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

* 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

codons. For example, a mutation from CGT to CGC would still lead to the production of arginine because both of these codons represent arginine. A nonsense mutation results in the introduction of a premature stop codon. For example, a mutation from TCA to TAA results in the production of the stop codon TAA rather than the amino acid serine. This could lead to a smaller protein being produced, as DNA after the stop codon would not be translated, and is likely to substantially affect the protein's function. For example, in colorectal cancer a high proportion of point mutations in *APC* lead to the introduction of a stop codon, leading to impaired binding to β-catenin and axin (Cottrell et al., 1992; Smith et al., 2002).

If the mutation leads to a change in the amino acid, this may or may not have a significant effect on the final protein produced and its function. Some changes may have a minimal effect or no effect at all on the protein. Other changes may result in abnormalities in protein folding, or in the protein's ability to bind to receptors or other molecules. Mutations in protein binding sites or active motifs (e.g. tyrosine kinase domains) may have a particularly significant effect on a protein's function. For example, the BRAF V600E mutation changes the amino acid at position 600 from a valine (V) to a glutamic acid (E). This mutation occurs in the activation segment of the kinase domain, destablising the conformation that usually maintains the kinase in an inactive state, and results in constitutive kinase activation (Cantwell-Dorris et al., 2011).

3.2. Insertions and deletions

Other variations are called indels, which is an abbreviation of insertions or deletions. This means that one or more bases has been inserted into the sequence, or deleted from the sequence. This can result in "frameshift" mutations if the number of nucleotides inserted or deleted is not multiple of 3 (the size of a codon). For example, if an extra base is inserted then this will mean that each codon thereafter will start in a slightly different place (new "frame") and therefore the amino acid sequence from this point onwards in the protein will change. Conversely, if the number of nucleotides inserted or deleted is a multiple of 3 ("in frame"), this will simply change, add or delete one or more amino acids, without altering the remaining sequence of the protein. In most cases, frameshift mutations result in the appearance of a premature stop codon and therefore tend to be deleterious as the mutant protein will not have the original function. In contrast, "in frame" mutations can simply enhance, transform or reduce the function of the original protein. For example, in-frame deletions in exon 19 of *EGFR* occur in approximately 48% of patients with EGFR-mutated non-small cell lung cancer (Mitsudomi and Yatabe, 2010). This leads to increased phosphorylation of EGFR without ligand stimulation and is associated with increased sensitivity to EGFR tyrosine kinase inhibitors such as erlotinib and gefitinib (Mitsudomi and Yatabe, 2010). Insertions in exon 20 of *EGFR* are usually in-frame insertions and/or duplications of 3–21 base pairs (Arcila et al., 2013). This can affect the binding of erlotinib and gefitinib and confers resistance to these agents (Arcila et al., 2013). In contrast, common deleterious mutations in *BRCA1* include small frameshift insertions or deletions which introduce a premature stop codon and therefore protein truncation and impaired protein function (Borg et al., 2010).

3.3. Structural variations and copy number changes

Genomic changes can also occur on a much larger scale than just changes in a single nucleotide. Copy number variations (CNV) occur when large areas of a gene or chromosome (or entire chromosomes) are deleted or duplicated, e.g. due to structural rearrangements such as inversions and translocations, tandem duplications or chromosomal gains or losses. This can lead to changes in the level of expression of genes in the area affected and are important in the development and progression of cancer (Zack et al., 2013). For example, *HER2* amplification is associated with a poor prognosis and response to trastuzumab in patients with breast cancer (Gajria and Chandarlapaty, 2011; Slamon et al., 1987) and *PTEN* deletion leads to dysregulation of the phosphatidylinositol-3-kinase (PI3K) pathway and a poor clinical outcome for patients with prostate cancer (Krohn et al., 2014).

In addition, chromosomal rearrangements and translocations can occur, leading to genes being re-localised into other areas of the genome. These translocations normally have functional consequences, as they can lead to fusion genes (where one part of a gene is "fused" with another) that will translate into chimeric proteins (with abnormal function), or they can lead to a gene being controlled by a different promoter or enhancer region that can alter significantly the expression of the protein. For example, *EML4-ALK* fusion genes are seen in up to 7% of non-small cell lung cancers (Koivunen et al., 2008; Kwak et al., 2010). These occur when an approximately 13 Mb section of chromosome 2 is inverted, resulting in a fusion between the anaplastic lymphoma kinase *(ALK)* gene and the echinoderm microtubule-associated protein-like 4 (*EML4*) gene, promoting dimerisation and constitutive kinase activity (Koivunen et al., 2008; Kwak et al., 2010).

3.4. Polymorphisms

A polymorphism is a DNA sequence variation that commonly occurs in the population. Whether a genetic variation is called a mutation or a polymorphism is determined by the frequency in which it occurs in the population, and an arbitrary cut-off level of 1% is used. This means that a variation is termed a polymorphism if it occurs in >1% of the population, and is termed a mutation if it occurs in <1% of the population (Crawford and Nickerson, 2005; Erichsen and Chanock, 2004). Single nucleotide polymorphisms (SNPs) are the most frequently occurring types of SNVs and represent a change in a single base. Many SNPs occur in non-coding regions of the genome and the majority of SNPs have no known clinical significance (Venter et al., 2001). However, other SNPs may be associated with response to drugs or to the risk of developing certain diseases, although it can be difficult to determine the causative SNP as many SNPs are correlated with one another (Crawford and Nickerson, 2005; Erichsen and Chanock, 2004). Variations other than SNVs can be polymorphic, including indels and copy number variations.

3.5. Germline and somatic mutations

A key aspect in the analysis of genetic variation in cancer is to identify the nature of the mutations, as these can be inherited or occur early in the embryogenesis process (also referred as germline) or acquired exclusively in the cancer cells (normally referred to as somatic). Examples of germline mutations are those that confer a significant risk of developing particular types of cancer, such as BRCA1/2 mutations associated with breast and ovarian cancer, or *TP53* mutations in Li-Fraumeni syndrome. Most mutations in cancer are somatically acquired and are therefore not present in normal tissue from the same patient.

4. The consequences of genetic abnormalities

Understanding the different types of genetic abnormalities is important because it helps in the understanding of the limitations of the technology employed to characterise them. Sanger sequencing is useful for substitutions and small insertions or deletions. However, it is more challenging to identify large insertions/deletions or other structural variants using Sanger sequencing and therefore other techniques such as fluorescence

in situ hybridisation (FISH) and comparative genomic hybridisation (CGH) may be employed.

It is important to remember that even if an abnormality leads to a significant impact on a particular protein, this may be of variable biological or clinical significance due to the complexity of the mechanisms involved in cancer. Some proteins may be key components of a particular cell-signalling pathway, and therefore anything that affects their function may have significant biological consequences and result in cancer growth or metastasis. When cancer cells are particularly driven by a specific gene, this phenomenon is referred to as "oncogenic addiction" (Weinstein and Joe, 2006). For example, *BRAF* mutant melanoma cells are dependent on the BRAF-mutant protein to continue their proliferation, and targeting mutant BRAF with vemurafenib has resulted in significant clinical efficacy (Chapman et al., 2011).

Other proteins may be of less importance as, for example, abnormalities in their function may be compensated for by other signalling pathways which may be correspondingly up- or down-regulated or they lose their significance due to intra-tumoural heterogeneity (Crockford et al., 2014; Gerlinger et al., 2012). It has become increasingly apparent that the molecular profile of a tumour can vary, not only between the primary tumour and metastatic sites, but also between different areas of the same tumour and between individual cancer cells (Meric-Bernstam and Mills, 2012). Therefore, establishing the clinical significance of a single genetic abnormality identified from a biopsy of a single tumour region is challenging. One method of visualising intratumour heterogeneity is the "trunk-branch" model, in which the trunk contains driver mutations which are present in all tumour subclones, whereas the branches contain a variety of mutations which are not present in all tumour regions (Yap et al., 2012). This is an important concept, as drugs targeting mutations in a particular "branch" will not be effective in tumour regions which do not contain these mutations. Therefore, although genetic sequencing can provide information on the frequency of a mutation in the DNA sample analysed, this may not be representative of other areas of the tumour.

Furthermore, the molecular profile of a tumour can change over time and in response to treatment, leading to the acquisition of genetic abnormalities that confer resistance to therapeutic agents. For example, the development of mutations in NRAS or MEK can result in acquired resistance to vemurafenib in patients with BRAF mutant melanoma (Trunzer et al., 2013), and amplification of KIT can reduce the sensitivity of non-small cell lung cancers to crizotinib (Katayama et al., 2012).

