Review

Pathobiology

Pathobiology 2017;84:292–305
DOI: 10.1159/000480089

Received: April 12, 2017
Accepted: August 6, 2017
Published online: October 31, 2017

Next-Generation Sequencing in Diagnostic Pathology

Mohammad Ilyas

Academic Unit of Pathology and Nottingham Molecular Pathology Node, Division of Cancer Stem Cells, School of Medicine, University of Nottingham, Queen's Medical Centre, Nottingham University, Nottingham, UK

Keywords
Next-generation sequencing · Diagnostic pathology

Abstract
Interrogation of tissue informs on patient management through delivery of a diagnosis together with associated clinically relevant data. The diagnostic pathologist will usually evaluate the morphological appearances of a tissue sample and, occasionally, the pattern of expression of a limited number of biomarkers. Recent developments in sequencing technology mean that DNA and RNA from tissue samples can now be interrogated in great detail. These new technologies, collectively known as next-generation sequencing (NGS), generate huge amounts of data which can be used to support patient management. In order to maximize the utility of tissue interrogation, the molecular data need to be interpreted and integrated with the morphological data. However, in order to interpret the molecular data, the pathologist must understand the utility and the limitations of NGS data. In this review, the principles behind NGS technologies are described. In addition, the caveats in the interpretation of the data are discussed, and a scheme is presented to "classify" the types of data which are generated. Finally, a glossary of new terminology is included to help pathologists become familiar with the lexicon of NGS-derived molecular data.

Introduction

Histopathology involves direct visual interrogation of diseased tissue in order to generate data which can be used to inform on patient management. The main test performed on the tissue is application of the 2 stains, i.e. haematoxylin and eosin. This incredibly simple procedure is followed by sophisticated interpretation of the morphological features of the tissue to yield information on the underlying diagnosis as well as information on the likely behaviour and the prognosis of the disease.

The use of haematoxylin-eosin staining has been around for over 100 years [1], and it provides a summative picture of the events which are occurring in the tissue. It does not however provide any detail on the individual processes which are contributing to the pathological change. A number of adjunctive tests have been developed over the years in order to identify the specific processes occurring within tissue samples. These tests range from simple histochemistry (for example, tests for specific mucins) to immunohistochemistry (testing for specific proteins) and, more recently, to molecular testing of nucleic acids. However, spectacular developments in sequencing technology mean that changes in the genome and transcriptome can be viewed in unprecedented detail. The sheer scale of these developments is demonstrated by the changes in cost for sequencing the human ge-

KARGER

E-Mail karger@karger.com
www.karger.com/pat

Mohammad Ilyas
Academic Unit of Pathology and Nottingham Molecular Pathology Node
Division of Cancer Stem Cells, School of Medicine, University of Nottingham
Queen's Medical Centre, Nottingham University, Nottingham NG7 2UH (UK)
E-Mail mohammad.ilyas@nottingham.ac.uk

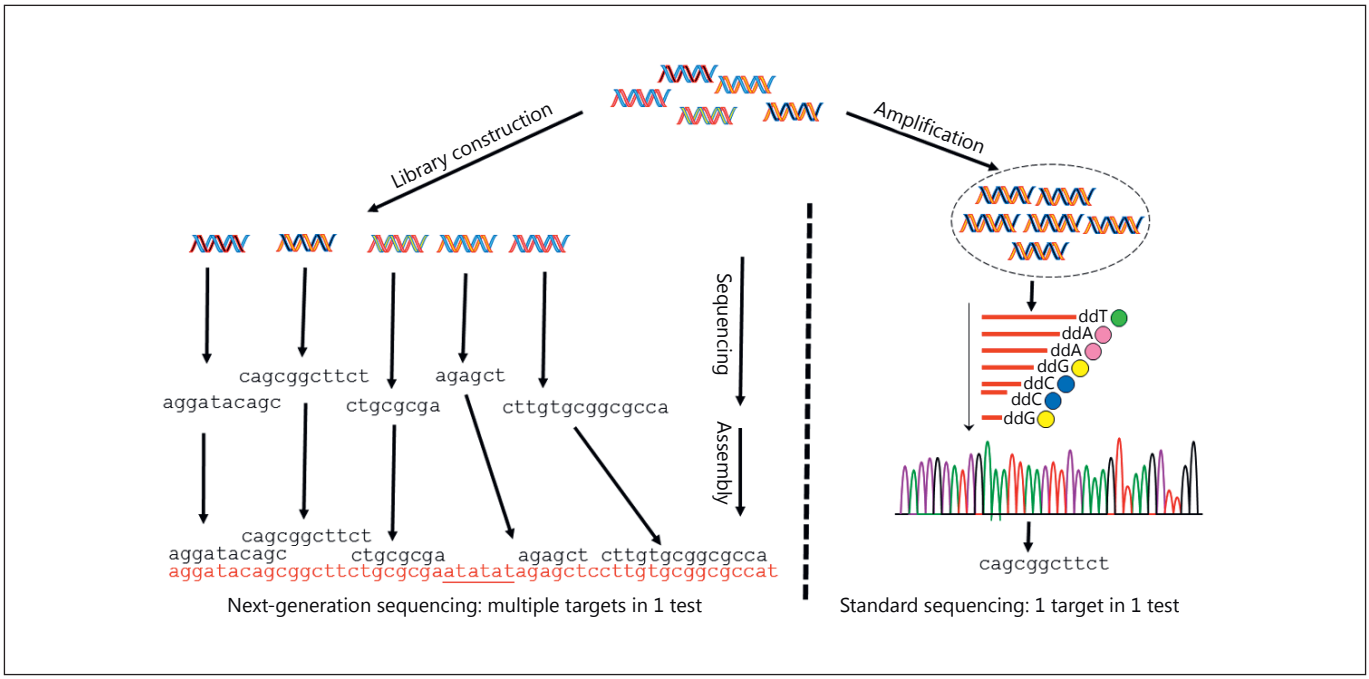

Fig. 1. The difference between NGS and standard sequencing. Any template may contain several targets of interest. In standard sequencing, 1 target is amplified in each test (by cloning or PCR), and the net signal from sequencing all the amplified molecules is taken for base calling. In NGS, multiple targets are interrogated in 1 test (5 targets shown here), and firstly a library containing the targets of interest is created. Individual molecules then undergo sequencing and are then compared against a reference sequence (shown here in red). The random nature of the library construction for non-targeted sequencing means that the library may contain different size fragments, and these may overlap. The underlined sequence in the reference AT-rich regions serves to indicate that reproducible biases can occur in the library construction so that some regions (such as GC- or AT-rich regions) may be less well represented. Even if only a single target were to be tested by NGS, the sequence of individual molecules would still be collected.

nome: assembly of the first human reference genome, using fluorescent Sanger sequencing, took approximately 13 years and cost around USD 3 billion [2]. Today, using next-generation sequencing (NGS) technologies, the same data could be captured in *under 2 weeks* and would cost under *USD 1,000*. These developments are therefore paving the way for a transformation of diagnostic pathology in which NGS-based tests will become a routine part of tissue interrogation.

