HUGO J (2011) 5:1–12
DOI 10.1007/s11568-011-9156-3

REVIEW

Review of massively parallel DNA sequencing technologies

Sowmiya Moorthie · Christopher J. Mattocks · Caroline F. Wright

Received: 26 May 2011 / Revised: 29 July 2011 / Accepted: 3 October 2011 / Published online: 27 October 2011

Abstract Since the development of technologies that can determine the base-pair sequence of DNA, the ability to sequence genes has contributed much to science and medicine. However, it has remained a relatively costly and laborious process, hindering its use as a routine biomedical tool. Recent times are seeing rapid developments in this field, both in the availability of novel sequencing platforms, as well as supporting technologies involved in processes such as targeting and data analysis. This is leading to significant reductions in the cost of sequencing a human genome and the potential for its use as a routine biomedical tool. This review is a snapshot of this rapidly moving field examining the current state of the art, forthcoming developments and some of the issues still to be resolved prior to the use of new sequencing technologies in routine clinical diagnosis.

Keywords Next generation sequencing · Targeting · Massively parallel

S. Moorthie · C. F. Wright
PHG Foundation, 2 Worts Causeway, Cambridge CB1 8RN, UK

C. J. Mattocks
National Genetics Reference Laboratory (Wessex), Salisbury District Hospital, Salisbury SP2 8BJ, UK

Present Address:
C. F. Wright (✉)
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
e-mail: caroline.wright@sanger.ac.uk

Introduction

The basic principles of DNA sequencing have remained constant since the development of the first practical method by Sanger et al. (1977). Even so, the classic Sanger method has undergone various modifications and refinements in the intervening years, most recently driven by the requirements of the Human Genome Project to facilitate automation and to increase throughput. Perhaps most significantly, contemporary methods use four different base-specific fluorescent dyes (Smith et al. 1986) instead of radioactive labels, and cumbersome gel electrophoresis has been replaced by automated capillary electrophoresis (Luckey et al. 1990). These developments have dramatically increased the efficiency of Sanger sequencing, which is now widely considered the gold standard for clinical diagnostic use. However, the technique remains too laborious and expensive for routine sequencing of anything more than a few genes. In an attempt to address this short-coming, a diverse array of new sequencing technologies have been developed and are currently in development. Although the widely perceived aim of practical and affordable whole genome sequencing is ambitious, requiring major improvements in run capacity, speed of processing and cost, progress to date has been remarkable (see reviews Mardis 2011; Metzker 2010; Pettersson et al. 2009; Tucker et al. 2009; Voelkerding et al. 2009).

Classification of technologies

The terminology surrounding the new sequencing technologies is diverse and often confusing with terms such as ‘next generation’, ‘massively parallel’ and ‘clonal’ sequencing being used as global classifiers for, what is, essentially the same thing. In an attempt to bring some

clarity to classification we have divided DNA sequencing technologies into three generations (Pettersson et al. 2009).

The first generation is synonymous with Sanger sequencing, which has been predominant since the 1970s. The defining characteristics of this technology are that each sequencing reaction represents a single, predefined target (up to about 1 kb) and this represents all copies of that target present in the original sample and thus its allelic content. The underlying principle of all post-Sanger DNA sequencing technologies, which is enabling the explosion in capacity and exponentially decreasing costs, is massive parallelisation. A fragmented input sample is captured on an array in such a way that each spatially identifiable location or feature is populated by a single target molecule. Depending on the technology a single sequencing array may comprise many millions or even billions of features, which are all sequenced in parallel in a single run, each feature generating a single sequencing 'read'. This is fundamentally different from Sanger sequencing in two key respects: first, the specific location of the reads is not pre-determined and so must be computationally determined (referred to as mapping or alignment); and second, because each read represents a single starting molecule, multiple coverage is required to analyse the full allelic content of the sample.

With second generation sequencing, widely referred to as 'next generation sequencing (NGS)' it is necessary to clonally amplify the isolated targets in order to generate sufficient signal for detection during the sequencing run. This process is usually performed in situ on a solid substrate and generates clusters of many thousands of identical DNA targets (sometimes called polonies) at each feature. With these technologies sequencing is performed through stepwise incorporation of suitably modified subunits. Generally speaking read lengths are shorter than those achieved by Sanger sequencing, although they are rapidly improving. This is an important consideration since short read length can make accurate assembly and alignment computationally challenging (Flicek et al. 2011; Li et al. 2008; Li and Durbin 2009). The key difference with third generation, or 'next next generation', sequencing is that the chemistry and/or detection has been refined so that no clonal amplification of the target is required before the run. These technologies are predominantly still in development and use a wide range of different detection methodologies.