Trials of molecular screening to stratify patients for targeted therapies have not yet shown any significant patient benefit, however these studies have important limitations including the inclusion of heavily pre-treated patients, the use of exploratory biomarkers to select treatment and the allocation of patients to potentially non-biologically active doses due to the nature of phase I studies (Andre et al., 2014; Dienstmann et al., 2012).

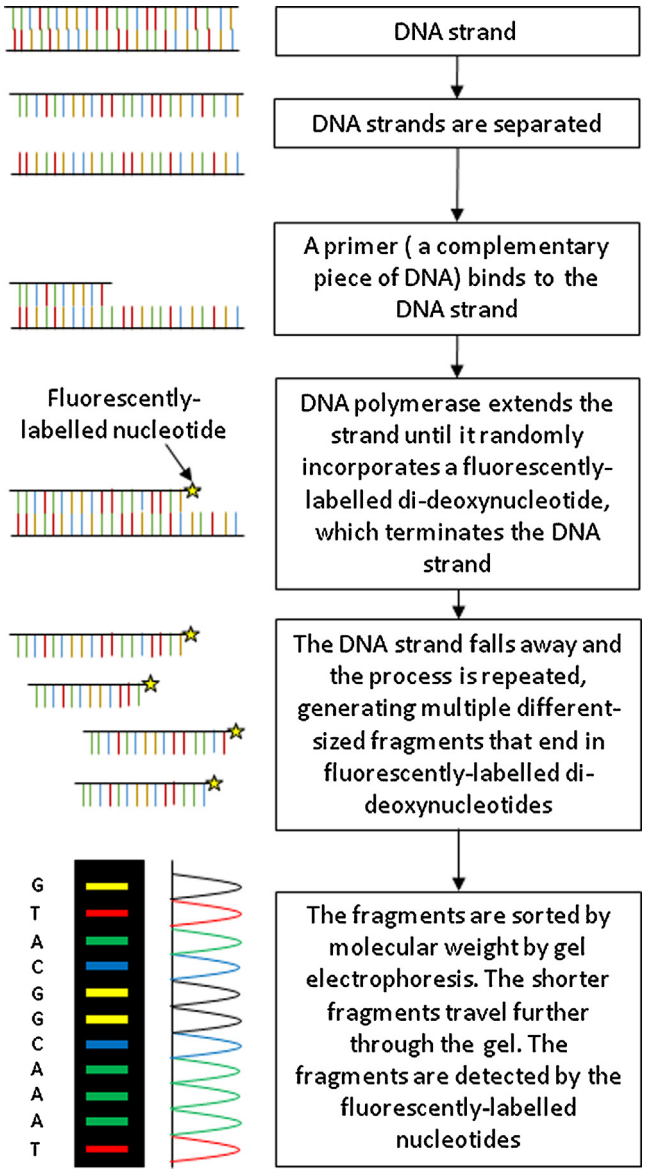

Fig. 1. Overview of Sanger sequencing.

5. Sequencing techniques

5.1. Sanger sequencing

Sanger sequencing, developed in the 1970s, is the most widely used method of DNA sequencing and was used to sequence the first genome (a bacteriophage) in 1977 and by the Human Genome Project (Lander et al., 2001; Sanger et al., 1977a,b). An overview of Sanger sequencing is shown in Fig. 1. The first step in Sanger sequencing involves amplifying the DNA sequence. Chemically altered bases called di-deoxy nucleotides are introduced together with normal nucleotides, resulting in random termination of the DNA when a di-deoxy base is incorporated. This results in production of all possible fragments of the target sequence. The fragments are sorted by their molecular weight (Sanger et al., 1977b). This was originally performed by gel electrophoresis, but this has now been replaced by capillary electrophoresis (Swerdlow and Gesteland, 1990). Each di-deoxy base is labelled with a fluorescent dye allowing the last base to be determined by a laser, producing an ordered read of the nucleotides present in the original DNA sequence.

5.2. Next generation sequencing

Sequencing "throughput" refers to the number of DNA sequences which can be read with each sequencing reaction. Sanger sequencing has a low throughput and therefore the sequencing power is low. To combat the limited throughput with Sanger sequencing, newer sequencing technologies were developed (Bentley et al., 2008; Eid et al., 2009; Margulies et al., 2005; Ronaghi et al., 1996; Rothberg et al., 2011; Shendure et al., 2005). These technologies have been collectively referred to as "next generation sequencing" (NGS) or "massively-parallel sequencing" (MPS). The main difference between Sanger sequencing and NGS

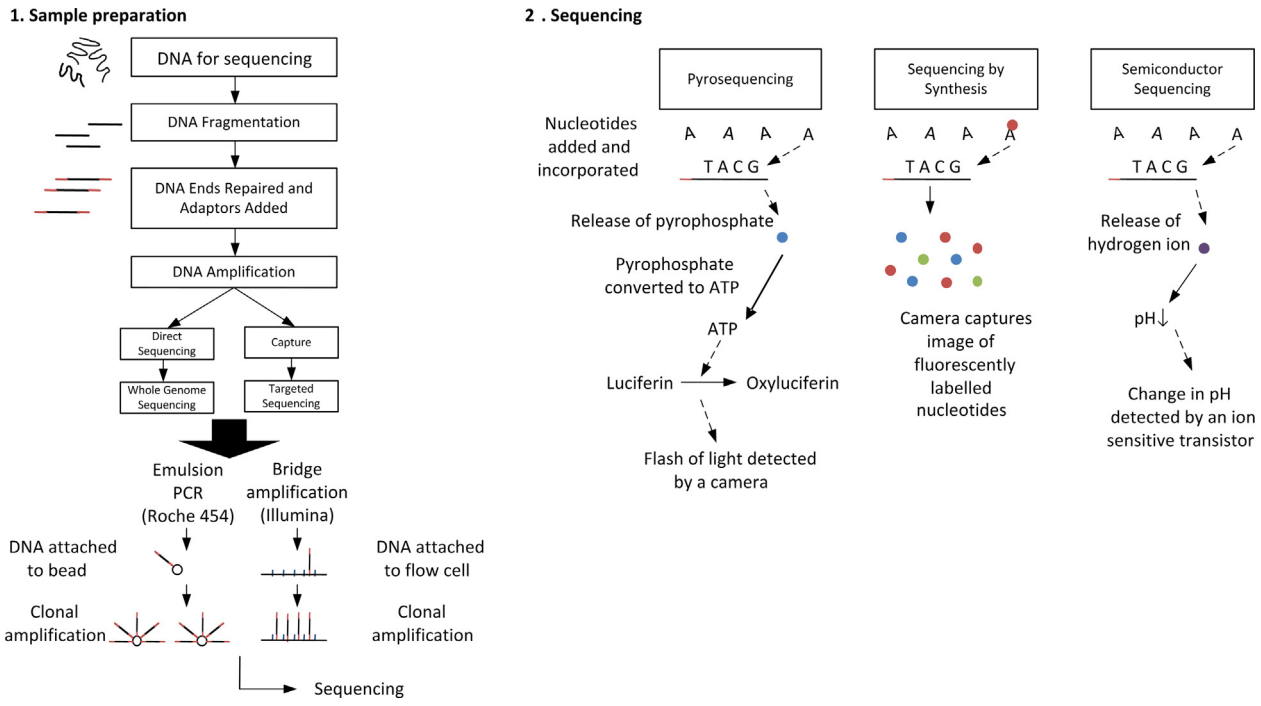

Fig. 2. Overview of next generation sequencing.

is that NGS sequences millions of small DNA fragments at the same time (i.e. in parallel) and therefore dramatically increases the throughput per reaction. The current next generation sequencing is "second generation sequencing," while newer "third generation" sequencing platforms also known as "single molecule sequencing" can remove the need for polymerase chain reaction (PCR) and need much less starting material but are currently less well established and their advantages over second generation platforms have not yet been established (Schadt et al., 2010).

