Critical Reviews in Oncology/Hematology 96 (2015) 463–474

Contents lists available at ScienceDirect

Critical Reviews in Oncology/Hematology

journal homepage: www.elsevier.com/locate/critrevonc

Review

Understanding next generation sequencing in oncology: A guide for oncologists

Sing Yu Moorcraft, David Gonzalez, Brian A. Walker*

The Royal Marsden NHS Foundation Trust, Surrey, United Kingdom

Contents

1. Introduction 464
2. Essential terminology 464
3. Genetic variations 464
   3.1. Single nucleotide variation (SNV) 464
   3.2. Insertions and deletions 465
   3.3. Structural variations and copy number changes 465
   3.4. Polymorphisms 465
   3.5. Germline and somatic mutations 465
4. The consequences of genetic abnormalities 465
5. Sequencing techniques 466
   5.1. Sanger sequencing 466
   5.2. Next generation sequencing 466
      5.2.1. Creation of a sequencing library 467
      5.2.2. Sequencing reactions and detection 467
6. Interpreting sequencing data 468
   6.1. Overview of data output 468
   6.2. Reference mapping 468
   6.3. Variant calling 468
   6.4. Variant annotation 469
7. Factors influencing the choice of sequencing method 469
   7.1. Applications of next generation sequencing 469
      7.1.1. Whole genome sequencing 469
      7.1.2. Exome sequencing 469
      7.1.3. Targeted sequencing 469
      7.1.4. Other next generation sequencing applications 469
   7.2. Sequencing depth, coverage and platform performance metrics 470
   7.3. Speed and cost of sequencing 471
   7.4. Material available for sequencing 471
8. Challenges in next generation sequencing 471
   8.1. Cost and infrastructure requirements 471
   8.2. Data interpretation and management 471
   8.3. Ethical considerations 472
9. Future applications of sequencing 472
10. Conclusion 472
   Conflicts of interest 472
   Acknowledgement 472
   References 472
   Biography 474

* Corresponding author. Present address: Centre for Molecular Pathology, The Royal Marsden NHS Foundation Trust, Sutton SM2 5PT, United Kingdom. Fax: +44 208 915 6566.
E-mail address: brian.walker@icr.ac.uk (B.A. Walker).

http://dx.doi.org/10.1016/j.critrevonc.2015.06.007

ARTICLE INFO

Article history:
Received 20 May 2014
Received in revised form 21 May 2015
Accepted 17 June 2015

Keywords:
Next generation sequencing
NGS
Bioinformatics
Genetics

ABSTRACT

DNA sequencing is now faster and cheaper than ever before, due to the development of next generation sequencing (NGS) technologies. NGS is now widely used in the research setting and is becoming increasingly utilised in clinical practice. However, due to evolving clinical commitments, increased workload and lack of training opportunities, many oncologists may be unfamiliar with the terminology and technology involved. This can lead to oncologists feeling daunted by issues such as how to interpret the vast amounts of data generated by NGS and the differences between sequencing platforms.

This review article explains common concepts and terminology, summarises the process of DNA sequencing (including data analysis) and discusses the main factors to consider when deciding on a sequencing method. This article aims to improve oncologists' understanding of the most commonly used sequencing platforms and the ongoing challenges faced in expanding the use of NGS into routine clinical practice.

1. Introduction

Advances in DNA sequencing technology have revolutionised genomic research. It took more than a decade and approximately US$3 billion to sequence the first draft of the human genome using Sanger sequencing, whereas whole genome sequencing can now be performed in less than 24 h for under $1,000 (Morey et al., 2013; National Human Genome Research Institute, 2013; Hayden, 2014).

