Journal of
Data Mining in Genomics & Proteomics

Al-Haggar et al., J Data Mining Genomics Proteomics 2013, 4:3
DOI: 10.4172/2153-0602.1000131

Review Article Open Access

Bioinformatics in High Throughput Sequencing: Application in Evolving Genetic Diseases

Mohammad MS Al-Haggar[1]*, Balkis A Khair-Allaha[2], Mohammad M Islam[3] and Abdalla SA Mohamed[3]

[1]Pediatrics Department, Genetics Unit, Faculty of Medicine, Mansoura University, Egypt
[2]Biomedical Engineering Department, MUCH (Mansoura University Children's Hospital), Mansoura, Egypt
[3]Systems and Biomedical Engineering Department, Faculty of Engineering, Cairo University, Giza, Egypt

Abstract

Bioinformatics is a computational biology, in terms of macromolecules applying "informatics" techniques to understand and organize the information associated with these molecules. These data are product of large-scale molecular biology projects, such as the various genomes sequencing projects, analysis of gene expression and analysis of genomics, proteomics and protein-protein interactions.

They are collected and stored in different databases. Analysis in bioinformatics available in molecular biology focuses on: macromolecular structures, genome sequences and gene expression data. Techniques developed by computer scientists have enabled researchers to sequence nearly 3 billion base pairs of the human genome. Recent scientific discoveries that resulted from the application of next generation DNA sequencing technologies have given rise to the science of genomics, and have enabled critical advances in other fields, including epidemiology, forensics, evolutionary biology and medical diagnostics. Technologies for high throughput sequencing, their limitations and their applications are spotted in this review. Sequencing known genes enables the discovery of novel mutations that could help scientists understanding the evolving features of some genetic diseases, occurrence of many genetic diseases due to mutant variants of one gene or clusters of genes, or even explains the overlapping features of some genetic diseases mapped to nearby or distant loci.

Keywords: Bioinformatics; High throughput sequencing; Re-sequencing, Techniques; Platforms

Introduction

Computational, mathematical, statistical and informatics technologies developed parallel to the biological research enabled scientists to interconnect, integrate and interpret the complex nature of any biological system. In fact, information extraction from complex data is a great problem in biological research, where computational systems, biostatistics and information technologies are finding their increasing applications. The assemblage and integration of all these technologies in solving the problems related to the biological systems have been termed as "bioinformatics" in mid 1980s [1].

According to this, Bioinformatics has become an integral part of research and development in the biomedical sciences, and also has an essential role both in deciphering genomic, transcriptomic and proteomic data generated by high throughput experimental technologies, and in organizing information gathered from traditional biology [2].

Defining Bioinformatics as a union of biology and informatics, meaning bioinformatics, involves the technology that uses computers for storage, retrieval, manipulation and distribution of information related to biological macromolecules, such as DNA, RNA and proteins. Bioinformatics has a major impact on many areas of biotechnology and biomedical sciences and applications, for example, in knowledge-based drug design, forensic DNA analysis and agricultural biotechnology.

Bioinformatics is limited to sequence, structural and functional analysis of genes and genomes and their corresponding products. It is often considered as computational molecular biology [3]. The ultimate goal of bioinformatics is to better understand a living cell, and how it functions at the molecular level. By analyzing raw molecular sequence and structural data, bioinformatics research can generate new insights and provide a "global" perspective of the cell. The reason that the functions of a cell can be better understood by analyzing sequence data is ultimately because the flow of genetic information is dictated by the "central dogma" of biology, in which DNA is transcribed to RNA, which is translated to proteins. Cellular functions are mainly performed by proteins whose capabilities are ultimately determined by their sequences. Therefore, solving functional problems using sequence and sometimes structural approaches has proved to be a fruitful endeavor [3].

Sequence based methods of analyzing individual genes or proteins have been elaborated and expanded, and developed for analyzing large numbers of genes or proteins simultaneously. With the complete genome sequences for an increasing number of organisms at hand, bioinformatics is beginning to provide both conceptual bases and practical methods for detecting systemic functional behaviors of the cell and the organisms [2].