NGS versus Standard Sequencing Technologies

NGS refers to a group of technologies which have, in common, the ability to perform and capture data from millions of sequencing reactions simultaneously – also called *massively parallel sequencing* [3–6]. Although the various NGS platforms differ in the way they acquire data, they are all able to capture the *individual sequence* of hundreds of millions of molecules. This is in contrast to standard sequencing technology (such as Sanger sequencing) in which the net signal derived from a pool of molecules is captured, thus giving a *collective sequence* (Fig. 1). NGS has a number of advantages, the foremost of which is the ability to sequence multiple targets in 1 reaction as opposed to the "1 target per reaction" limitation of standard technology. In addition, since the sequence of each molecule can be checked individually, low-frequency allelic variants can be identified rather than being lost in the net signal generated by the majority allele population. The limit of detection (i.e., the proportion of variant alleles which must be present in order to be detected) is variable for standard sequencing technologies and ranges from 5% for pyrosequencing to 20% for fluorescence-based Sanger sequencing [7]. It can however be enhanced through modifications such as COLD-PCR [8, 9]. The limit of detection of NGS will depend on the depth of coverage (see later) but it can reach well below 1% [10]. Having a low limit of detection gives greater flexibility in certain situations such as when the proportion of tumour

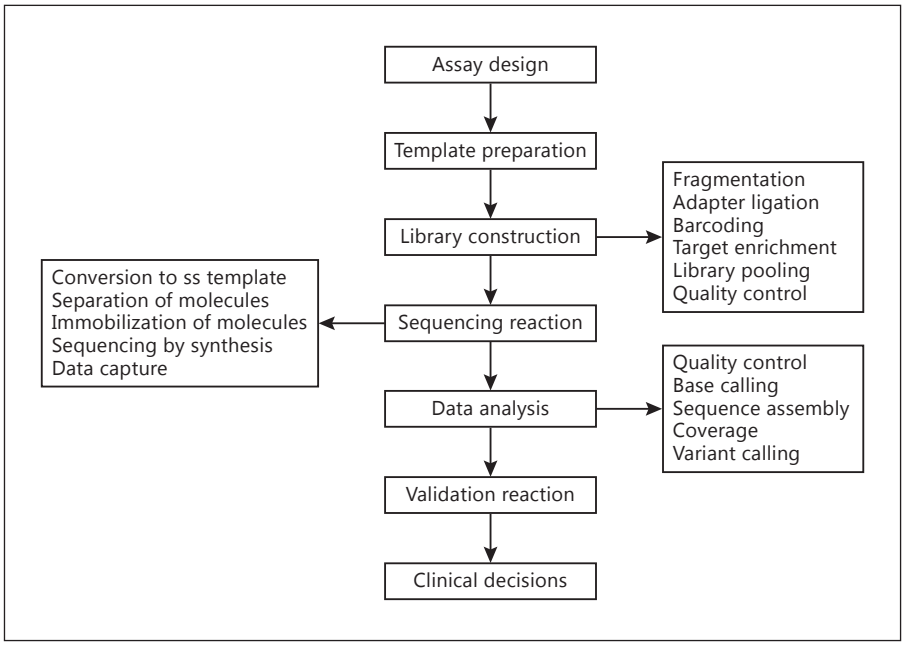

Fig. 2. NGS workflow. This shows the standard workflow for an NGS assay. There are numerous steps involved which require technical precision. The steps may vary depending on the type of assay and the information required.

cells in a tissue sample is low or when genetic heterogeneity may be an important consideration (such as predictive testing for treatment decisions).

What Targets Can Be Tested by NGS?

The genome and transcriptome can be examined at several different levels with NGS, and the targets chosen for sequencing should be appropriate for the underlying question. The human genome comprises approximately 3×10^9 bases which are organized into coding regions (exons) and non-coding regions (introns, promoters, regulatory elements, and structural elements). Sequencing all of these elements together is known as a *whole-genome sequencing (WGS)*.

WGS may not always be the most appropriate test to perform, and in many cases, where for example information about regulatory elements is not required, sequencing of selective parts of the gene can be performed. If one is only interested in the coding regions, then the most appropriate test would be sequencing only the exons of the known genes. This is known as *whole-exome sequencing (WES)* and would require sequencing only of approximately 2% of the genome [11]. This would be cheaper and would allow sequencing to be performed in greater depth. In other cases, information will be required only for a limited number of genes (or indeed the hotspots in those genes) and this is known as *targeted sequencing*.

NGS can also be applied to perform genome-wide analysis of specific modifications such as DNA methylation (known as Methyl-Seq) or DNA-protein interactions (such as histone modification or transcription factor binding, known as chromatin immunoprecipitation sequencing, ChIP-Seq).

RNA can be sequenced, and this is known as RNA-Seq. This would include all RNA species including mRNA, microRNA, ribosomal RNA, etc. [12–15].

Not all of the targets which can be sequenced by NGS have clinical utility, and some may remain permanently in the research arena. It is important to note, however, that if data emerge showing clinical value for specific targets (such as transcription factor binding sites), these can be quickly adopted into clinical practice.

The NGS Workflow

Assays must be designed in light of the clinical and laboratory requirements. The appropriate template can then be obtained, and testing can begin. The NGS workflow comprises 3 steps, i.e. library construction, sequencing reaction, and data analysis. These are considered next (Fig. 2–4).