Below we review the principles behind these alternative technologies, compare and contrast their characteristics, and provide an overview of some of the targeting techniques and bioinformatics tools that have been developed alongside them.

Second generation sequencing

Since late 2004, three principal NGS technologies have been commercially available (see Table 1). These technologies have been made available on an increasing range of platforms designed to suit different applications and capacity requirements from large genome centres down to the clinical laboratory. The underlying chemistries are briefly described below.

Reversible termination

This method closely resembles the Sanger sequencing-by-synthesis method, but uses special fluorescently labelled terminator nucleotides, which allow the chain termination process to be reversed (Bentley et al. 2008). It was originally developed by Solexa and is now commercialised by Illumina through the Genome Analyser and HiSeq systems; a further addition to the range will be the MiSeq, a lower capacity instrument due for release in mid-2011.

Template DNA molecules are generated by fragmentation of the sample followed by ligation of end specific universal adaptors. These fragments are then hybridised to a dense 'lawn' of universal probes immobilised to a glass surface known as a flow cell upon which both amplification and sequencing take place. Clonal amplification is performed using a process termed 'bridge amplification'; a surface PCR which uses two tethered universal primers to create dense clusters of identical DNA across the plate. The sequencing reaction begins with the addition of a universal sequencing primer, which hybridises to the adaptor sequences added in the first stage. The sequencing chemistry involves three stages. First, chain extension is performed using DNA polymerase and the four reversible nucleotide terminators, each labelled with a different fluorescent dye. Incorporation of a complementary nucleotide results in termination of polymerisation—this process is allowed to run to completion to ensure all templates on the flow cell are extended by a single base. Next, unincorporated nucleotides are washed off and the incorporated base on each cluster is identified by colour imaging. Finally, the dye and the terminating group are chemically cleaved to prepare the templates for the next round of incorporation and imaging. These three stages are repeated over several hundred cycles generating a temporal series of colour images, which can be computationally converted into sequence reads each corresponding to a feature on the array.

Pyrosequencing

Pyrosequencing is also based on a sequencing-by-synthesis technique, but rather than measuring fluorescence associated with specific nucleotides, it relies on indirect detection of incorporation events (Margulies et al. 2005). This technology was available in individual reaction form (Qiagen) before a massively parallelised version was

Table 1 Summary of existing NGS platforms

Platform	Illumina/Solexa *GAIIe, GAIIx, HiSeq 1000, HiSeq 2000, MiSeq*	Roche/454 *GS FLX, GS Junior*	Life Technologies/ Applied Biosystems *5500 SOLiD, 5500xl SOLiD*
Methodology	Reversible Termination	Pyrosequencing	Sequencing by Ligation
Loading	Adaptors on template DNA bind high density primers across surface of slide	Adaptors on template DNA bind primers on beads, one molecule per bead	Adaptors on template DNA bind primers on beads, one molecule per bead
Clonal Amplification	Surface PCR used to generate clusters by bridge amplification	Emulsion PCR used to create clusters on beads	Emulsion PCR used to create clusters on beads
Parallelisation	Random array on flow cell	Beads loaded onto high density plate	Beads immobilised on high density glass slide
Sequencing enzyme	DNA polymerase	DNA polymerase	DNA ligase
Generation of complementary strands	4 labelled terminator nucleotides added	One of 4 labelled nucleotide added, *n* incorporated, and phosphate released	16 labelled 8-base oligonucleotides added and hybridised
Detection	4 colours detected	Flash detected proportional to *n*	4 colours detected
Re-initiation	Terminator and dye removed		Last 3 bases and label removed

commercialised by Roche/454. Two platforms supporting this chemistry are now available: the Genome Sequencer FLX and the GS Junior, a low capacity version.