The precise sequencing process varies depending on the individual sequencing platform used, but the general steps are as follows (see Fig. 2):

5.2.1. Creation of a sequencing library

There are two main methods of detecting variations by NGS: a targeted amplicon-based approach or a hybridisation-capture approach. These approaches determine how the sequencing library is created. Amplicon-based sequencing involves PCR with two primers flanking the DNA regions of interest (the "amplicons") generating the sequencing library. Many samples can be analysed simultaneously by changing the ends of the primers, creating multiple 'barcodes' that can be used to identify the DNA fragments for each sample. Amplicon-based sequencing is quick but provides limited results, whereas hybridisation-capture enables the analysis of larger amounts of DNA without the limitation of knowing the precise sequence of the flanking regions.

In contrast, a hybridisation-capture approach uses a "capture probe" to bind to the DNA sequence of interest, allowing for enrichment of the regions of interest, to create the sequencing library. For the hybridisation-capture approach, the basics of sample preparation are the same regardless of whether the samples will later undergo genome, exome or targeted sequencing. The first step in the library preparation is to produce manageable fragments of DNA. This can be achieved either mechanically (usually by sonication) or by enzymatic digestion of the DNA.The fragment size required depends on the sequencing platform which will be used and the type of sample used. Following fragmentation, the DNA ends are repaired and adaptors are ligated onto the ends using a series of enzymatic steps. Tagmentation is a newer alternative to fragmentation and uses an enzyme to simultaneously fragment and tag the DNA with adaptors.

5.2.2. Sequencing reactions and detection

The next step is to amplify the DNA so that the signal is strong enough to be detected during the sequencing. This is done by a process called PCR. PCR involves heating the DNA so that it separates into two strands. An enzyme called Taq DNA polymerase uses the strands as templates to synthesize new DNA strands, which can be used to create more copies of the DNA. The process is repeated, creating more and more copies of the original DNA strands. The exact details of the PCR technique used varies on the sequencing platform, e.g. Roche and Life Technologies use emulsion-PCR (emPCR) and Illumina use bridge-PCR. In emPCR, DNA fragments, primer-coated beads and other reagents required for PCR are mixed together, resulting in one DNA fragment being captured onto each bead. These beads are suspended in a water-in-oil emulsion, with each emulsion droplet containing one bead and its associated DNA fragment. The DNA fragments are amplified by PCR and this means that each bead is coated with millions of copies of the DNA fragment (Dressman et al., 2003). The beads are then placed into wells ready for sequencing.

In bridge-PCR, a flow cell is coated with oligonucleotide probes. Each dsDNA fragment has two different adapters at each end. The DNA binds to the flow cell at one end and bends over to bind at the other end too, creating the bridge. The strands of the dsDNA separate and each is filled in with the complementary strand, recreating the dsDNA. The process is repeated several times creating a localised cluster of molecules that are identical to the first (Bentley et al., 2008). Both emPCR and bridge-PCR allow millions of microreactions to occur at the same time on each spatially separate template, but there is the potential for errors to be introduced at this stage as the DNA polymerases are not 100% accurate and therefore sequence changes can be introduced (Mardis, 2011).

Each NGS instrument processes the wells or flow cells containing the immobilised DNA templates. Nucleotides are added and detected, the reagents are then washed/removed and new

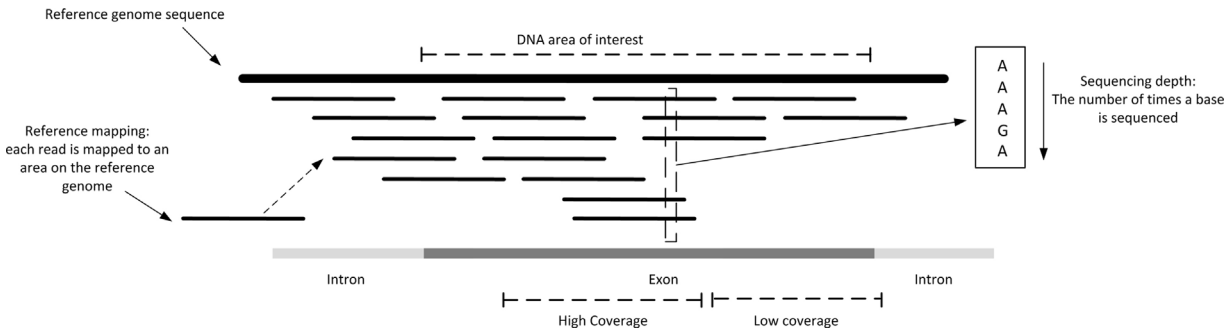

Fig. 3. Simplified overview of reference mapping, sequencing depth and coverage.

nucleotides are added. This process is repeated on a nucleotide by nucleotide basis until the whole DNA template has been sequenced. The different sequencing platforms use different techniques to detect the signal produced when each nucleotide is added. The Roche 454 platform uses a technique called pyrosequencing (Ronaghi et al., 1998). The addition of a nucleotide results in the release of pyrophosphate. The pyrophosphate is converted to ATP, which in turn converts luciferin to oxyluciferin and generates visible light. The light is detected by a camera and is seen as a peak in the raw data output. The amount of light generated is dependent on the amount of incorporated nucleotides, and therefore the higher the peak, the greater the number of nucleotides incorporated. In contrast, the Illumina platforms use a reversible dye terminator sequencing by synthesis technique. This technique involves attaching a fluorescent label to each nucleotide (one colour per base). A camera captures an image of the fluorescent colours and each colour is used to identify its corresponding nucleotide. (Liu et al., 2012). Unlike the Roche 454 and Illumina platforms, the Life Technologies Ion PGM and Ion Proton platforms do not detect optical signals, but instead measure changes in voltage. The addition of a nucleotide causes the release of a hydrogen ion. This lowers the pH of the surrounding solution and the change in pH is detected by an ion-sensitive transistor. The more nucleotides added, the lower the pH and the higher the voltage detected. This process is called semiconductor sequencing (Liu et al., 2012).

6. Interpreting sequencing data

6.1. Overview of data output

NGS generates large amounts of data and data interpretation remains a huge hurdle in the implementation of routine NGS. For example, the size of the data file generated by sequencing the whole exome of one patient is approximately 8 Gb, and a whole human genome is approximately 150 Gb (Strand Scientific Intelligence Inc., 2013). It has been suggested that every dollar spent on sequencing hardware needs to be matched with a comparable investment in informatics (Perkel, 2011), and there are a large variety of bioinformatics tools and software available.

"Base-calling" converts the raw data produced by the sequencing instrument into sequences of bases. The initial data output from any instrument is usually in the form of a text file in the FASTQ format (which also contains information on quality and other parameters) (Cock et al., 2010). This file contains millions of "reads." A "read" is a short sequence of letters that correspond to nucleotides (A, T, G and C). As each base call is an estimate of the true nucleotide, it can be wrong. A quality score assigned to a base reflects the confidence that it has been correctly identified. The quality score can vary, e.g. base quality tends to deteriorate towards the ends of reads. Each instrument has its own base quality score which is a derivation of the Phred quality score which was originally developed to determine the accuracy of Sanger sequencing. The read length and quality scores are used to determine a run's global quality.

6.2. Reference mapping

Reference mapping involves aligning the reads with specific chromosomal locations on a reference genome sequence obtained from online databases (see Fig. 3). Alignment software is typically built in to the instrument, but there are also a number of other third-party tools which can be used for alignment, such as MAQ (Li et al., 2008), BWA (Li and Durbin, 2009), Bowtie (Langmead et al., 2009), Novoalign (Novoalign, 2014) and SHRiMP (Rumble et al., 2009). The different platforms generate data in diverse formats, but commonly used formats include the sequence alignment/map (SAM) format which stores sequencing data as a text file and the binary alignment map (BAM) format, which stores the same information in a compressed, indexed, binary form. These SAM or BAM files are used to perform the variant calling analysis and to visualise the sequence reads in genome browsers such as the Integrative genomics viewer (IGV) (Robinson et al., 2011).

6.3. Variant calling

Once the sequencing data has been mapped, the next step is to look for abnormalities in the DNA sequence, a process termed variant calling. In order to identify variants, the sequenced DNA is compared to reference genome sequences. Germline variants are present in virtually every cell in the body. However, the majority of variants involved in cancer development are somatic variants. It is therefore important to compare the tumour DNA with germline DNA from the same patient. The germline sample (often a blood sample, buccal swab or non-neoplastic tissue) provides a baseline sequence for the patient and therefore enables somatic variants to be distinguished from germline variants.