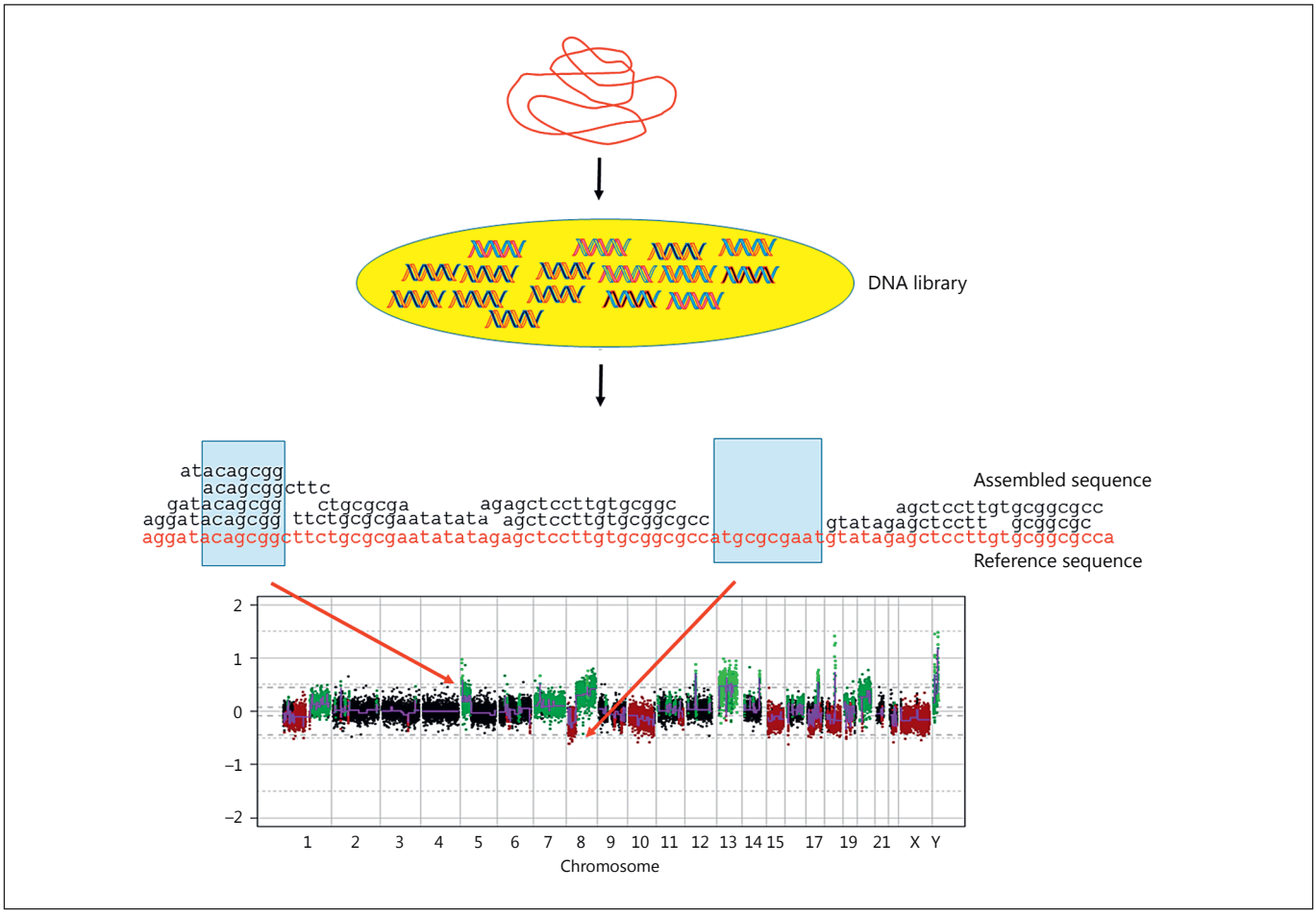

Fig. 3. Whole-genome sequencing and depth of coverage. Together with Figure 4, this shows how different types of assay can produce different results on the same material. In this example, the template (shown in red) is derived from a tumour which has amplifications, deletions, and point mutations. The DNA library is prepared and undergoes whole-genome sequencing. Assuming no bias, the whole genome will be sequenced, and library fragments from each site will be sequenced. However, a finite number of sequencing reads are produced and, in this example, only 10 sequencing reads are shown. Depth of sequencing refers to the number of times any base is sequenced, and it can be seen from this example that there is an average depth ×2. Areas of amplification and deletion (highlighted in blue) will be over- and underrepresented compared to the average, and thus copy number variations can be identified. If there is a point mutation which is present at a low frequency (for example 20%), it is likely to be missed.

Library Construction

Once the target regions have been decided, the next step is processing the template into a format which is suitable for sequencing – known as library construction. The importance of template quality is discussed later but it is well known that poor-quality template will result in greater numbers of sequencing errors. Given the huge numbers of sequencing reactions being performed simultaneously, it is easy to see that even minor increases in error rate could produce large amounts of unreliable data.

Library construction protocols will vary to some degree depending on the type of NGS platform to be deployed but in essence library construction involves 2 main steps: (i) fragmentation of the template into a suitable size (usually 200–500 bp) and (ii) attachment of specially modified DNA adapters to allow the sequencing reaction to occur and identify the origin of the sample (i.e. a sequence barcode). The barcoding will allow multiple samples to be pooled together and sequenced simultaneously. Prior to the sequencing reaction, the library should undergo both precise quantification and quality control.

Fragmentation can be performed using a variety of methods including physical methods (such as sonication,

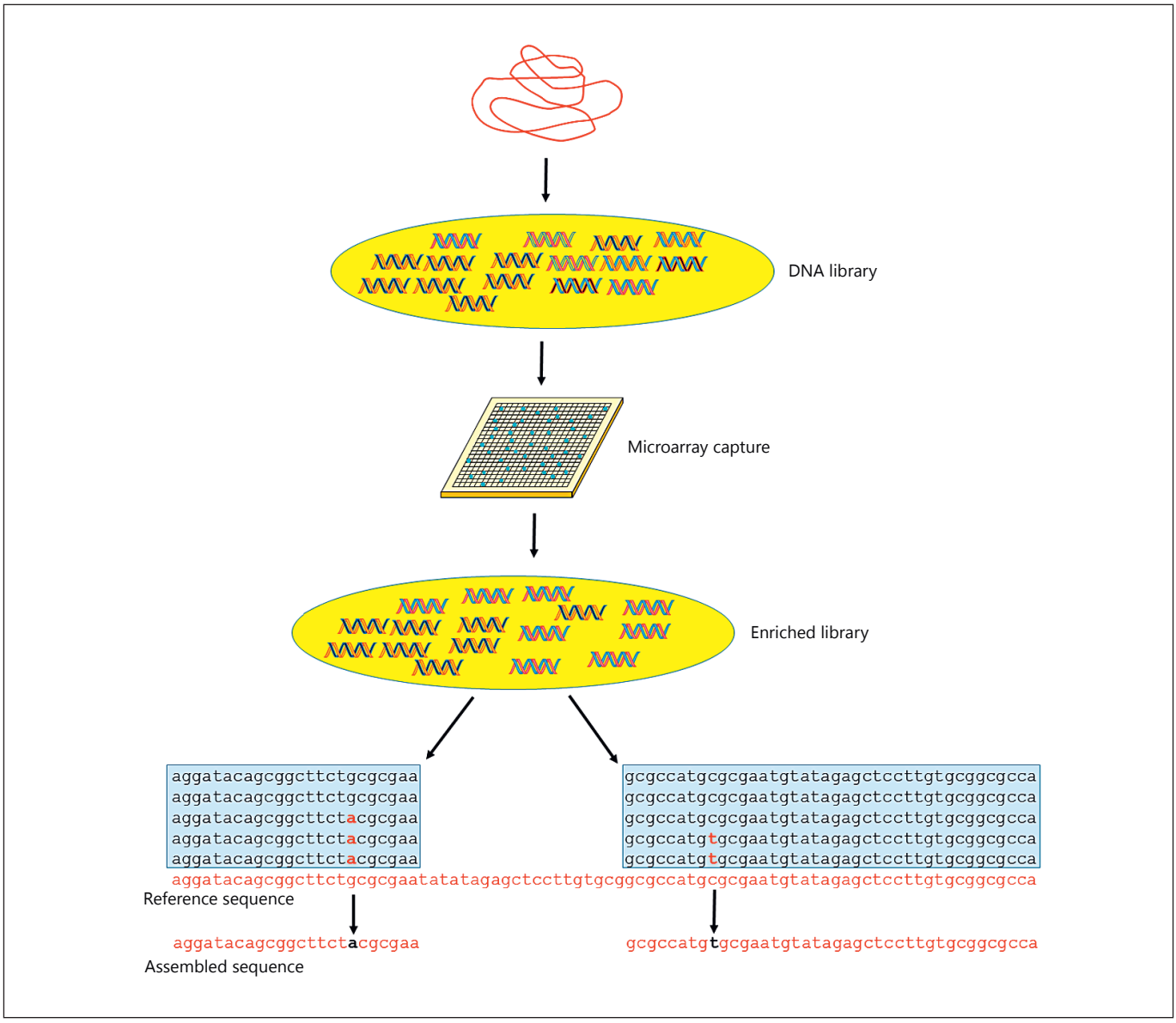

Fig. 4. Whole-exome sequencing and depth of coverage. Together with Figure 3, this shows how different types of assay can produce different results on the same material. In this example, the template (shown in red) is derived from a tumour which has amplifications, deletions, and point mutations. The DNA is fragmented and enriched for exonic sequences (representing approx. 2% of the DNA and in this case captured on a chip). The sample undergoes whole-exome sequencing, and although the total number of reads will be the same as for whole-genome sequencing, each target base will be read at greater depth. In this example, there are a total of 10 sequence reads but only 2 exons, and therefore each base will have a sequencing depth of ×5. This allows low-frequency mutations to be identified (down to a limit of detection of 20%, i.e. 1 mutant sequence in the 5 sequences that are read). However, inferences about structural changes (copy number changes, translocations, etc.) are far more difficult to make.

acoustic shearing), enzymatic methods (such as endonucleases) and, for single-stranded RNA, chemical methods (heat with divalent ions). The required size of the fragment – also known as the insert (as it is inserted between adapters) – will depend on the sequencing platform [16, 17]. The various fragmentation methods have similar efficiencies although the enzymatic methods are more prone to introducing insertion-deletion artefacts [18]. The fragmented DNA is then blunt ended and phosphorylated in order to allow the ligation of adapters. The

adapters contain sequences to allow clonal PCR amplification during the sequencing reaction.

