards a cloud definition, SIGCOMM Comput. Commun. Rev. 2008;39:50–55.

58. Thakur RS, Bandopadhyay R, Chaudhary B, Chatterjee S. Now and next-generation sequencing techniques: Future of sequence analysis using cloud computing. Front Genet. 2012;3:280.

59. Mu-Hsing Kuo A. Opportunities and challenges of cloud computing to improve health care services. J Med Internet Res. 2011;13(3):e67.

60. Vilaplana J, Solsona F, Filgueira AR, Rius J. The cloud paradigm applied to e-Health. BMC Med Inform Decis Mak. 2013;13:35.

61. Qiu J, Ekanayake J, Gunarathne T, Choi JY, Bae SH, Li H, Zhang B, Wu TL, Yang Ruan Y, Ekanayake S, Hughes A, Fox G. Hybrid cloud and cluster computing

paradigms for life science applications. BMC Bioinformatics. 2010;11(Suppl 12): S3.

62. Apache. Apache Hadoop; 2012. Available: http://hadoop.apache.org/

63. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters, 6th Conference on Symposium on Operating Systems Design & Implementation, USENIX Association, Berkeley, CA, EUA. 2004;10–10.

64. Borthakur D. The Apache Software Foundation: HDFS Architecture; 2008. Available:http://hadoop.apache.org/common/docs/r0.20.2/hdfs_design.pdf

65. Genalysis®, Lab-free DNA Testing. Available:http://dnae.co.uk/platforms/genalysis/

66. Lillehoj PB, Huang MC, Truong N, Ho CM. Rapid electrochemical detection on a mobile phone. Lab Chip; 2013. Accepted Manuscript.
DOI: 10.1039/C3LC50306B

67. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IA, Grechkin Y, Dubchak I, Anderson I, Lykidis A, s Mavromatis K, Hugenholtz P, Kyrpides NC. IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res. 2008;36(Database issue):D534–D538.

68. Hawkins RD, Gary C, Hon GC, Ren B. next-generation genomics: an integrative approach. Nat Rev Genet. 2010;11(7): 476–486.
DOI: 10.1038/nrg2795

69. Popitsch N, von Haeseler A. NGC: lossless and lossy compression of aligned high-throughput sequencing data. Nucleic Acids Res. 2013;41(1):e27.
DOI: 10.1093/nar/gks939

70. Jiang P, Li SQ. a data compression algorithm for wireless sensor networks based on an optimal order estimation model and distributed coding. Sensors (Basel). 2010;10(10):9065–9083.
DOI: 10.3390/s101009065.

71. Nalbanto glu ÖU, Russell DJ, Sayood K. data compression concepts and algorithms and their applications to bioinformatics. Entropy (Basel). 2010; 12(1):34.
DOI: 10.3390/e12010034

72. Jones DC, Ruzzo WL, Peng X, Katze MG. Compression of next-generation sequencing reads aided by highly efficient de novo assembly. Nucleic Acids Res. 2012;40(22):e171.
DOI: 10.1093/nar/gks754

73. Morihiro Hayashida, Tatsuya Akutsu. Comparing biological networks via graph compression. BMC Syst Biol. 2010; 4(Suppl 2):S13.
DOI: 10.1186/1752-0509-4-S2-S13

74. Bonfield JK, Mahoney MV. Compression of FASTQ and SAM format sequencing data. PLoSOne. 2013;8(3):e59190.

75. Samuel Lampa, Martin Dahlö, Pall I Olason, Jonas Hagberg, Ola Spjuth. Lessons learned from implementing a national infrastructure in Sweden for storage and analysis of next-generation sequencing data. Gigascience. 2013;2:9.

76. Suzuki S, Ishida T, Kurokawa K, and Akiyama Y. GHOSTM: A GPU-accelerated homology search tool for metagenomics. *PLoSOne*. 2012;7(5):e36060.
DOI: 10.1371/journal.pone.0036060

77. Fan K, Kudlur M, Dasika GS, Mahlke SA. Bridging the computation gap between programmable processors and hardwired accelerators. HPCA. 2009;313-322.

78. Ghoting A, Buehrer G, Parthasarathy. A characterization of data mining algorithms on a modern processor. Proceedings of the First International Workshop on Data Management on New Hardware. DaMon; 2005.

79. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147:195-197.

80. Gullapalli RR, Lyons-Weiler M, Petrosko P, Dhir R, Becich MJ, LaFramboise WA. Clinical integration of next-generation sequencing technology. Clin Lab Med. 2012;32(4):585-99.
DOI: 10.1016/j.cll.2012.07.005

81. Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J. Metagenomic Pyrosequencing and Microbial Identification. Clinical Chemistry 2009;55(5):856–866.

82. German JB, Hammock BD, Watkins SM. Metabolomics: Building on a century of biochemistry to guide human health. Metabolomics. 2005;1:3–9.

83. Godzik A. Metagenomics and the protein universe. Current Opinion in Structural Biology. 2011;21:398-403.

84. Wishart DS. Human metabolome database: Completing the “human parts list”. Pharmacogenomics. 2007;8:683–686.

85. Coen M, O'Sullivan M, Bubb WA, Kuchel PW, Sorrell T. Proton nuclear magnetic resonance-based metabolomics for rapid diagnosis of meningitis and ventriculitis. Clin Infect Dis. 2005;41:1582–1590.

86. Griffin JL, Bollard ME. Metabolomics: It's potential as a tool in toxicology for safety assessment and data integration. Curr Drug Metab. 2004;5:389–398.

87. Yang J, Xu G, Hong Q, Liebich HM, Lutz K, Schmulling RM, Wahl HG. Discrimination of Type 2 diabetic from healthy controls by using metabolomics method based on their serum fatty acid profiles. J Chromatogr B. 2004;813:53–58.