The completion of the first human genome drafts was just a start of the modern DNA sequencing era, which resulted in further invention, improved development toward new advanced strategies of high-throughput DNA sequencing, so called the "high-throughput-next generation sequencing" (HT-NGS). These developed HT-NGS strategies addressed our anticipated future needs of throughput

***Corresponding author:** Mohammad MS Al-Haggar, Pediatrics Department, Genetics Unit, Faculty of Medicine, Mansoura University, Egypt, Tel: +20502310661; E-mail: m.alhaggar@yahoo.co.uk

Received April 18, 2013; **Accepted** June 15, 2013; **Published** June 22, 2013

Citation: Al-Haggar MMS, Khair-Allaha BA, Islam MM, Mohamed ASA (2013) Bioinformatics in High Throughput Sequencing: Application in Evolving Genetic Diseases. J Data Mining Genomics Proteomics 4: 131. doi:10.4172/2153-0602.1000131

sequencing and cost, in a way which enabled its potential multitude of current and future applications in mammalian genomic research [4]. Additionally in these advanced laboratory methodologies, a scope of new generation of bioinformatics tools has further emerged as an essential prerequisite to accommodate further strategic development and improvement of output results.

Sequencing Overview

Sequencing has progressed far beyond the analysis of DNA sequences, and is now routinely used to analyze other biological components, such as RNA and protein, as well as how they interact in complex networks. In addition, increasing throughput and decreasing costs are making medical applications of sequencing a reality.

Next-generation sequencing (also 'Next-gen sequencing' or NGS) refers to DNA sequencing methods that came to existence in the last decade after earlier capillary sequencing methods that relied upon 'Sanger sequencing' [5]. As opposed to the Sanger method of chain-termination sequencing, NGS methods are highly parallelized processes that enable the sequencing of thousands to millions of molecules at once.

Popular NGS methods include pyrosequencing developed by 454 Life Sciences (now Roche), which makes use of luciferase to read out signals as individual nucleotides are added to DNA templates, Illumina sequencing that uses reversible dye terminator techniques that adds a single nucleotide to the DNA template in each cycle and SOLiD sequencing by Life Technologies that sequences by preferential ligation of fixed length oligonucleotides [4]. But these advances did not merely make the sequencing of DNA and RNA cheaper and more efficient; they have also helped create innovative new experimental approaches that penetrate deeply into the molecular mechanisms of genome organization and cellular function.

A prime example of the advances that have been facilitated by new sequencing technologies is the NHGRI-funded ENCODE project, which was launched in late 2003, based largely upon methods first developed in yeast [6,7]. The pilot phase of ENCODE relied heavily on microarray-based assays to analyze 1% of the human genome in unprecedented depth [8].

With credit to advances in high-throughput sequencing, researchers expanded the scope of this project, to include the whole human genome [9]. A total of ~1650 high-throughput experiments were performed to analyze transcriptomes and map elements, and identify methylation patterns in the human genome. This multi-institution consortia project has assigned biochemical activities to 80% of the genome, particularly annotating the portion of the genome that lies outside the well-studied protein-coding regions, including mapping over four million regulatory regions. This information has also enabled researchers to map genetic variants to gene regulatory regions and assess indirect links to disease [10]. Similar projects annotating the genome have also been performed for *Drosophila melanogaster* [1], *Caenorhabditis elegans* [11], and mouse [12].

Whole Genome Re-sequencing

The term "re-sequencing" refers to the act of sequencing multiple individuals from the same species, where a reference genome has been generated, and is used to assist in the interpretation of the data collected using next generation sequencing approaches. For example, re-sequencing of human genomes has been used to discover both mutations [13,14], and polymorphisms [1]. The existence of reference genome sequences has driven this application, which was the first one employed using Roche/454, Illumina/Genome-Analyzer and Applied Biosystems/SOLiD technologies.

Since that landmark study, whole genome re-sequencing continues to be used actively in various projects, including for example the 1000 Genomes Project [1], which aims to discover common sequence variants in healthy human populations, and also in various cancer studies (e.g. [13,14]), including those conducted under the auspices of the large TCGA (http://tcga.cancer.gov/) and ICGC (http://www.icgc.org/, 2010) consortia.

Applications for whole genome re-sequencing continue to emerge, and the steady decrease in cost per base and the increased throughputs associated with the latest technology advances will hopefully make this mode of data collection as appealing financially as it is scientifically.