Template DNA molecules can be generated either by fragmentation or standard PCR. If fragmentation is used, universal adaptors are ligated to the fragment end, similar to the method used by Illumina. In the case of PCR, the adaptors can be built into the primers. The prepared fragments are hybridised to special beads upon which both amplification and sequencing takes place; the beads are used in excess to ensure that each bead binds a maximum of one template molecule. A mix of these beads, PCR reagents and oil is then agitated to form an emulsion of tiny oil reaction chambers, each containing a single bead with a single molecule attached and all the components of a PCR. This is subjected to thermal cycling to clonally amplify the DNA template on the surface of each bead (known as emulsion PCR or emPCR). The sequencing reaction is performed on a specially fabricated 'PicoTitre Plate (PTP)'—this comprises millions of microscopic wells, each just big enough to contain a single template bead. After breaking the emulsion the beads are loaded onto the PTP along with other, much smaller beads that contain all the reagents necessary for the sequencing reaction except the nucleotides. Sequencing proceeds with the sequential addition of each individual nucleotide in turn (i.e. A, then C, then G, then T). If a nucleotide is incorporated by DNA polymerase into the growing DNA strand, an inorganic phosphate ion is released. This initiates an enzyme cascade resulting in the release of a flash of light. Since no terminators are used in this chemistry, incorporation of nucleotides into homopolymer stretches continues until a different base is encountered and the associated light flash is proportionally brighter. The location and intensity of light emitted is detected by a camera across the whole plate. Excess nucleotides are then washed off in preparation for the next cycle. This process is repeated several hundred times to build the temporal image sequence. Unlike other chemistries the number of incorporation cycles required to reach a particular read length is dependent on the sequence composition of the template. On average read length is expected to be $\sim 2.5\times$ cycle number but this could be less with more homopolymers. As before the temporal series of images can be computationally converted into sequence reads.

Sequencing by ligation

Unlike the previous techniques, this method does not involve polymerase based DNA synthesis, but instead uses ligation of fluorescently labelled hybridisation probes to determine the sequence of a template DNA strand two

bases at a time (Shendure et al. 2005). It has been commercialised by Life Technologies/Applied Biosystems through the SOLiD (Sequencing by Oligonucleotide Ligation and Detection) system.[1]

Template DNA molecules are prepared by fragmentation, adaptor ligation, hybridisation to beads and emPCR in a similar fashion to that described for the Roche/454 system above. After breaking the emulsion the beads are immobilised at high density on a glass slide. Sequencing proceeds with the addition of a universal primer, followed by fluorescently labelled oligonucleotide probes. Each probe comprises eight bases, of which only the first two define the probe whilst the following six are degenerate (i.e. able to pair with any nucleotide sequence on the template strand). After the complementary probe hybridises to the template DNA, it is chemically linked to the growing strand by the enzyme DNA ligase. The flow cell is then washed to remove excess probes and imaged to record the ligation cycle. Then, the three terminal degenerate bases, along with the fluorescent dye, are cleaved from the bound probes and the flow cell is washed again (this is known as a ligation cycle). This process is repeated a number of times, after which the newly synthesised strand is entirely denatured and removed from the template. At this point a new primer, which is one base shorter than that previously used (i.e. $n - 1$), is hybridised to the template and a new round of ligation cycles performed. In all, five rounds of ligation cycles are performed, each one using a primer one base shorter than the last. By this process, all bases on the template strand are interrogated twice. Counter-intuitively, although there are 16 possible permutations of the first two bases in the probe, only four coloured dyes are used; thus each colour represents four possible different two-base permutations. The arrangement of colours is such that if the first base is known, the second can be inferred. Since the first base in the sequence belongs to the universal primer added initially, the rest of the sequence can be sequentially inferred from the raw colour data, which is called colour space, by applying logical rules. This system, known as 2-base encoding, enables miss-called bases to be distinguished from true sequence variants as the former lead to logical impossibilities.

Third generation sequencing

A large number of companies are involved in developing much faster and higher throughput third or next–next generation DNA sequencing systems, some of which have already launched and some of which are still in stealth mode. Most of these are focussed on sequencing single molecules of DNA in real-time, and although many are based on sequencing-by-synthesis, there are several novel methodologies such as monitoring the passage of DNA through nanopores. A key advantage of single molecule sequencing is that no clonal amplification is required. This not only reduces preparation time, but effectively eliminates biases and errors introduced at this stage. In addition it is generally expected that these methods will generate much longer reads (potentially tens to hundreds of kilobases) which will enable much more accurate mapping, particularly in repetitive regions, and facilitate haplotyping. Moreover, some technologies have been demonstrated to be capable of distinguishing methylated cytosine bases, which could open the door to direct epigenetic analysis.