A good understanding of genomics is critical in oncology, due to the importance of genetic abnormalities in cancer development and progression. Genetic abnormalities can be predictors of a patient's prognosis (e.g. acquired *BRAF* mutation confers a poor prognosis in metastatic colorectal cancer (Sclafani et al., 2013)) or identify patients who have an increased susceptibility to cancer, e.g. inherited mutations in the *BRCA1* and *BRCA2* genes are associated with increased risk of developing breast cancer (Ford et al., 1998). In addition, genetic alterations can also determine suitability for anticancer drugs, particularly when they exhibit oncogenic addiction to specific cell-signalling pathways, e.g. vemurafenib for *BRAF*-mutant melanoma, crizotinib for *ALK*-translocated lung cancer and panitumumab for RAS wild-type colorectal cancer (Chapman et al., 2011; Douillard et al., 2013; Shaw et al., 2013). DNA sequencing is now widely used in the research setting, e.g. whole genome or whole exome sequencing has been performed on large cohorts in a number of cancers (including leukaemia, glioblastoma, oesophageal, pancreatic and colorectal cancers) as part of international collaborative projects such as the Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), and it has the potential of being utilised in clinical practice (Biankin et al., 2012; Cancer Genome Atlas Network, 2012; Dulak et al., 2013; Parsons et al., 2008).

However, many oncologists have received limited training in genomics and therefore may not be aware of the capabilities and challenges of sequencing technologies. This review aims to provide clinicians with the information required to understand the principles of DNA sequencing, including an explanation of the main terminology, an overview of the sequencing process and data interpretation, a comparison of the different sequencing platforms and a discussion of some of the ongoing challenges in incorporating sequencing into routine clinical practice. This review does not aim to provide detailed technical information regarding sequencing techniques, but this information can be found in other articles (Clark et al., 2011; Liu et al., 2012; Meldrum et al., 2011; Quail et al., 2012; Voelkerding et al., 2009).

2. Essential terminology

In order to understand DNA sequencing, it is essential to have a good understanding of the basics of genetics. Deoxyribonucleic acid (DNA) is the basic unit that encodes the genetic instructions required for functioning of all living organisms. DNA is a double-stranded helix comprised of four nucleotides containing different bases: adenine (A), guanine (G), cytosine (C) and thymine (T). These DNA strands, containing all the information that a cell needs to function, are organised in chromosomes. The double-strand structure is based on complementarity of the bases that form the DNA, e.g. adenine pairs with thymine and guanine pairs with cytosine, to form units called base pairs (bp). The DNA provides the template that is used to create ribonucleic acid (RNA), including messenger RNA (mRNA) by a process called transcription (The Translational Research and Personalised Medicine Working Group, 2015). This mRNA is subsequently translated into a chain of amino acids to form a protein by a process called translation. A codon is a set of three consecutive bases, and each codon can be translated into a particular amino acid or indicates the end of the protein (e.g. the codon GTC corresponds to the amino acid valine and TAG is one of the three stop codons).

A "genome" is a complete set of chromosomal DNA, and in humans it comprises approximately 3 billion base pairs organised into 23 pairs of chromosomes. However, not all of these base pairs are involved in coding for proteins, as the genome consists of protein-coding regions (exons) and non-coding regions (introns and intergenic regions). The complete set of protein-coding regions is termed the "exome" and represents approximately 1–2% of the genome.

Mutational signatures can also be identified. A mutational signature is a pattern of mutations caused by a particular mutational process, such as exposure to tobacco carcinogens or defective DNA repair (Alexandrov et al., 2013). Most cancer types contain at least two mutational signatures, and although some signatures are confined to one type of cancer, others are found in multiple cancer types.

3. Genetic variations

3.1. Single nucleotide variation (SNV)

SNVs, also referred to as "substitutions", occur when one base is substituted for another (e.g. an adenine for a cytosine). This changes the DNA at a single point (and therefore is also known as a "point" mutation). The effect of this point mutation can vary dramatically and also determines its classification. Mutations can be classed as "missense" (non-synonymous), "silent" (synonymous) or "nonsense." A missense mutation results in a change from one amino acid to another, e.g. a change from GTA to GAA would cause the amino acid to change from valine to glutamic acid. A silent mutation does not result in a change in the amino acid as there is redundancy, with many amino acids being coded for by a number of different