For WES, the principles of library construction still apply but a step of enrichment to pull out the coding sequences is required. This can be done by PCR using exon-specific primers but is usually done by using exon-specific hybridization probes [19]. This can be done using a microarray whereby the library is hybridized to an expression array. Alternatively, the library can be hybridized in solution with exon-specific probes which have been tagged to allow the exonic regions to be pulled out. Targeted NGS is a scaled-down version of WES, and PCR for the specific targets forms the enrichment step. Since, for these applications, the total amount of sequencing to be performed is much less than WGS, barcodes for sample identification can be added to adapters by PCR. Barcoded libraries can then be pooled to allow multiple samples to be tested in 1 reaction.

In some specialist applications the DNA may need modification prior to library construction. Examples would include methylation profiling using Methyl-Seq in which DNA will need to be bisulphite modified. Cross-linking of DNA and bound protein will be required for ChIP-Seq and then, after fragmentation, the protein-bound DNA fragments can be pulled out by immunoprecipitation with the appropriate antibodies.

Once the library has been constructed it needs to be cleaned up to remove self-ligated adapters and inappropriately sized fragments. This can be done by magnetic beads or purification from agarose gel. Quantification and quality assessment of the library are important steps and can be performed by quantitative PCR. This will allow an assessment of the amplifiability of the library and also, when using a bioanalyser, the range of insert size.

Sequencing Reaction

The first step of the sequencing reaction is to convert the library into single-stranded DNA and isolate individual molecules in order for them to be sequenced. The signal emitted from sequencing a single molecule will not be detectable using currently available chemistries, and thus each individual molecule must be "clonally amplified" so that sufficient signal is obtained. This involves immobilization of the single-stranded molecule and local PCR amplification of that molecule. The 2 most common methods of clonal amplification are "bridge amplification" and "bead amplification" [5, 6, 19].

Bridge amplification is proprietary technology of the Illumina platform. The library is poured into a flow cell which is covered with a "lawn" of oligonucleotides attached to the cell. These oligonucleotides are complementary to sequences within the adapters and therefore individual molecules in the library can be immobilized on the flow cell. Once immobilized, local PCR results in generation of what is known as a "cluster." The sequencing is then performed, and the sequence from each cluster (derived from a single molecule) can be captured.

For bead amplification, beads and the library are mixed together in a water-in-oil emulsion [20]. Oligonucleotides are attached to the bead which immobilize the DNA molecule and which allow PCR to be performed. The quantities of the beads and library are adjusted so that each droplet contains 1 DNA molecule and 1 bead. Each bead will therefore contain clonally amplified molecules, and it can then be placed in a well from which the sequence can be captured.

Both methodologies require precise quantification using specialized equipment. Excess DNA in the libraries may lead, for example, to a high cluster density with signals from one cluster bleeding into adjacent clusters or to more than 1 molecule becoming attached to a bead, thus leading to a mixed sequence.

Sequencing of the clonally amplified molecules is based on the principle of synthesizing a new complementary strand onto a single-stranded template. The sequence is captured as the bases are incorporated into the new strand, and this is known as "sequencing by synthesis." A variety of chemistries are used in the different platforms for base calling and include fluorescently tagged bases (such as with the Illumina platform) or monitoring the change in pH that occurs every time a base is incorporated into a newly synthesized strand (such as the Ion Torrent platform) [21]. Quantifying pyrophosphate release during DNA synthesis (as used in pyrosequencing) is less commonly used [22]. Some platforms will sequence in both directions (known as paired-end sequencing) whilst others will have protocols which sequence both strands in a library.

Data Analysis

A huge amount of data is generated by NGS, and interpreting the data correctly is a major challenge. A variety of different software packages are available which enable the data to be analysed. Some of these are produced by the manufacturers themselves whilst others are third-party packages. The primary output of the data is usually in the form of a FASTQ file which contains the raw sequence and information about the quality of the sequence. Information about sequence quality is denoted by the "Phred" score which is allocated to each base. It is derived

Table 1. Commonly used NGS-based assays

Assay type	Template	Utility
Whole-genome sequencing (WGS)	DNA: coding (exons) and non-coding regions (introns, regulatory and structural regions)	Generates huge amounts of data per sample but usually low depth of coverage Useful for CNV and SV although less useful for SNV and small insertion/deletion (indels)
Whole-exome sequencing (WES)	DNA: coding regions	Sequence from 2% of the genome and usually good depth of coverage Useful for SNVs and indels but less useful for CNV/SV
Targeted sequencing	DNA: specifically selected regions	Usually very deep sequencing Very good for SNVs and indels and for identifying minority clones in heterogeneous samples; cannot be used for CNV/SV
RNA-Seq	Nascent and mature mRNA Non-coding RNA – miRNA, lncRNA, snoRNA, rRNA footprinting	Useful for precisely quantified transcriptome profiling, expressed SNV and indel analysis, splice variant analysis, SV analysis Useful for profiling the various RNA species and translational profiling
Methylome sequencing (Methyl-Seq)	Methylated DNA	Useful for identifying regions of DNA (such as promoters) which have become methylated Requires pretreatment of DNA with bisulphite prior to NGS
Chromatin immunoprecipitation and sequencing (ChIP-Seq)	DNA-protein complexes	Useful for identifying target regions of DNA-binding proteins (such as transcription factors) or DNA regions affected by protein modification (such as histone acetylation) Requires DNA-protein cross-linking and then antibody-mediated immunoprecipitation prior to NGS

CNV, copy number variant; SV, structural variant; SNV, single-nucleotide variant; indel, insertion or deletion mutation; miRNA, microRNA; lncRNA, long non-coding RNA; snoRNA, small nucleolar RNA; rRNA, ribosomal RNA.

from assessment of a number of features in the raw data which are distilled into a single value. This value indicates the probability of a base having been accurately called and therefore indicates the confidence with which a variant has been accepted as "true." The Phred score was originally devised for assessing quality of Sanger sequencing but it is also used for NGS data. Since the chemistries used in NGS are variable, each platform will have its own metrics which are used for calculating the Phred score [23].

In contrast to Sanger sequencing, which may have read lengths of up to 1 kb, NGS platforms produce short reads (typically 25–500 bases). Thus, once poor-sequence reads have been excluded, the next first step is to stitch the sequence fragments together. This process is known as sequence assembly. Theoretically, for WGS, it is possible to create the genomic sequence directly from the sequence data, and this is called de novo assembly. In diagnostic practice, it is more likely that the test genome will be compared with a known reference genome in order to find variations.

The sequence will be mapped and aligned automatically by the bioinformatics tools resulting in BAM (binary alignment/map) files. These contain information on the sequence and its location in relation to the reference sequence. Accurate mapping and alignment are essential parts of the bioinformatics analysis and can be confounded by pseudogenes, homopolymers and, in the case of RNA-Seq, contamination of samples by DNA. The BAM files can then be analysed further using software such as the Integrated Genome Viewer. The software will also produce a variant call file (VCF) which will contain information about the variant detected, the location of the variant and the number of reads at that location containing wild-type and variant sequences.