The biggest challenge is identifying true variants and separating them from sequencing noise. There are many programs available for variant calling, such as Samtools (Li et al., 2009), GATK Unified Genotyper (DePristo et al., 2011), Illumina VariantStudio (Illumina, 2015), Torrent Variant Caller (Life Technologies Corporation, 2015) and MuTect (Cibulskis et al., 2013). These programs essentially use Bayesian algorithms to calculate the probability of each variant being a true variant based on the known sequencing errors and polymorphism rate. Various additional steps, such as removing duplicate reads and realigning reads around insertions and deletions, can help to increase the accuracy of variant detection. To improve call rates, tumour and normal samples can be realigned as pairs using software such as the GATK indel realigner. Different types of variants provide different challenges. SNVs are the most reliably detected variants, whereas indels and structural variants are relatively difficult to detect using NGS. Specialist software

has been developed to look for structural variants by looking for the flanking end regions of the NGS read data but detecting large amplifications and deletions remains challenging (Gullapalli et al., 2012).

Sequencing provides information on the frequency of individual variants. Due to tumour heterogeneity, not all cancer cells contain every variant (Gerlinger et al., 2012). Therefore, when sequencing a tumour sample, some reads will contain the variant and some will not (Walker et al., 2012). There is debate regarding the optimum cut-off point for deciding whether the frequency of the variant is high enough to be deemed to be present and of clinical significance. For example, in one sample a specific mutation may only be present in 20% of the reads, whereas in another sample the mutation may be present in 90% of reads, but in both cases the report issued to the clinician may only state that the mutation was detected.

6.4. Variant annotation

Once the variants have been detected, different informatics solutions can be used to annotate the results. These annotations may include gene and transcript identifiers, as well as predictions of the significance of the variant. Databases (e.g. ENSEMBL, COSMIC) and tools (e.g. Polyphen2, Oncotator, SnpEff or Alamut) contain details of known associations with disease and can facilitate automatic annotation of variants (McLaren et al., 2010) or help to predict the functional significance of variants (Karchin, 2009).

With reference mapping, variant calling and annotation, the method used for each step is not critical, and as the field is constantly moving the methods used will also change frequently. In a clinical setting, the important issue is that the method is validated against a gold standard. This standard may be samples that have been analysed for mutations using a known method, such as Sanger sequencing, for which you can repeat the test using NGS and get the same result or better. In this way, the sensitivity, specificity and confidence intervals of the new test and analysis pathway can be determined. Any changes to the bioinformatic analysis should be re-tested, possibly using the same dataset, to ensure that equal or better results can be obtained before the new protocol is implemented for diagnostic samples.

7. Factors influencing the choice of sequencing method

There are different types of sequencing, and deciding which sequencing method is most appropriate depends on factors such as the type of results required, type of material to be sequenced, the performance metrics of the various sequencing platforms (e.g. read length) and the overall cost. Summaries of the different types of sequencing and the different sequencing platforms are shown in Tables 1 and 2.

7.1. Applications of next generation sequencing

7.1.1. Whole genome sequencing

Whole genome sequencing (WGS) involves sequencing all of the base pairs in the genome, and therefore provides more data than whole exome sequencing (WES). However, these data are also more difficult to interpret because the clinical consequences of mutations in intronic regions are not yet clear. Although it seems logical that if the "whole genome" is sequenced then all variants will be detected, this is not technically true. Some variants can be missed due to variation in coverage across the genome, and in some cases these "missed" variants can be detected by WES due to the higher coverage achieved in certain areas with target-enriched sequencing (Clark et al., 2011). Therefore, as neither WES nor WGS detects all variants, some researchers advocate performing both exome and genome sequencing on the samples to ensure that as many variants as possible are detected. The main advantage of WGS is that it can detect structural aberrations that occur outside exonic areas in the genome, such as translocations and rearrangements. Additionally, variations occurring in DNA regions containing regulatory elements, such as enhancers or silencers can only be analysed by WGS.

7.1.2. Exome sequencing

Exome sequencing involves analysing the exons only. Mutations in an exon can lead to abnormalities in protein structure and function and therefore these mutations can be easier to interpret than intronic mutations. In addition, as less of the genetic material is sequenced, exome sequencing is generally faster to analyse and cheaper than WGS. Although, both WES and WGS can provide information regarding CNVs in genes, this remains challenging, particularly with data from WES (Tan et al., 2014).

7.1.3. Targeted sequencing

If only a small region of the genome is of interest, then targeted sequencing can be used to sequence the relevant individual genes or parts of genes. For example, some DNA sequences are highly susceptible to mutations, leading to a higher frequency of mutations in these areas, which are called mutation "hotspots" (Kandoth et al., 2013; Olivier et al., 2010; Rowan et al., 2000; Tennis et al., 2006; Ziegler et al., 1993). Therefore, sequencing the hotspot mutation region is likely to detect a large proportion of the relevant mutations and can be more efficient than sequencing large parts of a gene that are unlikely to be of clinical or biological interest. Various companies provide both pre-designed gene panels and customised panels designed around the requirements of the researcher/clinical setting. However, as sequencing becomes cheaper and faster, it may become feasible in the future to perform WGS and only analyse the genes of interest, storing the remainder of the data for later analysis if required, rather than sequencing a limited gene panel.

7.1.4. Other next generation sequencing applications

Whole genome sequencing does not explain the intricacies of cancer growth and metastasis, as cell biology is not purely determined by the genome sequence. For example, a single gene can lead to the production of multiple proteins via a process called alternate splicing. NGS techniques can be used to study other aspects of genomics. For example, NGS can be used to sequence RNA (often called RNA-Seq). As NGS uses DNA, this requires reverse transcription of the RNA to cDNA prior to sequencing (Wang et al., 2009). RNA-Seq provides information on the expression level (amount of RNA) of the RNA sequences and novel splice isoforms and can also detect gene fusions (Carrara et al., 2013; Edgren et al., 2011; Zhou et al., 2013).

Epigenetics is the study of changes in gene expression that do not involve changes in the DNA sequence itself (Verma and Srivastava, 2002), e.g. alterations to promoter regions (DNA regions that initiate the transcription of a particular gene) can lead to a gene being over or under-transcribed. For example, hypomethylation can be associated with gene activation or chromosomal instability, leading to chromosomal translocations (Feinberg and Tycko, 2004; Qu et al., 1999) and hypermethylation of promoter regions of tumour suppressor genes (such as RB1, MGMT and GSTP1) may lead to their inactivation (Esteller et al., 1999; Feinberg and Tycko, 2004; Greger et al., 1989; Ohtani-Fujita et al., 1993; Verma and Srivastava, 2002). Methylation can be studied using a technique called bisulphite sequencing, in which DNA is treated with sodium bisulphite. This results in the conversion of non-methylated cytosines to uracil, whereas methylated cytosines remain unchanged, enabling methylated regions to be distinguished from non-methylated regions (Laird, 2010).

Table 1
Applications of next generation sequencing.

	Description	Advantages	Disadvantages
Genome sequencing	Determines the sequence of most of the DNA from the individual's genome	• Provides information on non-coding regions and structural variants as well as coding regions	• Expensive and time consuming • Data can be more difficult to interpret • Challenges of what to do about incidental findings
Exome sequencing	Determines the sequence of the protein-coding DNA regions (exons)	• Faster and cheaper than genome sequencing • The majority of known pathological abnormalities occur in the exome • Functional consequences of variants are more easily understood	• Misses variations in non-coding regions and some structural variants • Challenges of what to do about incidental findings
Targeted panels	Determines the sequence of specific genes or parts of genes	• Usually cheaper than exome or genome sequencing, but this depends on the size of the gene panel • Focussed on particular regions of interest and so data interpretation is easier • No concern regarding incidental findings as only the regions of interest are sequenced • Can optimise the panel to capture problematic regions that are difficult to sequence using exome or genome approaches	• Does not provide information on regions outside of the gene panel

Table 2
Comparison of DNA sequencing platforms.