DNA Sequencing Using Bioinformatics Analysis

Bioinformatics analysis of sequencing data can be divided into several stages. The first step is technology dependent, and deals with processing the data provided by the sequencing instrument. Downstream analysis is then done ad hoc to the type of experiment. When sequencing new genomes, de novo assemblies are required, which are possibly followed up with genome annotations. Re-sequencing projects use the short reads for aligning (or mapping assembly) against a reference sequence of the source organism; these alignments are then analyzed to detect events relevant to the experiment being conducted (e.g. mutation discovery, detection of structural variants, copy number analysis). The first step of bioinformatics analysis starts during sequencing, and involves signal analysis to transform the sequencing instruments fluorescent measurements into a sequence of characters representing the nucleotide bases. As sequencers image surfaces densely packed with the DNA sequencing templates and sequencing products, image processing techniques are required for detection of the nascent sequences and conversion of this detected signal into nucleotide bases. Most technologies assign a base quality to each of the nucleotides, which is usually a value representing the confidence of the called bases. Although each vendor has methods specific to their technology to evaluate base quality, most provide the user with a Phred-like Score value: a quality measurement based on a logarithmic scale encoding the probability of error in the corresponding base call [15].

To achieve contiguous stretches of overlapping sequence (contigs) in de novo sequencing projects, software that can detect sequence overlaps among large numbers of relatively short sequence reads is required. The process of correctly ordering the sequence reads, called assembly,is complicated by the short read length; the presence of sequencing errors; repeat structures that may reside within the genome; and the sheer volume of data that must be manipulated to detect the sequence overlaps. To address such complications, hybrid methods involving complimentary technologies have been successful. For example, by mixing 200 bp 454 sequences reads with Sanger sequences, Goldberg et al. [16] successfully sequenced the genomes of several marine organisms. A different approach eliminated the need for Sanger sequencing by mixing two distinct next generation sequencing technologies [17]. By taking advantage of 454's longer reads (250 bp) with short Illumina reads (36 bp), Reinhardt et al. [17] were able to de novo sequence a 6.5 Mb bacterial genome. These studies provided practical examples of how the strengths of different technologies can be used to alleviate their respective short comings.

Homology with previously sequenced organisms can help when sequencing new genomes. The use of this strategy was demonstrated

during sequencing of the mouse genome [18]; by taking advantage of the conserved regions between mouse and human, Gregory et al. [18] were able to build a physical map of mouse clones, establishing a framework for further sequencing. A similar approach can be used to produce better assemblies with next generation sequencing. For example, to sequence the genome of the fungus *Sordaria macrospora* [19], short reads from 454 and Illumina instruments were first assembled using Velvet [20], and the resulting contigs were then compared to draft sequences of related fungi (*Neurospora crassa*, *N. discreta* and *N. tetrasperma*).

This process helped produce a better assembly by reducing the number of contigs from 5,097 to 4,629, while increasing the N50 (the contig length N, for which 50% of the genome is contained in contigs of length N or larger), from 117 kb to 498 kb.

More recently, new algorithms have been developed, which can assemble genomes using only short reads. Most of these methods are based on de Bruijn graphs. Briefly, the logic involves decomposing short reads into shorter fragments of length k (k-mers). The graph is built by creating a node for each k-mer and drawing a link, or "edge," between two nodes when they overlap by k-1 bp. These edges specify a graph in which overlapping sequences are linked. Sequence features can increase the resulting graph's complexity. The graph can, for example, contain loops due to highly similar sequences (e.g. gene family members or repetitive regions), and so-called bubbles can be created when single base differences (e.g. due to polymorphisms or sequencing errors) result in the creation of non unique edges in the graph, which yield not one, but two possible paths around the sites of the sequence differences.

Graph complexity and size increase for large genomes, and given that the graph needs to be available in memory for efficient analysis, not all implementations can handle human size genomes. Some publicly available implementations, such as Velvet [20] and Euler-SR [21], have been successfully used to assemble bacterial genomes. Another implementation, ABySS [22], makes use of parallel computing through the Message Passing Interface (MPI), to distribute the graph between many nodes in a computing cluster. In this way, ABySS can efficiently scale up for the assembly of human size genomes, using a collection of inexpensive computers. Two newer assemblers [23], and ALLPATHS-LG [24], are able to assemble human-sized genomes using large memory multi-cpu servers, requiring 150 Gb and 512 Gb RAM, respectively.