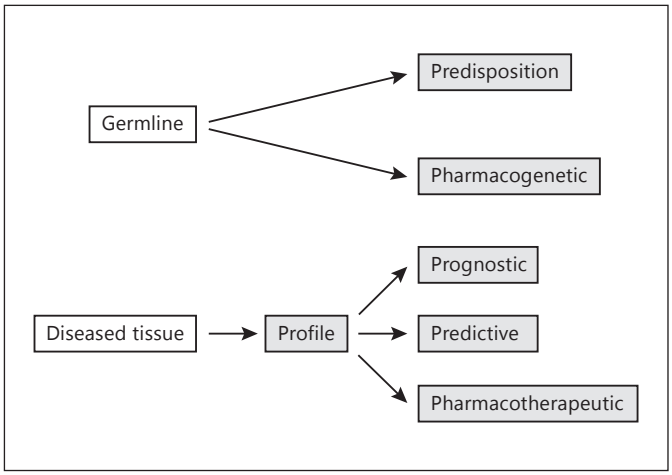

Fig. 5. A classification system for genetic biomarkers. Each biomarker can be put into a category depending on what kind of information it gives. Biomarkers from the germline give information about predisposition to a disease and metabolism of chemical therapies. Biomarkers from a diseased tissue give a profile of the disease and may also give prognostic, predictive, or pharmacotherapeutic information. A biomarker may occupy more than 1 category. For example, mutation of the *MSH2* gene will be a predisposition biomarker for Lynch syndrome. Tumours arising in a context of Lynch syndrome will have microsatellite instability, and therefore this mutation may also be a prognostic biomarker. Tumours with microsatellite instability also have a resistance for 5-fluorouracil therapies, and therefore it could also be a predictive marker. Finally, tumours with microsatellite instability may respond better to immunotherapy, and so it may also be a pharmacotherapeutic marker.

Interpretation of these data also requires a knowledge of what exactly has been sequenced, i.e. what percentage of the desired sequence has actually been tested (known as *breadth of coverage*) and how many molecules from each point of interest have been sequenced (known as *depth of coverage* or *read depth*). Since each sequencing reaction has a finite capacity, this will inevitably lead to a tension between the breadth of coverage required and the depth of sequencing possible (Fig. 3, 4). The best way to understand this is to consider 2 extreme examples. If a single short target (such as a PCR product from 1 exon) is sequenced on a platform which can perform 15 million sequencing reactions, then 100% breadth of coverage can be obtained of the target sequence with a sequencing depth of ×15 million. If however, there are 15 million targets, statistically only 1 molecule of each target can be sequenced for a breadth of coverage of 100%, thus limiting the depth of coverage to ×1. In this scenario, given the stochastic nature of DNA immobilization during the sequencing reaction, it is possible that some of the targets may have 2 molecules sequenced (depth of ×2) at the expense of other targets. If some targets, by chance, are not sequenced, there will be <100% breadth of coverage.

These 2 extreme examples serve to exemplify the different types of inferences which can be made from the data. In the first example, if there is a mutation in the target, even if the mutation is only present in 1% of the alleles, this will still lead to 150,000 reads containing mutant sequence. Nothing however can be inferred about structural changes. In the second example, if there is no bias, each target will be sequenced a similar number of times, and an average depth of sequencing can be derived. If however there is bias (such as gene amplification) some targets may have a read depth at variance with the average depth thereby allowing an inference of copy number change. However, low-frequency mutations/variants will not be identified since only 1 molecule from each target will be sequenced.

The optimum would be whole-genome and transcriptome coverage at great depth to allow all mutation types to be detected and all RNA species to be profiled. Given the rate of progress of the technologies, this may soon be possible although currently different assays give different information. The strengths of the different assays are listed in Table 1 and a key to terminology is given in Table 2.

The Clinical Potential of NGS and a Classification System for Biomarkers

The sheer volume of data produced by NGS technologies could become overwhelming especially for the non-specialist. A classification is proposed in which the "actionable" biomarkers are put into 6 distinct classes in accordance with the type of information they provide. This is called the "6 Ps of genetic biomarkers" (Fig. 5), and adoption of this type of classification would allow the clinician to see the clinical implication of the biomarker. An individual biomarker may be found in more than one of the following categories:

Predisposition Biomarkers

This refers to the germline variants which are associated with subsequent risk of disease. Since the whole genome can be sequenced, it is no longer necessary to identify large kindreds in order to map and then clone the mutant gene. NGS – both WGS and WES – has proven to be extremely useful in identifying the causative mutations in several rare inherited syndromes from only a few af-

Table 2. Glossary of commonly used technical and interpretative terms

Technical terminology	
Fragmentation	Breaking up of DNA/RNA (either physically or chemically) into 200- to 500-bp fragments
Adapters	Sequences of DNA ligated onto the DNA/RNA fragments to allow PCR and sequencing
Barcodes	Specific sequences added to DNA/RNA fragments allowing sample identification in multiplexed reactions
Library	All the fragments from a sample with adapters and barcodes added; libraries can be pooled
Hybrid capture	A means of enriching libraries using probes complementary to regions of interest (e.g. the exome)
Emulsion PCR	PCR performed in a droplet of water in an oil/water emulsion; usually used for clonal amplification in conjunction with beads so that each drop contains 1 bead, 1 DNA molecule (immobilized on the bead), and PCR components
Bridge amplification	PCR performed on an Illumina flow cell; DNA molecules are immobilized and clonally amplified to form a cluster
Flow cell	Chamber in which Illumina platform NGS is performed; bridge amplification is followed by sequencing by synthesis
Sequencing by synthesis	Sequence is recorded as fluorescently tagged bases are incorporated into the new strand
Semiconductor sequencing	Sequence is recorded by measuring changes in pH which occur following incorporation of bases in new strands
Pyrosequencing	Sequence is recorded by measuring pyrophosphate released following incorporation of bases in new strands
Read length	The length of sequence recorded from each sequencing reaction; it normally ranges from 25 to 500 bp
Paired-end sequencing	Sequencing performed in both directions
FASTQ file	Output file containing the raw sequence and information on sequence quality
SAM/BAM file (sequence/binary alignment map)	Output file containing the raw sequence aligned to a reference sequence; BAM – binary version of SAM
Interpretative terminology	
Depth of coverage	How often any particular point in the target regions has been sequenced
Breadth of coverage	How much of the target region has been sequenced
Reference sequence	The sequence against which newly generated sequencing data are compared
Sequence assembly	Alignment of the newly generated sequence from DNA fragments to the reference sequence
De novo assembly	Creation of a new sequence from the newly generated sequence of the DNA fragments; sequence overlap at the edges of the fragments allows a contiguous sequence to be created
SNV (single-nucleotide variant)	Variation at a single base
SNP (single-nucleotide polymorphism)	Single base variation occurring at a frequency of >1% in the population (if <1%, it is regarded as mutation)
Synonymous/non-synonymous change	Single base change which, in coding regions, may change the amino acid sequence (non-synonymous) or which may be silent (synonymous); in non-coding regions, synonymous changes may alter splicing, transcription, etc.
VUS (variant of unknown significance)	Gene variants can be put into 5 classes: pathogenic (known, disease causing); likely pathogenic (novel, likely disease causing); VUS (novel, uncertain whether pathogenic or benign); likely benign (novel, unlikely disease causing); benign (known, not associated with disease)
SV (structural variants)	Changes in large DNA fragments due to translocation, inversion, deletion, and duplication/amplification
CNV (copy number variation)	Amplification or deletion leading to either >2 or <2 copies of a specific gene/genomic sequence; there is a lot of CNV within the general population, and individuals may have duplicated/deleted regions without phenotype
Indel (insertion or deletion mutation)	Insertion or deletion of bases; indels are generally more difficult to detect using NGS, especially if located at the edges of a sequence; large indels (>30 bp) are problematic as shortened sequences may be filtered out as low quality

fected individuals [24–27]. As well as being used to identify the predisposition biomarkers in rare syndromes, NGS can identify – either through deliberate screening or as incidental findings – susceptibility alleles for a variety of more common syndromes. This is important in the syndromes which have variable penetrance and variable expression. Thus, a patient may have the germline variant but may not yet have an overt or typical manifestation of the disease. Similarly, a patient may be found to be a carrier of a mutant allele for an autosomal recessive syndrome. Both situations have implications for both the patient and extended family members but they also raise ethical issues (see below).