Sequencing platform	Sequencing	Key applications	Read length (bp)	Data output per run	Run time	Main advantages	Main disadvantages
Sanger	Capillary electrophoresis	Confirmation of NGS results, targeted re-sequencing	400–900 bp	1.9–84 kb	20 min–3 h	Long read length, high quality	High cost per Mb, low throughput
Roche-454™ GS FLX	Pyrosequencing	Genome & exome sequencing	400–1000 bp	450–700 Mb	10–23 h	Longer read lengths improve mapping in repetitive regions	High cost per Mb Homopolymer errors
Illumina HiSeq 2500™	Polymerase-mediated incorporation of fluorescent nucleotides	Genome & exome sequencing	2 × 150 bp[a] (rapid run) 2 × 125 bp[a] (high output run)	10–180 Gb (rapid run) 50–1000 Gb (high output run)	7–40 h (rapid run) #7'3#6 days (high output run)	Low cost per Mb. High throughput	Short reads. Long run time (in high output run mode)
Illumina MiSeq™	Polymerase-mediated incorporation of fluorescent nucleotides	Small genome & targeted gene panels	2 × 300 bp[a]	0.3–15 Gb	5–65 h	Short run times	Higher cost per Mb compared to HiSeq™
Life Technologies Ion PGM™	H+ ion sensitive transistor	Exome sequencing & targeted gene panels	100–200 bp	30 Mb–2 Gb	2 h	Short run times Low cost instrument	Homopolymer errors
Life Technologies Ion Proton™	H+ ion sensitive transistor	Genome & exome sequencing	100–200 bp	10 Gb	2 h	Short run times	Homopolymer errors

[a] Paired-end reads: each end of the DNA fragment is sequenced for the stated number of bases.

In addition, the structure of chromatin (a DNA-protein complex) can be altered by various processes, including histone deacetylation. Chromatin can be studied using a technique called chromatin immunoprecipitation sequencing (ChIP-Seq). ChIP-Seq involves cross-linking proteins to DNA, then shearing the DNA strands into short fragments, which are then immunoprecipitated using bead-attached antibodies against the protein of interest. The DNA and protein are then unlinked and the DNA is purified and sequenced to identify the sequence that bound to the protein (Jothi et al., 2008; Park, 2009).

Epigenetic inhibitors have been developed, e.g. azacitidine is used in chronic myelomonocytic leukaemia and acute myeloid leukaemia (Dawson and Kouzarides, 2012). There is also increasing evidence that epigenetic changes may influence prognosis and response to therapy (Dubrowinskaja et al., 2014), and therefore epigenetics is likely to become increasingly important in clinical practice.

7.2. Sequencing depth, coverage and platform performance metrics

Sequencing depth is the number of times a base pair is sequenced. For example, a depth of 30× means that the base is sequenced 30 times (see Fig. 3). However, depth may not be uniform across a DNA sequence. For example, depth can be affected by areas with high or low GC content (Clark et al., 2011; Taub et al.,

2010). Depth may also vary due to factors such as the accuracy of the chosen platform, the variant detection methods, the material being sequenced and the required sensitivity or specificity (Ajay et al., 2011; Koboldt et al., 2010; Nielsen et al., 2011; Voelkerding et al., 2009). For example, if the variant allele representation is low (e.g. due to tumour heterogeneity or contamination with normal cells), then a greater depth may be required, although this can in turn increase the false positive rate. In general, the accuracy of all second-generation platforms is similar (98–99.5%) and relies on adequate depth, although there are some systematic biases between platforms (Liu et al., 2012; Quail et al., 2012). Sequencing coverage means how much of the targeted DNA region is covered (usually at a specified depth) and a coverage of >80% is usually acceptable. For example, 80% coverage at 20× means that 80% of the base pairs in the targeted DNA region have been sequenced at least 20 times.

Sanger sequencing reactions can read DNA fragments of 400–900 base pairs in length and is used to sequence small DNA fragments. Standard Sanger sequencing can detect most mutations in the targeted region, but mutations occurring at a frequency of <20% allele frequency may not be detected. A comparison between NGS and Sanger sequencing showed that these approaches had a similar sensitivity and specificity for the detection of BRAF mutations (Ihle et al., 2014), whereas another study showed that the Roche-454 platform had superior sensitivity and specificity for the detection of KRAS mutations compared to Sanger sequencing in formalin fixed paraffin embedded (FFPE) specimens (Altimari et al., 2013).

A general overview of the advantages and disadvantages of each platform is provided in Table 2, and previous articles have compared the performance of the different technologies (Clark et al., 2011; Ihle et al., 2014; Liu et al., 2012; Quail et al., 2012). If the results are intended for clinical use, then it is also important to consider if the technology has the regulatory approvals for this purpose and whether the laboratory performing the sequencing has the required certification.

7.3. Speed and cost of sequencing

After Sanger sequencing, the platforms developed by various companies for sequencing were mainly focussed at use in the research setting. Instruments such as the HiSeq™ from Illumina are able to output large amounts of data for each run of the machine relatively cheaply. However, the effect of this is that each run takes up to 6 days and the equipment is expensive. This is not therefore optimal for use in the routine clinical setting and has led to the development of "bench-top" sequencing technology such as the Illumina MiSeq™ and the Life Technologies Ion PGM. These are cheaper to buy and are also capable of generating results within a few hours, although one consequence of this is a reduced throughput. These bench-top platforms are therefore commonly used for targeted sequencing of a limited number of genes in specific gene panels.

The number of samples that can be analysed per run of the instrument is dependent on the size of DNA to be sequenced and the depth of sequencing required. The larger the region of DNA to be sequenced and/or the greater the depth required, the fewer samples that can be sequenced per run. The cost of the run is fixed, as each run uses the same amount of reagents, regardless of the number of samples analysed. Therefore, the more samples that can be analysed simultaneously, the lower the sequencing cost per sample. This means that samples may be collected into batches until there are sufficient samples to justify a run, and this should be taken into account when deciding whether sequencing would be feasible in clinical practice. Economies of scale are therefore extremely important when calculating the cost of sequencing and the most appropriate platform. Sanger sequencing has a higher cost per base pair than NGS, but this does not necessarily mean that Sanger sequencing is more expensive than NGS. The cost per base pair reflects the cost if the sequencing platform is used to its maximum capacity. Therefore, if the DNA regions of interest are small (e.g. KRAS mutational analysis) and only a few patient samples will be analysed at any one time, then Sanger sequencing would be cheaper than NGS as the NGS platform would not be used at its full capacity.

7.4. Material available for sequencing

The quantity and quality of material available for sequencing may impact on the optimal choice of sequencing platform. Most NGS platforms have library preparations that are optimised for a specific DNA quantity and quality, and although this is not frequently a problem with fresh or frozen specimens, it can be particularly challenging with DNA derived from FFPE samples. Furthermore, the formaldehyde fixation used in preparing FFPE samples and storage of these samples at room temperature can lead to deamination artefacts and therefore false positives, although duplicating the analysis may overcome this problem (Kerick et al., 2011; Schweiger et al., 2009). In addition, the tumour sample can be contaminated with DNA from normal cells, particularly for samples with low tumour content, although macro-dissection of a population highly enriched for cancer cells can help to minimise this problem.

8. Challenges in next generation sequencing

8.1. Cost and infrastructure requirements

Although, sequencing technology has dramatically improved both in terms of the speed to obtain results and the cost of sequencing, substantial challenges remain in incorporating this technology into clinical practice. Investment is required in the equipment (including its ongoing maintenance) and reagents needed to perform the sequencing itself as well as the preparation of the samples. The costs vary depending on the instrument purchased and so it is vital to consider what type of sequencing the equipment will mainly be used for. In addition, it is important to factor in the cost of employing the staff required to perform the sequencing and interpret the data, including bioinformatics support, as well as significant investment in hardware to perform the analysis.

8.2. Data interpretation and management

There are also a number of challenges associated with the bioinformatic analysis of the raw data. NGS is not 100% accurate, as false positives and negatives occur, e.g. due to sequencing errors, mis-incorporation of bases during PCR amplification or sequence detection errors. In addition, most widely used variant callers are designed based on a diploid genome. This is not ideal for tumour samples, as copy number changes are common and therefore diploidy is not guaranteed. To try to overcome this problem, cancer-specific tools such as SNVmix have been developed (Meldrum et al., 2011). In addition, even if a variant is accurately detected its biological or clinical significance may not be established.

The data needs to be managed so that the clinically relevant information can be extracted and conveyed in an understandable way to the clinician and it is also necessary to consider how the large amount of data generated by sequencing will be stored. This is complicated by the variety of different data formats generated by the different sequencing platforms and the importance of ensuring the privacy of this data, particularly if they are transmitted over the internet. Traceability of the data is a key consideration for storage purposes, as the same sequencing data can be analysed in different

ways and it is likely to evolve with time, while the patients are still alive, and may change clinical management in the future.