For re-sequencing experiments, high-throughput aligners are required to map reads to the reference genome. Many applications have long been available for sequence alignments; however, the amount and size of the short reads created by next generation sequencing technologies required the development of more efficient algorithms. Some methods use "hashing" approaches, such is the case of Maq [23], in which the reads are reduced in complexity to unique identifier keys ("hashed"). These can then be used to scan a table made from a similarly "hashed" representation of the reference genome to identify putative read alignments to the reference. Other methods, based on Burrows-Wheeler transformation, have become popular for read alignment. These include BWA [25], Bowtie [26], and Soap [26]. Although these algorithms are relatively fast compared to Maq [27], they are somewhat limited when it comes to splitting a read to achieve gapped alignments, which can occasionally be required due to insertion/deletion sequence differences ("indels") between sequence data and the reference. The Mosaik aligner [28] attempts to approach this by using a Smith and Waterman (1981) algorithm to align the short reads.

Genomic Sequencing in Medical Fields

Genomic sequencing will have an enormous impact on the field of medicine. Until recently, cost and throughput limitations have made general clinical applications infeasible. Currently, though, the price of about 5000 USD for a normal human genome sequence (not counting analysis), and fast throughput (several days to a few weeks) is rapidly making the medical sequencing practical. Indeed, high-throughput sequencing has already been used to help diagnose highly genetically heterogeneous disorders, such as X-linked intellectual disability, congenital disorders of glycosylation and congenital muscular dystrophies [29]; to detect carrier status for rare genetic disorders [29,30]; and to provide less invasive detection of fetal aneuploidy through the sequencing of free fetal DNA [31]. Nonetheless, medical sequencing could potentially be applied in a wide range of areas, such as cancer, hard-to-diagnose diseases and personalized medicine.

From Low Throughput to High Throughput Sequencing Bioinformatics in Clinical Settings

Many genetic diseases are characterized by cutting-edge features; overlapping features of some genetic diseases may necessitate an extensive study, not only at the DNA, but also at the protein level. Low throughput sequencing known target genes enables the discovery of novel mutations that could help scientists understanding the evolving features of some genetic diseases, occurrence of many genetic diseases due to mutation variants of one gene or cluster of genes, or even the overlapping features of different genetic diseases mapped to nearby or distant loci.

Amplification of all the coding sequences, including flanking introns in *CTNS* gene using a Big Dye Primer Cycle Sequencing kit and an ABI 310 Genetic Analyzer (PE Applied Biosystems, Foster City, California, USA), yielded a novel nonsense mutation (c.734G4A); homozygous in probands, but heterozygous in the parents [32]. This mutation substitutes tryptophan by a premature stop codon at the position 245 in cystinosin (W245X). This novel truncating *CTNS* mutation could explain the detection of congenital heart defects, for the first time–not previously reported in literature, in the two patients with severe infantile cystinosis (Figure 1A and 1B).