Pharmacogenetic Biomarkers

This refers to the germline variants which will predict how an individual will handle a particular kind of drug. This will have implications on whether or not the patient should be given the drug and also on the dosage of the drug that is administered [28–32].

Profile Biomarkers

This refers to the variants which are found within a diseased tissue. In tumours, for example, a variety of different mutations can be detected (see above), and these can be used to classify tumours – both as an addition to morphological classification and as an alternative to morphological classification. Some mutations are characteristic of certain tumours, and therefore they can be used to confirm or refute a morphological diagnosis [33–36]. For both tumour and non-tumour tissue, profile biomarkers would include a description of molecules expressed in that tissue. Certain disease states may have specific expression profiles, and thus the expressed profile could be used for classification.

Prognostic Biomarkers

This refers to the biomarkers which give information on the outcome of the disease. Knowledge of how a disease may behave allows management decisions to be made on, for example, whom and when to intervene with therapy.

Predictive Biomarkers

This refers to those biomarkers which give information as to whether the patient will respond to specific therapies. Predictive biomarkers allow the patient to be stratified into the appropriate treatment group and for precision medicine to the practised. Thus, certain mutations may be associated with resistance to specific therapy (such as *KRAS* mutation in a tumour indicating resistance of cetuximab) [37, 38]. Other mutations, however, may denote a sensitivity to specific therapies (such as a *BRAF* mutation in melanoma indicating a likely response to vemurafenib) [39].

Pharmacotherapeutic Biomarkers

This refers to the biomarkers identified in diseased tissue (whether they are gene mutations or expression patterns) which are amenable to direct therapeutic targeting. These may or may not be biologically relevant to the disease, e.g. there may be a passenger mutation which creates a tumour-specific neo-antigen, there may be cell surface molecules aberrantly expressed as an epiphenomenon of a driver mutation, or there may be novel targets which show synthetic lethality with driver mutations.

Limitations of NGS

As with all techniques, there are a number of factors which can confound the interpretation of NGS data. These range from variations in the templates (i.e. pre-analytical confounders), technical issues around the methodology (analytical confounders) and features within the data (post-analytical confounders).

Pre-Analytical Confounders

NGS can be performed on nucleic acid derived from any template although, in clinical practice, the most likely templates to be used will be frozen tissue, formalin-fixed paraffin-embedded tissue, and, increasingly more often, plasma (i.e. liquid biopsy). The first of the confounding factors is the content of the tissue, and steps should be taken to ensure that the sample must be appropriate for the test that is being performed. For example, if tumour tissue is being tested in order to obtain a mutation profile, then the ratio of tumour cells to non-tumour cells must be above the limit of detection of the methodology. Similarly, if RNA-Seq is being performed to profile mRNA or microRNA, it should be noted that this may be affected by external factors such as warm ischaemia (when the blood supply is cut off from the tissue during surgery) [40]. As a general rule, the highest quality of DNA and RNA is obtained from tissue which is frozen as quickly as possible. Short delays in freezing will not affect the DNA but there may be subtle changes in the RNA expression profile (a phenomenon known as cold ischaemia) [41]. The circulating free DNA obtained from plasma is usually of high quality but is frequently fragmented due to

cleavage prior to release from cell nuclei into the circulation. Methods requiring DNA fragments >200 bp are less likely to succeed on a template derived from plasma [42].

Probably the largest pre-analytical confounder in NGS is the effect caused by processing of fresh tissue into formalin-fixed paraffin-embedded tissue [43, 44]. This will cause fragmentation and cross-linking of the nucleic acid resulting in low quality and low-molecular-weight DNA. This is generally not suitable for WGS but can be used for WES and for targeted sequencing. Formalin-fixed paraffin-embedded tissue-derived DNA is more prone to AT/GC drop-out, random PCR errors and to deamination artefacts [45]. The latter is caused by loss of the amino residue in cytosine resulting in a sequence change to thymine during PCR. Sequencing platforms which analyse both strands are able to avoid these errors.

Analytical Confounders

Our experience has been that, under optimal conditions, NGS is a robust and reproducible technique. Each NGS platform will have its own strengths and weaknesses, and there will be platform-specific sources of variation. It is beyond the scope of this review to discuss the specific confounders but there are some generic confounders which can be identified. Template concentration – if too low or too high – can cause errors to occur during each of the steps of library construction. Inaccurate dilution of the libraries themselves can induce errors (such as excessively dense clusters) and there may be batch-to-batch variation due to changes occurring in the consumables. Although library construction is theoretically non-biased, there is also a tendency for underrepresentation of CG- and AT-rich areas [46]. NGS technologies do not deal well with repeat sequences, and those using pyrophosphate as a means of sequence detection are particularly prone to errors in mononucleotide repeat sequences. Platforms which sequence very short fragments are more likely to miss large (>20 bp) deletions [47]. A potential problem common across the various platforms is if some of the molecules in a clonally amplified group of molecules fall out of phase during the sequencing reaction; thus, different residues will be added to those being added to the other molecules in the group causing a loss of signal quality.

Post-Analytical Confounders and Interpretation in Context

The progress in sequencing technologies has led to the development of robust platforms for acquiring large volumes of sequencing data. The greater challenge now lies (i) in the optimization of the bioinformatics pipelines to ensure that the data are accurate and (ii) in the discrimination of clinically meaningful variation from irrelevant "noise."

Firstly, once the sequence data have been obtained, they need to be filtered to remove low-quality reads. The sheer number of sequencing reactions that are performed means that there will inevitably be errors. For platforms which sequence both strands, they can also filter out spontaneous PCR errors and deamination artefacts by ensuring that a variant is only called when it is present in both strands. Next, the generated sequence has to be accurately aligned to a reference sequence (or accurately stitched together for de novo assembly).

Once the sequence assembly is complete and true variants have been identified, it is important to interpret the data with the laboratory and clinical context in mind. For example, if mutations are not found in a tumour, it is essential to confirm that there was sufficient tumour in the template initially. A sample containing insufficient tumour cells should not get through for testing but samples may be near the borderline limit of detection. In this case, especially for resistance mutations, a decision needs to be made regarding the confidence with which a "true negative" call can be made. A factor informing this decision may be the depth of sequencing achieved for that specific site.

If a sequence variant is found, it is important to know whether this represents a true pathogenic event. Even fairly large-scale structural changes (such as copy number variations) can occur without any obvious phenotypic effect. Given the variation within the human population, it could be argued that structural variant analysis in a tumour should be undertaken in comparison with matched normal tissue from the same patient. This would give the most accurate information but would double the cost of the test. As more sequencing is performed, more deviations from the reference sequence will inevitably be identified. Some will be known pathogenic mutations whilst others will be referred to as variants of uncertain significance. Guidelines have been published on inferring the pathogenicity of a newly identified sequence variant [48, 49]. It is constantly emphasized that the simple presence of a variant – even one that has been called with confidence and is predicted to severely affect gene function – is not enough to assume that it is clinically pathogenic. There is a concern that simply taking mutations at face value may lead to overcalling of pathogenic mutations and inappropriate clinical management. Given the impending explosion of NGS-derived data, there is a need

to establish international databases in order to pool knowledge and facilitate interpretation. Since NGS can pick up pathogenic changes incidentally, there needs to be a local decision as what information will be relayed back to the clinician – whether it will be all changes, only validated pathogenic changes, or only changes in the targets of interest.