There are numerous third generation sequencing platforms at very different stages of development—ranging from basic research through to a launched product—which may be destined for different applications based on the precise idiosyncrasies of the sequencing chemistry and resultant performance metrics (e.g. error rate, read length, yield per run, cost per base, etc.—see later). The third generation technologies can be divided into categories based on the method they use to detect the DNA sequence:

Fluorescence

Most of the third generation sequencing platforms under development using fluorescence detection are based on the standard sequencing-by-synthesis method. The first single molecule sequencing platform to market was the Heliscope from Helicos Bioscience (Harris et al. 2008), launched in 2009, which is based on a similar methodology to that described for the Solexa/Illumina second generation platform. However this is a single molecule method and all the nucleotides have the same fluorescent label which acts as the terminating moiety. This means that the nucleotides need to be added individually and sequentially in order to identify which base is added when (Bowers et al. 2009). Following washing of excess nucleotides and polymerase, the slide is imaged to identify where bases were incorporated. The dye is then cleaved in preparation for the next nucleotide addition and the process repeated for each nucleotide, with each cycle extending the DNA strand by a single base. The data are then analysed to build up a sequence read for each location.

Another single molecule approach is the SMRT™ chemistry which has been made commercially available on the PACBIO RS platform from Pacific Biosciences. One of the main problems with single molecule sequencing is that the incorporation event that needs to be detected is so small

[1] An open platform based on sequencing-by-ligation has also been developed by *George Church/Dover Systems*, known as the *Polonator* (Shendure et al. 2005), but is not described further as it had a substantially smaller impact than the other technologies.

it is difficult to detect above background noise. This is the main reason for the use of clonal amplification in second generation systems. Pacific Biosciences have solved this problem by performing the sequencing in specially designed wells called zero mode wave guides which effectively eliminate the background noise (Eid et al. 2009). Template DNA forms a complex with the polymerase and nucleotide incorporation is detected by laser excitation and fluorescence monitoring in each well. The difference between this method and others that use fluorophores is that the dye is attached to the phosphate of the nucleotide rather than the base itself. Thus it is cleaved and released as a natural part of DNA synthesis, resulting in release of the dye without interruption to the sequencing process. The sequencing is therefore both single molecule and real-time.

Life Technologies have developed an approach which uses a DNA polymerase modified by the addition a quantum dot [Qdot® (Karow 2010b)]. This is a tiny nanocrystal that absorbs photons of light, then re-emits photons at a different wavelength. Template DNA is immobilised to the surface of a glass slide, and sequencing is initiated by the addition of a primer, the modified DNA polymerase and nucleotides with base specific fluorescent labels. As bases are incorporated the nucleotide labels are energised by the Qdot on the polymerase in a process known as fluorescence resonance energy transfer (FRET), which generates a very strong localised fluorescence signal (around 100-fold greater than standard dyes). FRET can only occur when the two fluorescent moieties (polymerase and nucleotide) are in close proximity to each other i.e. at incorporation thus elegantly eliminating interfering background fluorescence in the reaction chamber. At the end of a sequencing run both the polymerase and newly synthesised DNA strand can be removed, allowing the immobilised template DNA to be sequenced repeatedly. Life technologies claim that this system allows the read length and sequencing accuracy to be tailored to the application by adjusting the mode of repetitive sequencing.

There are also numerous other smaller companies developing third generation DNA sequencing platforms based on fluorescence detection, such as GnuBio, which is developing a microfluidics device that uses microdroplets as miniature reaction vessels, thus vastly reducing the cost of the reagents.

Electronic

Various third generation DNA sequencing platforms are being developed that are capable of converting a DNA sequence directly into an electrical signal. This essentially amounts to direct generation of digital information promising the enticing prospect of label-free sequencing. Platforms using such technology would be extremely cheap to produce and be both fast and scalable.

The system recently released by Ion Torrent/Life Technologies uses a sequencing-by-synthesis method almost identical to pyrosequencing. The key difference is that incorporations are detected by monitoring the release of H^+ ions (protons) which are also released as a by-product of nucleotide incorporation (Karow 2010a). A proprietary semi-conductor chip, which is essentially a miniature pH meter, is divided into wells in which the sequencing reactions take place. If a nucleotide is incorporated in a particular well, a single H^+ is release into solution and a concomitant change in acidity (pH) is detected as a voltage shift by sensors. The magnitude of the pH change can be related to the number of molecules of a particular base incorporated. Currently this system does not detect single molecules and amplification is required prior to sequencing, but the synthesis reaction is detected in real-time and no modified reagents are required.