On the other hand, as *LMNA* gene (OMIM: 150330) mutations that codes for lamin A/C (HGNC id: 663) had been associated with more than 13 disease variants, involving heart, nerve, adipose tissue, skeleton…etc. in different patterns of which mandibulo-acral dysplasia (OMIM: 248370) and Hutchinson-Gilford progeria syndrome (OMIM: 176670).A novel p.Arg527Leu *LMNA* mutation in two unrelated Egyptian families causes overlapping mandibuloacral dysplasia and progeria syndrome had been recently discovered; the affected patients had features of mandibulo-acral dysplasia (stunted growth, hypoplastic mandible, stiff spine, acro-osteolysis of distal phalanges), with some progeroid features, such as pinched nose, premature loss of teeth, loss of hair and scleroderma-like skin atrophy [33]. Patients were homozygous; however, their parents were heterozygous for p.Arg527Leu *LMNA* mutation (Figure 2A and 2B). Computational predictions of such substitution effects suggested an alteration in the protein stability, and thus a great tendency for protein aggregation; such changes might influence its interaction with other proteins. This bioinformatics prediction has been recently proven by the detection of minor ultra-structural changes in heterozygous parents, compared to the severe changes in affected patients, as elucidated on electron microscopic examination of skin biopsy samples (Al-Haggar M, personal communication, accepted for publication in Journal of clinical pathology, May 19, 2013).

```
EXON 8  100 bp
154   T GTC ATT GGT CTG AGC TTC GAC TTC GTG GCT CTG AAC CTG ACG 168
169 GGC TTC GTG GCC TAC AGT GTA TTC AAC ATC GGC CTC CTC TGG GTG 183
184 CCC TAC ATC AAG 187

EXON 9  120 bp
188 GAG CAG TTT CTC CTC AAA TAC CCC AAC CGA GTG AAC CCC GTG AAC 202
203 AGC AAC GAC GTC TTC TTC AGC CTG CAC GCG GTT GTC CTG ACG CTG 217
218 ATC ATC ATC GTG CAG TGC TGC CTG TAT GAG 227

EXON 10  171 bp        Codon 245 TGG → TAG (Stop codon)*
228 CGC GGT GGC CAG CGC GTG TCC TGG CCT GCC ATC GGC TTC CTG GTG 242
243 CTC GCG TGG CTC TTC GCA TTT GTC ACC ATG ATC GTG GCT GCA GTG 257
258 GGA GTG ACC ACG TGG CTG CAG TTT CTC TTC TGC TTC TCC TAG ATC 272
273 AAG CTC GCA GTC ACG CTG GTC AAG TAT TTT CGA CAG 284

EXON 11  118 bp
285 GCC TAG ATG AAC TTT TAC TAC AAA AGC ACT GAG GGC TGG AGC ATT 299
300 GGC AAC GTG CTC CTG GAC TTC ACC GGG GGC AGC TTC AGC CTC CTG 314
315 CAG ATG TTC CTC CAG TCC TAC AAC AAC G  324

EXON 12  1303 bp
324  AC CAG TGG ACG CTG ATC TTC GGA GAC CCA ACC AAG TTT GGA CTC 338
339 GGG GTC TTC TCC ATC GTC TTC GAC GTC GTC TTC TTC ATC CAG CAC 353
354 TTC TGT TTG TAC AGA AAG AGA CCG GGG TAT GAC CAG CTG AAC TAG 367
```

Figure 1A: The last five exons of *CTNS* gene showing G>A substitution in codon 245 (exon 10), at base position 734 of *CTNS* gene coding region (c.734 G>A), causing substitution of (TGG) for tryptophan to (TAG) a stop codon, at position 245 of cystinosin protein (p.W245X), leading to loss of 122 amino acid residues.

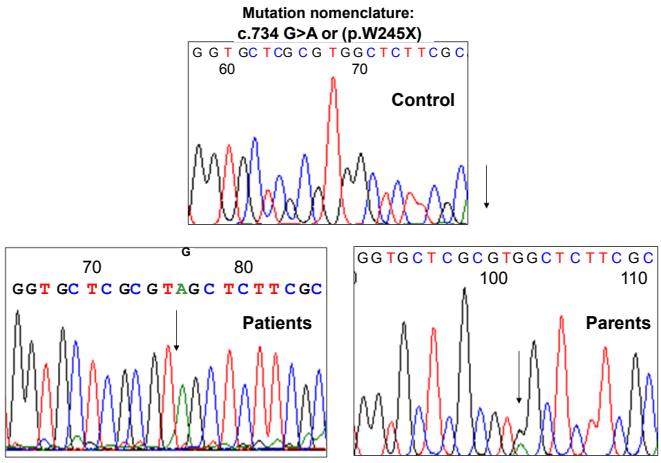

Figure 1B: Sequencing results of *CTNS* gene; note homozygous mutation in the patients (A base) and heterozygous mutation in the consanguineous parents (G-A bases overlap), compared to wild sequence in control (G base).

More extensively, whole genome sequencing is sometimes mandatory for elaboration of 'mysterious' clinical diseases, i.e. if dissection of known gene(s) candidates for a clinical state yielded no positive results. In other words, whole-genome and exome sequencing is likely to prove useful in the diagnosis of rare diseases, and in selecting the optimal individualized treatment option for patients. This approach typically involves the use of families; sequencing of affected individuals and relatives along with inheritance patterns is used to deduce variants that are associated with a disease. Whole exome sequencing performed on a four member family led to the discovery of the causative gene for Miller's syndrome, an extremely rare condition that gives rise to micrognathia and cleft lips among other features [34]. Nicholas Volker received a bone marrow transplant after his genome sequence indicated he had a mutation on the X chromosome that led to an inherited immune disorder that was giving him multiple problems. With the new diagnosis at hand, Volker was successfully treated, and his severe inflammatory bowel disease alleviated [35]. Richard Gibbs describes using complete genome sequences of twins diagnosed with dopa-responsive dystonia to identify the appropriate treatment option, which eventually resulted in significant clinical improvements of the twins [36].