Once a known pathogenic variant is identified, its interpretation will still be context dependent. Thus, a *BRAF* mutation in melanoma may mean that a patient will respond to the Braf inhibitor vemurafenib – this would therefore be put into the "predictive" group as well as the "profile" group of genetic biomarkers. However, *BRAF* mutation does not have the same implication in colorectal cancers – it is a driver mutation but colorectal cancers tend not to respond to Braf inhibitor treatment [39, 50]. The lack of response of colorectal cancers to vemurafenib is due to activation of pathways which bypass the Braf in inhibition. Thus, *BRAF* mutation will be in the profile of a colorectal cancer but it would not be included in the predictive group.

In other instances, it may be that the tumour has become "amnesic" for certain driver mutations and therefore may not respond to targeted therapy [51]. Finally, heterogeneity within a tumour continues to be a challenging question. If a variant is found at a lower frequency than the other variants, it most likely represents a subclone. If the variant is known to give resistance to a chemotherapy, then a decision will have to be made about the risk of selecting out resistant cells if that therapy is used.

NGS and Laboratory Management

Guidelines on the management of an NGS-based diagnostic service have been published, and an in-depth discussion is beyond the scope of this review [52–55]. It is worthwhile pointing out that the complexity of the NGS methodology provides major challenges. For a laboratory providing an "end-to-end" service beginning with receipt of the tissue sample and concluding with a list of variants, there are a number of "wet" bench steps and "dry" bioinformatics steps. Each step – both wet and dry – needs to be quality controlled, and parameters need to be set for accepting results (DNA quality, Phred score, number of reads mapping to target areas, depth of coverage, etc.). The quality control cannot deal with template-intrinsic confounders (such as formalin fixation or low tumour cell content) but it needs to ensure that there are no artefacts arising from the technical processes of the assay. The technical performance of the assay (e.g. limit of detection, short-term and long-term precision) needs to be assessed and documented. Given the multiple steps involved, specimen tracking needs to be accurate, and the final sequence data need to be validated. Each laboratory must decide on when and how validation is to be performed but variant detection protocols for all types of variants (i.e. single-nucleotide polymorphisms, indels etc.) should be validated. Every time one of the steps is changed, whether it is a wet step (such as a new DNA extraction kit) or a dry step (such as a new bioinformatics algorithm [56]), then the test needs to be validated to confirm that bias has not been introduced. The degree to which a diagnostic pathologist should be involved in the management of an NGS-based diagnostic service is a moot point. At the very least, however, the pathologist should have some understanding of the processes involved in the generation of the data and the potential sources of artefact.

New Technologies

NGS technologies have revolutionized biological research and, in due course, will transform the way that diagnostic pathology and clinical medicine are practised. Technology however continues to evolve, and newer technologies are being developed which offer an alternative methodology and which overcome some of the limitations of current technologies. Current technologies rely on clonal PCR to generate a signal which is sufficient for detection. A methodology involving single-molecule sequencing would theoretically be much better and would require a lower quantity of starting template. Two technologies have developed single-molecule sequencing. The technology produced by Pacific Biosystems requires library construction but each molecule of the library is sequenced using immobilized specially modified polymerase enzyme, and each nucleotide is detected during incorporation into the new strand [57]. The technology produced by Oxford Nanopore does not require new molecule synthesis. Single-stranded DNA molecules are fed through tiny pores in a special electrically resistant membrane. Specialized proteins feed single-stranded DNA through the pores which have current running through them [58]. The molecule disrupts the current as it passes through the pore, and from the pattern of disruption, the DNA sequence can be inferred. Chips have been produced by Oxford Nanopore which can be plugged into

a USB socket of a computer, and the DNA sequence can be read in real time. These new technologies are able to produce much longer reads (up to tens of kilobases) than current NGS technologies. They are however more error prone.

Conclusion

NGS technologies are opening up new possibilities for tissue interrogation and molecular diagnostics. Whilst the information which they will generate is undoubtedly going to contribute to patient management, it should not be taken at face value. The diagnostic pathologist is at the interface of clinical and laboratory medicine and is best placed to provide interpretation of the data within the context of the clinical question. Pathologists need to understand and embrace these technologies in order to maximize their clinical utility.

Disclosure Statement

There are no conflicts of interest.

References

1 Mallory FB: On certain improvements in histological technique. I. A differential stain for amoebae coli. II. Phosphotungstic-acid-haematoxylin stain for certain tissue elements. III. A method of fixation for neuroglia fibres. J Exp Med 1897;2:529–533.
2 McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, et al: A physical map of the human genome. Nature 2001; 409:934–941.
3 Frese KS, Katus HA, Meder B: Next-generation sequencing: from understanding biology to personalized medicine. Biology 2013;2: 378–398.
4 Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al: Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012;2012: 251364.
5 Mardis ER: Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 2008;9:387–402.
6 Mardis ER: Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif) 2013;6:287–303.
7 Ibrahem S, Seth R, O'Sullivan B, Fadhil W, Taniere P, Ilyas M: Comparative analysis of pyrosequencing and QMC-PCR in conjunction with high resolution melting for KRAS/BRAF mutation detection. Int J Exp Pathol 2010;91:500–505.
8 Li J, Wang L, Mamon H, Kulke MH, Berbeco R, Makrigiorgos GM: Replacing PCR with COLD-PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. Nat Med 2008;14:579–584.
9 Milbury CA, Li J, Makrigiorgos GM: PCR-based methods for the enrichment of minority alleles and mutations. Clin Chem 2009;55: 632–640.
10 Pochon X, Bott NJ, Smith KF, Wood SA: Evaluating detection limits of next-generation sequencing for the surveillance and monitoring of international marine pests. PLoS One 2013; 8:e73935.
11 Ecker JR, Bickmore WA, Barroso I, Pritchard JK, Gilad Y, Segal E: Genomics: ENCODE explained. Nature 2012;489:52–55.
12 Hrdlickova R, Toloue M, Tian B: RNA-Seq methods for transcriptome analysis. Wiley Interdiscip Rev RNA 2017, Epub ahead of print.
13 Ku CS, Naidoo N, Wu M, Soong R: Studying the epigenome using next generation sequencing. J Med Genet 2011;48:721–730.
14 Mundade R, Ozer HG, Wei H, Prabhu L, Lu T: Role of ChIP-Seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell Cycle 2014;13:2847–2852.
15 Sun Z, Cunningham J, Slager S, Kocher JP: Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. Epigenomics 2015;7: 813–828.
16 Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al: Library construction for next-generation sequencing: overviews and challenges. Biotechniques 2014;56:61–64, 66, 68, passim.
17 Podnar J, Deiderick H, Huerta G, Hunicke-Smith S: Next-generation sequencing RNA-Seq library construction. Curr Protoc Mol Biol 2014;106:4.21.1–19.
18 Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D: Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. PLoS One 2011;6:e28240.
19 Metzker ML: Sequencing technologies – the next generation. Nat Rev Genet 2010;11:31–46.
20 Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, Griffiths AD: Amplification of complex gene libraries by emulsion PCR. Nat Methods 2006;3:545–550.
21 Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al: An integrated semiconductor device enabling non-optical genome sequencing. Nature 2011;475:348–352.
22 Voelkerding KV, Dames SA, Durtschi JD: Next-generation sequencing: from basic research to diagnostics. Clin Chem 2009;55: 641–658.
23 Ulahannan D, Kovac MB, Mulholland PJ, Cazier JB, Tomlinson I: Technical and implementation issues in using next-generation sequencing of cancers in clinical practice. Br J Cancer 2013;109:827–835.
24 Galmiche L, Serre V, Beinat M, Assouline Z, Lebre AS, Chretien D, et al: Exome sequencing identifies MRPL3 mutation in mitochondrial cardiomyopathy. Hum Mutat 2011;32: 1225–1231.
25 Liu L, Li XB, Zi XH, Shen L, Hu Zh M, Huang Sh X, et al: A novel hemizygous SACS mutation identified by whole exome sequencing and SNP array analysis in a Chinese ARSACS patient. J Neurol Sci 2016;362:111–114.
26 Morita H: Identification of a mutation causing hypertrophic cardiomyopathy using whole exome sequencing: a proof-of-concept. J Cardiol 2016;67:131–132.
27 Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, Sougnez C, et al: Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. N Engl J Med 2010; 363:2220–2227.
28 Del Re M, Restante G, Di Paolo A, Crucitta S, Rofi E, Danesi R: Pharmacogenetics and metabolism from science to implementation in clinical practice: the example of dihydropyrimidine dehydrogenase. Curr Pharm Des 2017;23:2028–2034.
29 Johnson JA, Caudle KE, Gong L, Whirl-Carrillo M, Stein CM, Scott SA, et al: Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for pharmacogenetics-guided warfarin dosing: 2017 update. Clin Pharmacol Ther 2017;102:397–404.
30 Meulendijks D, Cats A, Beijnen JH, Schellens JH: Improving safety of fluoropyrimidine chemotherapy by individualizing treatment based on dihydropyrimidine dehydrogenase activity – ready for clinical practice? Cancer Treat Rev 2016;50:23–34.