88. Williamson MP, Humm G, Crisp AJ. ^{1}H nuclear magnetic resonance investigation of synovial fluid components in osteoarthritis, rheumatoid arthritis and traumatic effusions. Br J Rheumatol. 1989;28:23–27.

89. Wishart DS, Querengesser LMM, Lefeb¬vre BA, Epstein NA, Greiner R, Newton JB. Magnetic resonance diagnostics: A new technology for high-throughput clinical diagnostics. Clin Chemistry. 2001;47:1918–1921.

90. Moolenaar SH, Engelke UF, Wevers RA. Proton nuclear magnetic resonance spectroscopy of body fluids in the field of inborn errors of metabolism. Ann Clin Biochem. 2003;40:16–24.

91. Wishart DS. Metabolomics: The principles and potential applications to transplantation. Am J Transplant. 2005;5:2814–2820.

92. Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. J. Pathol Inform. 2012;3:40.

93. Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J, Peacock SJ. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog. 2012;8:8.

94. Wishart DS. Current progress in computational metabolomics. Brief Bioinform. 2007;8:279–293.

95. Mutz K, Heilkenbrinker A, Lönne M, Walter J, Stahl F. Transcriptome analysis using next-generation sequencing. Current Opinion in Biotechnology. 2013; 24:22–30.

96. Jiang B, Bussey H, and Roemer T. Novel strategies in antifungal lead discovery. Current Opinion in Microbiology. 2002;5: 466-471.

97. Haas J, Katus HA, Meder B. Next-generation sequencing entering the clinical arena. Molecular and Cellular Probes. 2011;25:206-211.

98. Brown JR. Next generation sequencing for antibacterial drug discovery. Int. Drug Discov. 2010;5:18-23.

99. Boyd. Diagnostic applications of high-throughput DNA sequencing. Annual review of pathology: Mechanisms of Disease. 2013;8:381-410.

100. Daemen A, Gevaert O, Ojeda F, Debucquoy A, Suykens JA, Sempoux C, Machiels JP, Haustermans K, De Moor B. A kernel-based integration of genome-wide data for clinical decision support. Genome Med. 2009;1(4):39.

101. Goel N, Singh S, Aseri TC. A comparative analysis of soft computing techniques for gene prediction. In Press: Anal. Biochem; 2013;
Available:http://dx.doi.org/10.1016/j.ab.2013.03.015

102. Caccia D, Dugo M, Callari M, and Bongarzone I. Bioinformatics tools for secretome analysis. In Press. Biochimica et Biophysica Acta; 2013.
Available:http://dx.doi.org/10.1016/j.bbapap.2013.01.039

103. Carriço JA, Sabat AJ, Friedrich AW, Ramirez M. Bioinformatics in bacterial molecular epidemiology and public health: Databases, tools and the next-generation sequencing revolution. Euro Surveill. 2013;18(4):20382.

104. Laczik M, Tukacs E, Uzonyi B, Domokos B, Doma Z, Kiss M, Horváth A, Batta Z, Maros-Szabó Z, Török Z. Geno viewer, a SAM/BAM viewer tool. Bioinformation. 2012;8(2):107–109.

105. Galperin MY, Cochrane GR. The 2011 nucleic acids research database issue and the online molecular biology database collection. Nucleic Acids Res. 2011;39: D1–D6

106. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14): 1754-60.

107. Nilsson RH, Bok G, Ryberg M, Kristiansson E, Hallenberg N. A software pipeline for processing and identification

of fungal ITS sequences. Source Code Biol Med. 2009;4:1.

108. Bellazi R, Diomidous M, Sarka IN, Takabayashi K, Ziegler A, McCray AT. Data analysis and data mining: Current issues in biomedical informatics. Methods Inf Med. 2011;6:536-544.

109. Haux R. Medical informatics: past, present, future. Int J Med Inform. 2010;79(9):599-610.

110. Tse-Laurence MA, Bidartondo MI. Mapping fungi from belowe ground: Online genetic resources and ectomycorrhizal geopraphic distributions. iForest. Biosciences and Forestry. 2011;4:252-255.

111. Ali M, Park JK, von Siedlein L, Acosta CJ, Deen JI, Clemens JD. Organizational aspects and implementation of data systems in large-scale epidemiological studies in less developed countries. BMC Public Health. 2006;6:86.

112. Mukhi S, Stuart Chester TL, Klaver-Kibria JDA, Nowicki DL, Whitlock ML, Mahmud SM, Louie M, Lee BE. Innovative technology for web-based data management during an outbreak. Online J Public Health Inform. 2011;3:1.

113. Lacson R, Pitzer E, Kim J, Galante P, Hinske C, Ohno-Machado L. DSGeo: software tools for cross-platform analysis of gene expression data in GEO. J Biomed Inform. 2010;43(5):709-15.

114. Vu TD, Eberhardt U, Szoke S, Groenewald M, and Robert V. A laboratory information management system for DNA barcoding workflows. Integr. Biol. 2012; 4:744-755.

115. Wang J, Tan A, Tian T. Next Generation Microarray Bioinformatics: Methods and Protocols. Methods in Molecular Biology. 2012;802. DOI:10.1007/978-1-61779-400-1_1, Springer Science+Business Media, LLC 2012.

116. Sirowy S, Forin A. Where's the beef? Why FPGAs are so fast. Microsoft Research, Technical Report, MSR-TR-2008-130; 2008.

117. Townsend M, Hardie W, Weiss R, Shakian A, Ray A, Jernigan K, Vengurlekar N, Avril P, Baier M, Hamburger M. Optimizing storage and protecting data with oracle database 11g. Oracle White Paper; 2011.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/14483