Conclusion

Bioinformatics mainly deals with four facets of analysis: DNA sequence analysis, protein structure prediction, functional genomics and proteomics, and systems biology. High-throughput sequencing, with its rapidly decreasing costs and increasing applications, is replacing many other research technologies. Nonetheless, significant challenges remain with NGS; these include data processing and

```
EXON 8  108 bp
461 GAC CAG TCC ATG GGC AAT TGG CAG ATC AAG CGC CAG AAT GGA GAT 475
476 GAT CCC TTG CTG ACT TAC CGG TTC CCA CCA AAG TTC ACC CTG AAG 790
491 GCT GGG CAG GTG GTG ACG 496

EXON 9  120 bp
497 ATC TGG GCT GCA GGA GCT GGG GCC ACC CAC AGC CCC CCT ACC GAC 511
512 CTG GTG TGG AAG GCA CAG AAC ACC TGG GGC TGC GGG AAC AGC CTG 526
527 CGT ACG GCT CTC ATC AAC TCC ACT GGG GAA 536

EXON 10  90 bp   1580 G>T
537 GAA GTG GCC ATG CGC AAG CTG GTG CGC TCA GTG ACT GTG GTT GAG 551
552 GAC GAC GAG GAT GAG GAT GGA GAT GAC CTG CTG CAT CAC CAC CAC 566

EXON 11  270 bp
567 GGC TCC CAC TGC AGC AGC TCG GGG GAC CCC GCT GAG TAC AAC CTG 581
582 CGC TCG CGC ACC GTG CTG TGC GGG ACC TGC GGG CAG CCT GCC GAC 596
597 AAG GCA TCT GCC AGC GGC TCA GGA GCC CAG GTG GGC GGA CCC ATC 611
612 TCC TCT GGC TCT TCT GCC TCC AGT GTC ACG GTC ACT CGC AGC TAC 626
627 CGC AGT GTG GGG GGC AGT GGG GGT GGC AGC TTC GGG GAC AAT CTG 641
642 GTC ACC CGC TCC TAC CTC CTG GGC AAC TCC AGC CCC CGA ACC CAG 656

EXON 12  1001 bp
657 AGC CCC CAG AAC TGC AGC ATC ATG TAA // 665
```

Figure 2A: The last five exons of *LMNA* gene showing G>T substitution in codon 527 (exon 9), at base position 1580 of *LMNA* gene coding region (c.1580G>T), causing substitution of (CGT) for Arginine to (CTT) for Leucine, at position 527 of Lamin A/C protein (p.Arg527Leu).

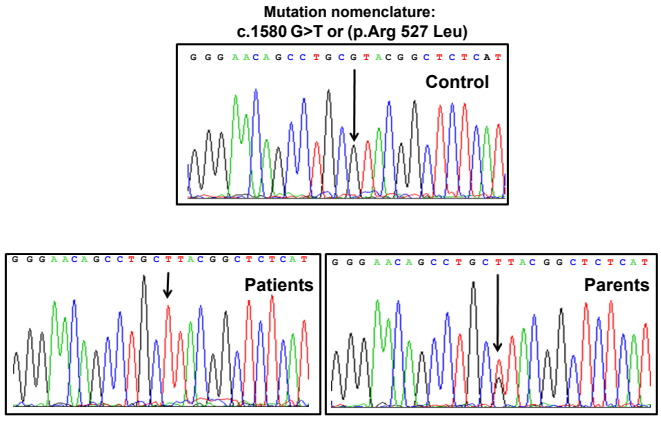

Figure 2B: Sequencing results of *LMNA* gene; note homozygous mutation in the patients (T base) and heterozygous mutation in the consanguineous parents (G-T bases overlap), compared to wild sequence in control (G base).

storage. Another significant challenge is genome interpretation, which includes not only the analysis of genomes for functional elements, but the understanding of the significance of variants in individual genomes on human phenotypes and disease. All these add to the still-impractical costs of vast sequencing applications in the clinic.

The benefits of sequencing applications in the medical clinic definitely look promising, and also it is necessary, in the future, to develop ways to map sequencing data onto currently difficult-to-map regions, such as highly repetitive and low-expressed regions. Sequencing technology is rapidly improving, but the analytical capabilities to understand everything that is being generated by the sequencers is lagging far behind. We need to advance the computational technologies and skills in Bioinformatics as we progress towards the systemic use of high-throughput sequencing in research and medicine.

