Development of Genomic Based Diagnostics in Various Application Domains

(Extended Abstract)

© Zoltan Szallasi

Department of Bio and Health Informatics, Technical University of Denmark,
Kemitorvet 208, 2800
Lyngby, Denmark,
Computational Health Informatics Program, Boston Children's Hospital, USA,
Harvard Medical School,
Boston, USA

zszallasi@chip.org

Abstract. We will review the revolution brought about by low cost next generation sequencing in a wide array of diagnostic and industrial applications with a special emphasis on computational requirements and big data challenges.

Keywords: next generation sequencing, big data challenges.

1 Introduction

Next generation sequencing ((NGS) has fundamentally changed modern biological research. It is, in fact, an excellent example of how gradual improvements on a powerful initial idea, Sanger's original dideoxynucleotide sequencing, can lead to such levels of quantitative increase in data production that fundamentally changes a given research field.

Virtually any nucleic acid related research question can be investigated in a comprehensive, high resolution fashion free from experimental confounding factors such as nucleic acid cross hybridization. This has produced a deluge of data on the scale of hundreds of Terabytes even for a single research laboratory. This review will survey both the various application domains of next generation sequencing and their associated computational and analytical challenges.

2 Biochemical considerations of next generation sequencing

It was recognized early on that next generation sequencing will allow querying both the genome (DNA) and the transcriptome (RNA) on a wide range of resolution. The exact sequence of nucleotides (e.g. single nucleotide polymorphisms, single nucleotide variations) and the overall architecture of the entire genome can be determined in a single experiment (one run of whole genome sequencing) [1].

A wide array of starting materials can be used for next generation sequencing. Any form of nucleic acid (DNA or RNA), from any sources (from inside the cell, from cell free biological fluids or ancient fragmented DNA) can be sequenced and quantified. Nucleic acids can be preselected as in e.g. whole exome sequencing (exon capture) or by other capture mechanisms such as specific protein beacons in ChipSeq analysis. The variations are virtually unlimited and novel approaches are constantly being added to the toolbox of biological research. This universality has led to an enormous variety of application domains.

3 Application domains of next generation sequencing

3.1 Next generation sequencing in microbiology

Since DNA is universal, next generation sequencing can be used to detect and investigate any life form from viruses to humans. This is readily exploited in the various forms of sequencing based microbial diagnostics and it has also led to a new research field, metagenomics [2]. In this, the fact that often a diverse group of microbiota live together as an "organic whole" has led to the realization that those species do not need to be isolated individually before sequencing, but the pooled DNA can be sequenced together and the sequence tags can be "sorted out" after the sequencing reaction. This ingenious idea has led to significant advances in our understanding of, for example, the microbial community of our gut flora. This method also allows monitoring sewage quality and may help to monitor and prevent disease outbreaks.

3.2 Next generation sequencing in human diseases

Genome wide association studies are a powerful method to identify germline variants associated with increased disease risk.

Remarkably, germline DNA of the fetus can be efficiently detected in maternal blood, leading to the powerful tool of prenatal testing [3]. By this, genomic, chromosomal aberrations of the fetus can be detected without virtually any risk to the mother or the fetus.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

3.3 Next generation sequencing in cancer

Cancer is a genetic disease. Accumulating mutations at various levels of the germline genome lead to malignant transformation. It is therefore, obvious, that one of the main targets of next generation sequencing is cancer diagnostics. Both germline and somatic mutations are readily identified by NGS. A great number of oncogenic mutations, many of them targetable by therapy, have thus been identified [4]. NGS data also allow us to reconstruct the evolutionary history of cancer, an issue of potentially great significance [5].

Recently, sequence analysis of liquid biopsies, essentially cell free DNA obtained from various bodily fluids, emerged as a minimally invasive tool to obtain vital information about the presence of cancer in a patient individual before or during therapy.

4 Computational and analytical challenges of next generation sequencing data

The uniform nature of the biochemical reactions and ingenuity of technical development has led to an unexpected situation. The price drop/throughput increase of next generation sequencing has significantly outpaced Moore's law over the past decade. Therefore, next generation sequencing has become more of a computational rather than a biochemical problem. The speed of data accumulation is so fast that data storage and data analysis is becoming a more and more challenging problem in modern sequencing based projects. Whole genome sequencing on a single cancer sample can easily take up hundreds of GBs of data storage. Therefore, a single study, such as the one analyzing the whole genome of 560 breast cancer cases [6], can easily produce data on at the level of hundreds of Terabytes. Such amount of data cannot possibly be downloaded for reanalysis in an efficient manner, therefore alternative solutions, such as cloud based computing had to be found.

Management of vast amounts of data is only one, mainly technical aspect of the challenges at hand.

While next generation sequencing based genomics easily qualifies as one of the main areas of big data science, in many aspects it is also markedly different from those. While in, e.g., financial data the individual variables are connected by poorly understood causative factors and in physics the entire data space is regulated by well defined, homogenous laws of physics, in biology, genomics the situation lies somewhere in between. Variables, such as genes, proteins etc. are connected by the principles of physical chemistry, but the actual parameters of those significantly vary across the various pairs of biological entities. This fact places the analysis of biological systems in the realm of robust, complex systems for which the analytical principles are poorly understood. Therefore, in order to effectively analyze the massive amounts of genomic information one needs to "front-load" the computational analysis with as much biological knowledge as possible.

We will present several strategies along those lines. In particular, we will discuss how genomics, next generation sequencing based whole genome analysis helps us to understand DNA repair pathway aberrations, and their diagnostic and therapeutic implications in cancer. We will also discuss how genomics is exploited to understand the main principles of therapeutic immune responses against cancer and how genomics, machine learning and high throughput screening are combined in an interdisciplinary environment to design effective vaccines against cancer.

5 The industrial impact of next generation sequencing

In order to satisfy the need for NGS based diagnostics a whole industry has developed during the past decade. Conferences such as the 2017 Next Generation D_x Summit, (http://www.nextgenerationdx.com) provide an excellent overview of the major trends and players.

References

[1] Mardis ER (2013) Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif) 6:287–303. doi: 10.1146/annurev-anchem-062012-092628

[2] Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Comput Biol 6:e1000667. doi: 10.1371/journal.pcbi.1000667

[3] Hui L, Bianchi DW (2017) Noninvasive Prenatal DNA Testing: The Vanguard of Genomic Medicine. Annu Rev Med 68:459–472. doi: 10.1146/annurev-med-072115-033220

[4] Lawrence MS, Stojanov P, Mermel CH, et al (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505:495–501. doi: 10.1038/nature12912

[5] Jamal-Hanjani M, Wilson GA, McGranahan N, et al (2017) Tracking the Evolution of Non-Small-Cell Lung Cancer. N Engl J Med 376:2109–2121. doi: 10.1056/NEJMoa1616288

[6] Nik-Zainal S, Davies H, Staaf J, et al (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534:47–54. doi: 10.1038/nature17676