8.3. Ethical considerations

There are also a number of ethical issues regarding DNA sequencing, such as who should have access to the sequencing data, which results should be returned to patients and the potential consequences for the patient and their families (McGuire et al., 2008).

Ownership of sequencing data and who has the right to access this data remains contentious. As a case in point, is it possible for an organisation to explore the data for research purposes, and if so, is this limited to the disease area (e.g. cancer) for which the patient presents? Should genetic data be available to all of the clinicians treating the patient or just to the researchers (Dressler, 2009)? Should genetic results influence non-medical issues such as patients' insurance? In addition, patients may be identifiable from their genetic data, particularly if other family members have been sequenced or they have rare variants.

Variants unrelated to the original presenting condition may be discovered incidentally and there remains debate regarding whether patients should be informed about incidental findings that may be of clinical significance, particularly if they are unaware that this sort of information might be obtained from sequencing performed for an entirely different purpose (Dressler, 2009). Various applications exist to allow patients to download and explore their own sequencing data (Illumina, 2014), but who should they turn to if they have questions about their results and what they mean for them and/or their families? If sequencing was performed for research purposes, then it is unlikely that the results have been validated in an approved clinical laboratory and it is unclear how validation should be arranged and funded if a patient requests this. When consenting patients for genetic sequencing, should all these issues be explained to the patients and their views sought on which results should be returned to them (Allen and Foulkes, 2011; Rotimi and Marshall, 2010; Lunshof et al., 2008)?

As previously discussed in Section 6.3, sequenced tumour DNA can be compared with germline DNA in order to distinguish somatic variants from germline variants. Variants may be discovered which have potential significance for a patient's family and this raises questions about whether there are any obligations to the patients' relatives (Dressler, 2009). For example, what should the clinician do if a result is of potential significance for a patient's family but the patient has either declined to be informed of the result or is deceased? These are only some of the many, as yet unresolved questions regarding sequencing that would benefit from an informed public debate.

9. Future applications of sequencing

There are numerous possible applications for the use of NGS in the treatment of patients with cancer, particularly in the research setting. However, in the immediate future, whole genome and exome sequencing are likely to be mainly used for exploratory research, with a more focussed, targeted NGS approach being used in clinical practice due to practical and cost considerations. NGS is likely to be increasingly used for screening patients for abnormalities in a panel of targetable genes and thereby identifying patients who might be suitable for targeted therapies. This NGS-based screening would also facilitate recruitment into clinical trials of targeted drugs, and there are a number of clinical trials currently investigating molecular profiling, including the NCI-MPACT trial (NCT01827384), the MOSCATO 01 trial (NCT01566019) and the MOST trial (NCT02029001). Furthermore, NGS (and other technologies such as digital PCR) can be used to identify specific mutations which can be measured in circulating tumour DNA or circulating tumour cells (Diaz and Bardelli, 2014). Testing for these mutations could potentially be used to monitor patients for signs of relapse or identify patients with residual disease as well as aiding assessment of response to treatment, although these applications would need to be validated in prospective clinical trials.

10. Conclusion

Sequencing costs are expected to continue to fall and therefore NGS is likely to become more widely utilised, particularly in clinical practice. In the research setting, the uses of NGS are manifold, including investigating mechanisms of drug resistance and identifying new drug targets. In the clinical setting, it is hoped that that NGS will facilitate familial genetic testing, better disease prognostication and the use of drugs targeted to the characteristics of individual patients. There remain many challenges to overcome, but cancer genomics is an exciting component of cancer research and treatment.

Conflicts of interest

None of the authors have any conflicts of interest.

Acknowledgement

We acknowledge support from the NIHR RM/ICR Biomedical Research Centre.

References

S.S. Ajay, S.C. Parker, H.O. Abaan, K.V. Fajardo, E.H. Margulies, Accurate and comprehensive sequencing of personal genomes, Genome Res. 21 (2011) 1498–1505.
L.B. Alexandrov, S. Nik-Zainal, D.C. Wedge, S.A. Aparicio, S. Behjati, A.V. Biankin, et al., Signatures of mutational processes in human cancer, Nature 500 (2013) 415–421.
C. Allen, W.D. Foulkes, Qualitative thematic analysis of consent forms used in cancer genome sequencing, BMC Med. Ethics 12 (2011) 14.
A. Altimari, D. de Biase, G. De Maglio, E. Gruppioni, E. Capizzi, A. Degiovanni, et al., 454 Next generation-sequencing outperforms allele-specific PCR, Sanger sequencing, and pyrosequencing for routine KRAS mutation analysis of formalin-fixed, paraffin-embedded samples, OncoTargets Ther. 6 (2013) 1057–1064.
F. Andre, T. Bachelot, F. Commo, M. Campone, M. Arnedos, V. Dieras, et al., Comparative genomic hybridisation array and DNA sequencing to direct treatment of metastatic breast cancer: a multicentre, prospective trial (SAFIR 01/UNICANCER), Lancet Oncol. (2014).
M.E. Arcila, K. Nafa, J.E. Chaft, N. Rekhtman, C. Lau, B.A. Reva, et al., EGFR exon 20 insertion mutations in lung adenocarcinomas: prevalence, molecular heterogeneity, and clinicopathologic characteristics, Mol. Cancer Ther. 12 (2013) 220–229.
D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, et al., Accurate whole human genome sequencing using reversible terminator chemistry, Nature 456 (2008) 53–59.
A.V. Biankin, N. Waddell, K.S. Kassahn, M.C. Gingras, L.B. Muthuswamy, A.L. Johns, et al., Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes, Nature 491 (2012) 399–405.
A. Borg, R.W. Haile, K.E. Malone, M. Capanu, A. Diep, T. Torngren, et al., Characterization of BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study, Hum. Mutat. 31 (2010) E1200–E1240.
Cancer Genome Atlas Network, Comprehensive molecular characterization of human colon and rectal cancer, Nature 487 (2012) 330–337.
E.R. Cantwell-Dorris, J.J. O'Leary, O.M. Sheils, BRAFV600E: implications for carcinogenesis and molecular therapy, Mol. Cancer Ther. 10 (2011) 385–394.
M. Carrara, M. Beccuti, F. Lazzarato, F. Cavallo, F. Cordero, S. Donatelli, et al., State-of-the-art fusion-finder algorithms sensitivity and specificity, BioMed. Res. Int. 2013 (2013) 340620.
P.B. Chapman, A. Hauschild, C. Robert, J.B. Haanen, P. Ascierto, J. Larkin, et al., Improved survival with vemurafenib in melanoma with BRAF V600E mutation, N. Engl. J. Med. 364 (2011) 2507–2516.
K. Cibulskis, M.S. Lawrence, S.L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, et al., Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples, Nat. Biotechnol. 31 (2013) 213–219.