References

1. Singh DP, Prabha R, Rai A, Arora DK (2012) Bioinformatics-assisted microbiological research: Tasks, developments and upcoming challenges. Am J Bioinformatics 1: 10-19.
2. Kanehisa M, Bork P (2003) Bioinformatics in the post-sequence era. Nat Genet 33: 305-310.
3. Xiong J (2006) Essential bioinformatics. Cambridge University Press, UK.
4. Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. J Appl Genet 52: 413-435.
5. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74.
6. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, et al. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409: 533-538.
7. Horak CE, Snyder M (2002) ChIP-chip: A genomic approach for identifying transcription factor binding sites. Methods Enzymol 350: 469-483.
8. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.
9. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57-74.
10. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, et al. (2012) Annotation of functional variation in personal genomes using regulomedb. Genome Res 22: 1790-1797.
11. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, et al. (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science 330: 1775-1787.
12. Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, et al. (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol 13: 418.
13. Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, et al. (2009) Recurring Mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med 361: 1058-1066.
14. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, et al. (2009) Mutational evolution in a lobular breast tumor profiled at single nucleotide resolution. Nature 461: 809-813.
15. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-Calling of automated sequencer traces using Phred. I. accuracy assessment. Genome Res 8: 175-185.
16. Goldberg SM, Johnson J, Busam D, Feldblyum T, Ferriera S, et al. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci USA 103: 11240-11245.
17. Reinhardt JA, Baltrus DA, Nishimura MT, Jeck WR, Jones CD, et al. (2009) De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. Genome Res 19: 294-305.
18. Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawa K, et al. (2002) A physical map of the mouse genome. Nature 418: 743-750.
19. Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E, et al. (2010) De novo assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora,* a model organism for fungal morphogenesis. PLoS Genet 6: e1000891.
20. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821-829.
21. Chaisson MJ, Pevzner PA (2008) Short read fragment assembly of bacterial genomes. Genome Res 18: 324-330.
22. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, et al. (2009) ABySS: A parallel assembler for short read sequence data. Genome Res 19: 1117-1123.
23. Li R, Zhu H, Ruan J, Qian W, Fang X, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20: 265-272.
24. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, et al. (2010) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci 108: 1513-1518.
25. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
26. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.
27. Li R, Yu C, Li Y, Lam TW, Yiu SM, et al. (2009) SOAP2: An improved ultrafast tool for short read alignment. Bioinformatics 25: 1966-1967.
28. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, et al. (2008) Whole-genome sequencing and variant discovery in *C. elegans.* Nat Methods 5: 183-188.
29. Zhang W, Cui H, Wong LJ (2012) Application of next generation sequencing to molecular diagnosis of inherited diseases. Top Curr Chem.
30. Tester DJ, Ackerman MJ (2011) Genetic testing for potentially lethal, highly treatable inherited cardiomyopathies/ channelopathies in clinical practice. Circulation 123: 1021-1037.
31. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR (2008) Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci USA 105: 16266-16271.
32. Al-Haggar M, Taranta A, El-Hawary A, Al-Said A, Shaban A, et al. (2012) Novel truncating mutation in the CTNS gene in an Egyptian family with cases of infantile nephropathic cystinosis and congenital heart malformations. Middle East J Med Genet 1: 71-75.
33. Al-Haggar A, Madej-Pilarczyk A, Kozlowski L, Bujnicki JM, Yahia S, et al. (2012) A novel homozygous p.Arg527Leu LMNA mutation in two unrelated Egyptian families causes overlapping mandibuloacral dysplasia and progeria syndrome. Eur J Hum Genet 20: 1134-1140.
34. Ng S, Buckingham K, Lee C, Bigham AW, Tabor HK, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42: 30-35.
35. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, et al. (2011) Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med 13: 255-262.
36. Bainbridge MN, Wiszniewski W, Murdock DR, Friedman J, Gonzaga-Jauregui C, et al. (2011) Whole-genome sequencing for optimized patient management. Sci Transl Med 3: 87re3.

This article was originally published in a special issue, **Bioinformatics for Highthroughput Sequencing** handled by Editor: Dr. Heinz Ulli Weier, Lawrence Berkeley National Laboratory, USA