31 Mizzi C, Dalabira E, Kumuthini J, Dzimiri N, Balogh I, Basak N, et al: A European spectrum of pharmacogenomic biomarkers: implications for clinical pharmacogenomics. PLoS One 2016;11:e0162866.
32 Skrzypczak-Zielinska M, Borun P, Bartkowiak-Kaczmarek A, Zakerska-Banaszak O, Walczak M, Dobrowolska A, et al: A simple method for TPMT and ITPA genotyping using multiplex HRMA for patients treated with thiopurine drugs. Mol Diagn Ther 2016; 20:493–499.
33 Arrigoni G, Doglioni C: Atypical lipomatous tumor: molecular characterization. Curr Opin Oncol 2004;16:355–358.
34 Deng G, Bell I, Crawley S, Gum J, Terdiman JP, Allen BA, et al: BRAF mutation is frequently present in sporadic colorectal cancer with methylated hMLH1, but not in hereditary nonpolyposis colorectal cancer. Clin Cancer Res 2004;10:191–195.
35 Johnson V, Volikos E, Halford SE, Eftekhar Sadat ET, Popat S, Talbot I, et al: Exon 3 beta-catenin mutations are specifically associated with colorectal carcinomas in hereditary non-polyposis colorectal cancer syndrome. Gut 2005;54:264–267.
36 Shimada S, Ishizawa T, Ishizawa K, Matsumura T, Hasegawa T, Hirose T: The value of MDM2 and CDK4 amplification levels using real-time polymerase chain reaction for the differential diagnosis of liposarcomas and their histologic mimickers. Hum Pathol 2006; 37:1123–1129.
37 Lievre A, Bachet JB, Le Corre D, Boige V, Landi B, Emile JF, et al: KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. Cancer Res 2006;66: 3992–3995.
38 Ramos FJ, Macarulla T, Capdevila J, Elez E, Tabernero J: Understanding the predictive role of K-ras for epidermal growth factor receptor-targeted therapies in colorectal cancer. Clin Colorect Cancer 2008;7(suppl 2):S52–S57.
39 Davies MA: Molecular approaches to tumor inhibition in melanoma. Clin Adv Hematol Oncol 2015;13:831–833.
40 Dash A, Maine IP, Varambally S, Shen R, Chinnaiyan AM, Rubin MA: Changes in differential gene expression because of warm ischemia time of radical prostatectomy specimens. Am J Pathol 2002;161:1743–1748.
41 Grizzle WE, Otali D, Sexton KC, Atherton DS: Effects of cold ischemia on gene expression: a review and commentary. Biopreserv Biobank 2016;14:548–558.
42 Devonshire AS, Whale AS, Gutteridge A, Jones G, Cowen S, Foy CA, et al: Towards standardisation of cell-free DNA measurement in plasma: controls for extraction efficiency, fragment size bias and quantification. Anal Bioanal Chem 2014;406:6499–6512.
43 Atanesyan L, Steenkamer MJ, Horstman A, Moelans CB, Schouten JP, Savola SP: Optimal fixation conditions and DNA extraction methods for MLPA analysis on FFPE tissue-derived DNA. Am J Clin Pathol 2017;147:60–68.
44 Gilbert MT, Haselkorn T, Bunce M, Sanchez JJ, Lucas SB, Jewell LD, et al: The isolation of nucleic acids from fixed, paraffin-embedded tissues – which methods are useful when? PLoS One 2007;2:e537.
45 Kim S, Park C, Ji Y, Kim DG, Bae H, van Vrancken M, et al: Deamination effects in formalin-fixed, paraffin-embedded tissue samples in the era of precision medicine. J Mol Diagn 2017;19:137–146.
46 Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC: Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PLoS One 2013;8:e62856.
47 Muzzey D, Evans EA, Lieber C: Understanding the basics of NGS: from mechanism to variant calling. Curr Genet Med Rep 2015;3: 158–165.
48 Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, et al: Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med 2017;19:249–255.
49 MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al: Guidelines for investigating causality of sequence variants in human disease. Nature 2014;508:469–476.
50 Kopetz S, Desai J, Chan E, Hecht JR, O'Dwyer PJ, Maru D, et al: Phase II pilot study of vemurafenib in patients with metastatic BRAF-mutated colorectal cancer. J Clin Oncol 2015;33: 4032–4038.
51 Felsher DW: Oncogene addiction versus oncogene amnesia: perhaps more than just a bad habit? Cancer Res 2008;68:3081–3086; discussion 3086.
52 Aziz N, Zhao Q, Bry L, Driscoll DK, Funke B, Gibson JS, et al: College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. Arch Pathol Lab Med 2015;139:481–493.
53 Cree IA, Deans Z, Ligtenberg MJ, Normanno N, Edsjo A, Rouleau E, et al: Guidance for laboratories performing molecular pathology for cancer patients. J Clin Pathol 2014;67:923–931.
54 Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, et al: Guidelines for diagnostic next-generation sequencing. Eur J Hum Genet 2016;24:2–5.
55 Weiss MM, Van der Zwaag B, Jongbloed JD, Vogel MJ, Bruggenwirth HT, Lekanne Deprez RH, et al: Best practice guidelines for the use of next-generation sequencing applications in genome diagnostics: a national collaborative study of Dutch genome diagnostic laboratories. Hum Mutat 2013;34:1313–1321.
56 Hwang S, Kim E, Lee I, Marcotte EM: Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep 2015;5:17875.
57 Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, et al: Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods 2015;12:780–786.
58 Venkatesan BM, Bashir R: Nanopore sensors for nucleic acid analysis. Nat Nanotechnol 2011;6:615–624.