M.J. Clark, R. Chen, H.Y. Lam, K.J. Karczewski, R. Chen, G. Euskirchen, et al., Performance comparison of exome DNA sequencing technologies, Nat. Biotechnol. 29 (2011) 908–914.
P.J. Cock, C.J. Fields, N. Goto, M.L. Heuer, P.M. Rice, The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants, Nucleic Acids Res. 38 (2010) 1767–1771.
S. Cottrell, D. Bicknell, L. Kaklamanis, W.F. Bodmer, Molecular analysis of APC mutations in familial adenomatous polyposis and sporadic colon carcinomas, Lancet 340 (1992) 626–630.
D.C. Crawford, D.A. Nickerson, Definition and clinical importance of haplotypes, Annu. Rev. Med. 56 (2005) 303–320.
A. Crockford, M. Jamal-Hanjani, J. Hicks, C. Swanton, Implications of intratumour heterogeneity for treatment stratification, J. Pathol. 232 (2014) 264–273.
M.A. Dawson, T. Kouzarides, Cancer epigenetics: from mechanism to therapy, Cell 150 (2012) 12–27.
M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data, Nat. Genet. 43 (2011) 491–498.
L.A. Diaz Jr., A. Bardelli, Liquid biopsies: genotyping circulating tumor DNA, J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol. 32 (2014) 579–586.
R. Dienstmann, D. Serpico, J. Rodon, C. Saura, T. Macarulla, E. Elez, et al., Molecular profiling of patients with colorectal cancer and matched targeted therapy in phase I clinical trials, Mol. Cancer Ther. 11 (2012) 2062–2071.
J.Y. Douillard, K.S. Oliner, S. Siena, J. Tabernero, R. Burkes, M. Barugel, et al., Panitumumab-FOLFOX4 treatment and RAS mutations in colorectal cancer, N. Engl. J. Med. 369 (2013) 1023–1034.
L.G. Dressler, Disclosure of research results from cancer genomic studies: state of the science, Clin. Cancer Res. 15 (2009) 4270–4276.
D. Dressman, H. Yan, G. Traverso, K.W. Kinzler, B. Vogelstein, Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 8817–8822.
N. Dubrowinskaja, K. Gebauer, I. Peters, J. Hennenlotter, M. Abbas, R. Scherer, et al., Neurofilament heavy polypeptide CpG island methylation associates with prognosis of renal cell carcinoma and prediction of antivascular endothelial growth factor therapy response, Cancer Med. (2014).
A.M. Dulak, P. Stojanov, S. Peng, M.S. Lawrence, C. Fox, C. Stewart, et al., Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity, Nat. Genet. 45 (2013) 478–486.
H. Edgren, A. Murumagi, S. Kangaspeska, D. Nicorici, V. Hongisto, K. Kleivi, et al., Identification of fusion genes in breast cancer by paired-end RNA-sequencing, Genome Biol. 12 (2011) R6.
J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, et al., Real-time DNA sequencing from single polymerase molecules, Science 323 (2009) 133–138.
H.C. Erichsen, S.J. Chanock, SNPs in cancer research and treatment, Br. J. Cancer 90 (2004) 747–751.
M. Esteller, S.R. Hamilton, P.C. Burger, S.B. Baylin, J.G. Herman, Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is a common event in primary human neoplasia, Cancer Res. 59 (1999) 793–797.
A.P. Feinberg, B. Tycko, The history of cancer epigenetics, Nat. Rev. Cancer 4 (2004) 143–153.
D. Ford, D.F. Easton, M. Stratton, S. Narod, D. Goldgar, P. Devilee, et al., Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The breast cancer linkage consortium, Am. J. Hum. Genet. 62 (1998) 676–689.
D. Gajria, S. Chandarlapaty, HER2-amplified breast cancer: mechanisms of trastuzumab resistance and novel targeted therapies, Expert Rev. Anticancer Ther. 11 (2011) 263–275.
M. Gerlinger, A.J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, et al., Intratumor heterogeneity and branched evolution revealed by multiregion sequencing, N. Engl. J. Med. 366 (2012) 883–892.
V. Greger, E. Passarge, W. Hopping, E. Messmer, B. Horsthemke, Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma, Hum. Genet. 83 (1989) 155–158.
R.R. Gullapalli, K.V. Desai, L. Santana-Santos, J.A. Kant, M.J. Becich, Next generation sequencing in clinical medicine: challenges and lessons for pathology and biomedical informatics, J. Pathol. Inf. 3 (2012) 40.
E.C. Hayden, Technology: the $1000 genome, Nature 507 (2014) 294–295.
M.A. Ihle, J. Fassunke, K. Konig, I. Grunewald, M. Schlaak, N. Kreuzberg, et al., Comparison of high resolution melting analysis, pyrosequencing, next generation sequencing and immunohistochemistry to conventional Sanger sequencing for the detection of p. V600E and non-p. V600E BRAF mutations, BMC Cancer 14 (2014) 13.
Illumina, 2015. Illumina VariantStudio. <http://wwwilluminacom/informatics/research/biological-data-interpretation/variantstudiohtml>.
Illumina, 2014. MyGenome App. <http://wwwilluminacom/clinical/clinical_informatics/mygenome_appilmn>.
R. Jothi, S. Cuddapah, A. Barski, K. Cui, K. Zhao, Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data, Nucleic Acids Res. 36 (2008) 5221–5231.
C. Kandoth, M.D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, et al., Mutational landscape and significance across 12 major cancer types, Nature 502 (2013) 333–339.
R. Karchin, Next generation tools for the annotation of human SNPs, Briefings Bioinf. 10 (2009) 35–52.
R. Katayama, A.T. Shaw, T.M. Khan, M. Mino-Kenudson, B.J. Solomon, B. Halmos, et al., Mechanisms of acquired crizotinib resistance in ALK-rearranged lung cancers, Sci. Trans. Med. (2012) a17.
M. Kerick, M. Isau, B. Timmermann, H. Sultmann, R. Herwig, S. Krobitsch, et al., Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity, BMC Med. Genomics 4 (2011) 68.
D.C. Koboldt, L. Ding, E.R. Mardis, R.K. Wilson, Challenges of sequencing human genomes, Briefings Bioinf. 11 (2010) 484–498.
J.P. Koivunen, C. Mermel, K. Zejnullahu, C. Murphy, E. Lifshits, A.J. Holmes, et al., EML4-ALK fusion gene and efficacy of an ALK kinase inhibitor in lung cancer, Clin. Cancer Res. 14 (2008) 4275–4283.
A. Krohn, F. Freudenthaler, S. Harasimowicz, M. Kluth, S. Fuchs, L. Burkhardt, et al., Heterogeneity and chronology of PTEN deletion and ERG fusion in prostate cancer, Mod. Pathol.: Off. J. U. S. Can. Acad. Pathol. Inc. (2014).
E.L. Kwak, Y.J. Bang, D.R. Camidge, A.T. Shaw, B. Solomon, R.G. Maki, et al., Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer, N. Engl. J. Med. 363 (2010) 1693–1703.
P.W. Laird, Principles and challenges of genomewide DNA methylation analysis, Nat. Rev. Genetics 11 (2010) 191–203.
E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, et al., Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25.
H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (2009) 1754–1760.
H. Li, J. Ruan, R. Durbin, Mapping short DNA sequencing reads and calling variants using mapping quality scores, Genome Res. 111 (2008).
H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al., The sequence alignment/map format and SAM tools, Bioinformatics 25 (2009) 2078–2079.
Life Technologies Corporation, 2015. Torrent Browser Analysis Report Guide <http://mendeliontorrentcom/ion-docs/Torrent-Variant-Caller-Pluginhtml>.
L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, et al., Comparison of next-generation sequencing systems, J. Biomed. Biotechnol. 2012 (2012) 251364.
J.E. Lunshof, R. Chadwick, D.B. Vorhaus, G.M. Church, From genetic privacy to open consent, Nat. Rev. Genet. 9 (2008) 406–411.
E.R. Mardis, A decade's perspective on DNA sequencing technology, Nature 470 (2011) 198–203.
M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, et al., Genome sequencing in microfabricated high-density picolitre reactors, Nature 437 (2005) 376–380.
A.L. McGuire, T. Caulfield, M.K. Cho, Research ethics and the challenge of whole-genome sequencing, Nat. Rev. Genet. 9 (2008) 152–156.
W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, F. Cunningham, Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor, Bioinformatics 26 (2010) 2069–2070.
C. Meldrum, M.A. Doyle, R.W. Tothill, Next-generation sequencing for cancer diagnostics: a practical perspective, Clin. Biochem. Rev.: Aust. Assoc. Clin. Biochem. 32 (2011) 177–195.
F. Meric-Bernstam, G.B. Mills, Overcoming implementation challenges of personalized cancer therapy, Nat. Rev. Clin. Oncol. 9 (2012) 542–548.
T. Mitsudomi, Y. Yatabe, Epidermal growth factor receptor in relation to tumor development: EGFR gene and cancer, FEBS J. 277 (2010) 301–308.
M. Morey, A. Fernandez-Marmiesse, D. Castineiras, J.M. Fraga, M.L. Couce, J.A. Cocho, A glimpse into past, present, and future DNA sequencing, Mol. Genet. Metab. 110 (2013) 3–24.
National Human Genome Research Institute, 2013. DNA Sequencing Costs. <http://wwwgenomegov/sequencingcosts/>.
R. Nielsen, J.S. Paul, A. Albrechtsen, Y.S. Song, Genotype and SNP calling from next-generation sequencing data, Nat. Rev. Genet. 12 (2011) 443–451.
Novoalign, 2014. <http://www.novocraft.com>.
N. Ohtani-Fujita, T. Fujita, A. Aoike, N.E. Osifchin, P.D. Robbins, T. Sakai, CpG methylation inactivates the promoter activity of the human retinoblastoma tumor-suppressor gene, Oncogene 8 (1993) 1063–1067.
M. Olivier, M. Hollstein, P. Hainaut, TP53 mutations in human cancers: origins, consequences, and clinical use, Cold Spring Harbor Perspect. Biol. 2 (2010) a001008.
P.J. Park, ChIP-seq: advantages and challenges of a maturing technology, Nat. Rev. Genet. 10 (2009) 669–680.
D.W. Parsons, S. Jones, X. Zhang, J.C. Lin, R.J. Leary, P. Angenendt, et al., An integrated genomic analysis of human glioblastoma multiforme, Science 321 (2008) 1807–1812.
J. Perkel, Sequence analysis 10: a newbie's guide to crunching next-generation sequencing data, Scientist (2011).
G.Z. Qu, P.E. Grundy, A. Narayan, M. Ehrlich, Frequent hypomethylation in Wilms tumors of pericentromeric DNA in chromosomes 1 and 16, Cancer Genet. Cytogenet. 10 (1999) 34–39.
M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, et al., A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, BMC Genomics 13 (2012) 341.
J.T. Robinson, H. Thorvaldsdottir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, et al., Integrative genomics viewer, Nat. Biotechnol. 29 (2011) 24–26.
M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlen, P. Nyren, Real-time DNA sequencing using detection of pyrophosphate release, Anal. Biochem. 242 (1996) 84–89.

M. Ronaghi, M. Uhlen, P. Nyren, A sequencing method based on real-time pyrophosphate, Science 281 (363) (1998) 5.
J.M. Rothberg, W. Hinz, T.M. Rearick, J. Schultz, W. Mileski, M. Davey, et al., An integrated semiconductor device enabling non-optical genome sequencing, Nature 475 (2011) 348–352.
C.N. Rotimi, P.A. Marshall, Tailoring the process of informed consent in genetic and genomic research, Genome Med. 2 (2010) 20.
A.J. Rowan, H. Lamlum, M. Ilyas, J. Wheeler, J. Straub, A. Papadopoulou, et al., APC mutations in sporadic colorectal tumors: a mutational hotspot and interdependence of the two hits, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 3352–3357.
S.M. Rumble, P. Lacroute, A.V. Dalca, M. Fiume, A. Sidow, M. Brudno, SHRiMP accurate mapping of short color-space reads, PLoS Comput. Biol. 5 (2009) e1000386.
F. Sanger, G.M. Air, B.G. Barrell, N.L. Brown, A.R. Coulson, C.A. Fiddes, et al., Nucleotide sequence of bacteriophage phi X174 DNA, Nature 265 (1977a) 687–695.
F. Sanger, S. Nicklen, A.R. Coulson, DNA sequencing with chain-terminating inhibitors, Proc. Natl. Acad. Sci. U. S. A. (1977b) 5463–5467.
E.E. Schadt, S. Turner, A. Kasarskis, A window into third-generation sequencing, Hum. Mol. Genet. 19 (2010) R227–R240.
M.R. Schweiger, M. Kerick, B. Timmermann, M.W. Albrecht, T. Borodina, D. Parkhomchuk, et al., Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis, PloS One 4 (2009) e5548.
F. Sclafani, G. Gullo, K. Sheahan, J. Crown, BRAF mutations in melanoma and colorectal cancer: a single oncogenic mutation with different tumour phenotypes and clinical implications, Crit. Rev. Oncol. Hematol. 87 (2013) 55–68.
A.T. Shaw, D.W. Kim, K. Nakagawa, T. Seto, L. Crino, M.J. Ahn, et al., Crizotinib versus chemotherapy in advanced ALK-positive lung cancer, N. Engl. J. Med. 368 (2013) 2385–2394.
J. Shendure, G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, et al., Accurate multiplex polony sequencing of an evolved bacterial genome, Science 309 (2005) 1728–1732.
D.J. Slamon, G.M. Clark, S.G. Wong, W.J. Levin, A. Ullrich, W.L. McGuire, Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene, Science 235 (1987) 177–182.
G. Smith, F.A. Carey, J. Beattie, M.J. Wilkie, T.J. Lightfoot, J. Coxhead, et al., Mutations in APC, kirsten-ras, and p53-alternative genetic pathways to colorectal cancer, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 9433–9438.
Strand Scientific Intelligence Inc., 2013. Guide to Storage and Computation Requirements.
H. Swerdlow, R. Gesteland, Capillary gel electrophoresis for rapid, high resolution DNA sequencing, Nucleic Acids Res. 18 (1990) 1415–1419.
R. Tan, Y. Wang, S.E. Kleinstein, Y. Liu, X. Zhu, H. Guo, et al., An evaluation of copy number variation detection tools from whole-exome sequencing data, Hum. Mutat. 35 (2014) 899–907.
M.A. Taub, H. Corrada Bravo, R.A. Irizarry, Overcoming bias and systematic errors in next generation sequencing data, Genome Med. 2 (2010) 87.
M. Tennis, S. Krishnan, M. Bonner, C.B. Ambrosone, J.E. Vena, K. Moysich, et al., p53 mutation analysis in breast tumors by a DNA microarray method. Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive, Oncology 15 (2006) 80–85.
The Translational Research and Personalised Medicine Working Group, ESMO Glossary in Molecular Biology of Cancer and Molecular Techniques, ESMO Press, Switzerland, 2015.
K. Trunzer, A.C. Pavlick, L. Schuchter, R. Gonzalez, G.A. McArthur, T.E. Hutson, et al., Pharmacodynamic effects and mechanisms of resistance to vemurafenib in patients with metastatic melanoma, J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol. 31 (2013) 1767–1774.
J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, et al., The sequence of the human genome, Science 291 (2001) 1304–1351.
M. Verma, S. Srivastava, Epigenetics in cancer: implications for early detection and prevention, Lancet Oncol. 3 (2002) 755–763.
K.V. Voelkerding, S.A. Dames, J.D. Durtschi, Next-generation sequencing: from basic research to diagnostics, Clin. Chem. 55 (2009) 641–658.
B.A. Walker, C.P. Wardell, L. Melchor, S. Hulkki, N.E. Potter, D.C. Johnson, et al., Intraclonal heterogeneity and distinct molecular mechanisms characterize the development of t(4;14) and t(11;14) myeloma, Blood 120 (2012) 1077–1086.
Z. Wang, M. Gerstein, M. Snyder, RNA-seq. A revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.
I.B. Weinstein, A.K. Joe, Mechanisms of disease: oncogene addiction – a rationale for molecular targeting in cancer therapy, Nat. Clin. Pract. Oncol. 3 (2006) 448–457.
T.A. Yap, M. Gerlinger, P.A. Futreal, L. Pusztai, C. Swanton, Intratumor heterogeneity: seeing the wood for the trees, Sci. Trans. Med. 4 (2012), 127ps10.
T.I. Zack, S.E. Schumacher, S.L. Carter, A.D. Cherniack, G. Saksena, B. Tabak, et al., Pan-cancer patterns of somatic copy number alteration, Nat. Genet. 45 (2013) 1134–1140.
J.B. Zhou, T. Zhang, B.F. Wang, H.Z. Gao, X. Xu, Identification of a novel gene fusion RNF213SLC26A11 in chronic myeloid leukemia by RNA-seq, Mol. Med. Rep. 7 (2013) 591–597.
A. Ziegler, D.J. Leffell, S. Kunala, H.W. Sharma, M. Gailani, J.A. Simon, et al., Mutation hotspots due to sunlight in the p53 gene of nonmelanoma skin cancers, Proc. Natl. Acad. Sci. U. S. A. 90 (1993) 4216–4220.

Biographies

Sing Yu Moorcraft is a clinical research fellow in Medical Oncology in the Gastrointestinal and Lymphoma Unit at the Royal Marsden NHS Foundation Trust. Her research focuses on the feasibility of using next generation sequencing to identify patients who might be suitable for targeted therapies.

David Gonzalez is a consultant clinical scientist and head of the Molecular Diagnostics department at the Centre for Molecular Pathology in the Royal Marsden NHS Foundation Trust. His work focuses on the molecular characterisation of tumours for clinical diagnostics and translational research purposes. He is a principal investigator in the CR-UK Stratified Medicine Programme, one of the largest UK initiatives to bring personalised therapies to cancer patients.

Brian A. Walker is a senior research scientist within the Molecular Diagnostics department at the Centre for Molecular Pathology in the Royal Marsden NHS Foundation Trust. He is responsible for the development of innovative and time saving tools in personalised molecular medicine. He has previously worked at the Institute of Cancer Research, specialising in Multiple Myeloma, where he used cutting edge research tools to identify novel genetic markers to identify patients with a poor prognosis.