A majority of other platforms currently under development for using electronic detection are not based on the sequencing-by-synthesis method, but on an entirely new method using either biological or solid state nanopores. These technologies monitor changes in electrical current as DNA strands or individual bases pass through a nanopore. The sequencing chamber is divided into two sub-chambers by a synthetic membrane or some other septa. Each sequencing chamber contains a single nanopore penetrating the septum providing a single channel between the two chambers. The nanopores themselves can either be small holes in an inorganic membrane (solid-state nanopores), such as silicon nitride (Aksimentiev et al. 2004) or grapheme (Garaj et al. 2010), or modified natural channel proteins like α-haemolysin (Howorka et al. 2001; Olasagasti et al. 2010) embedded in a lipid bilayer or synthetic membrane. Nanopore sequencing technologies are based on one of two approaches—either the DNA strand itself passes through the nanopore (strand sequencing), or individual bases are cleaved from the target DNA and fed sequentially through the nanopore. A voltage is placed across the membrane to drive the translocation of negatively charged DNA molecules through the pore. As DNA bases pass through the pore, the current is blocked and since each base blocks the current by a different amount the strand composition can be determined.

Numerous companies are currently developing nanopore-based DNA sequencing platforms, including Oxford Nanopore Technologies, NABSys, base4innovation, and IBM/Roche. Whilst this technique is extremely promising, there are still challenges to be overcome both technical, such as controlling the passage of bases though the nanopore to allow sequencing of consecutive bases, and those related to system performance, such as pore shelf life and parallelisation (Branton et al. 2008; Kircher and Kelso 2010). Perhaps the most advanced to date is the platform

under development by Oxford Nanopore Technologies, which uses α-haemolysin nanopores modified with a cyclodextrin ring covalently bound in the barrel. The DNA is digested by an exonuclease and the individual bases are drawn through the pore one at a time driven by an electrical potential (Astier et al. 2006; Wu et al. 2007). It has been demonstrated that this system can also distinguish 5-methyl cytosine thus enabling direct methylation analysis (Clarke et al. 2009).

Recently an alternative method for detection and identification of nucleotides has been described: here the transverse conductivity of the molecule is measured as it passes between two electrodes embedded in a solid state nanopore (Tsutsui et al. 2010, 2011). The authors suggest alternative methods for translocation of the DNA through the nanopore such as 'magnetic tweezers' (Peng and Ling 2009).

Atomic

Another novel technique for DNA uses transmission electron microscopy to directly visualise strands of DNA that have been suitably modified with heavy metal atoms to distinguish the bases (Krivanek et al. 2010). This method is being developed and commercialised by several companies, including Halcyon Molecular and ZSGenetics. The use of scanning tunnelling microscopy to sequence DNA molecules has also been described (Tanaka and Kawai 2009).

Targeting methods

Although NGS platforms have massively increased throughput, sequencing the entire genome is still neither practical nor affordable for most clinical applications. Moreover whole genome sequencing may not be desirable in a medical setting for reasons of interpretation and reporting. Consequently, many studies employ new sequencing technologies for targeted sequencing of specific regions of interest as opposed to whole genomes. This ranges from the analysis of gene families or large regions that are associated with a specific disease or pharmacogenetic effects, to the analysis of all coding exons in the genome (the 'exome') (Teer and Mullikin 2010; Majewski et al. 2011). Since NGS platforms sequence the entire input sample, it is necessary to have a method of selecting the desired DNA before sequencing. There are a three general approaches to targeting (Summerer 2009; ten Bosch and Grody 2008)—PCR-based methods, circularisation methods and hybridisation capture. The relative advantages and disadvantages of these approaches are contrasted in Table 2.

The current method of targeting for capillary sequencing is the polymerase chain reaction (PCR) (Saiki et al. 1985). This can equally be used for preparation of targets for NGS but owing to the massive capacity of these platforms very large numbers of PCRs are required to fill a run. The processing required can be limited by utilising multiplex PCR or long range PCR (Fredriksson et al. 2007; Varley and Mitra 2008). Commercial solutions to this problem include the RainStorm technology from RainDance Technologies which uses an emPCR approach to simultaneously amplify up to 4,000 short DNA sequences in separate microdroplets (Tewhey et al. 2009), and the Access Array from Fluidigm which uses proprietary microfluidics to setup an array of 2,304 PCRs (48 samples × 48 assays). It should be noted that neither of these systems is actually multiplex as the individual reactions are separated—this enables much higher levels of parallelisation than are achievable in a single reaction.

Circularisation methods are designed to resolve the interference issues that limit the level of multiplexing achievable by standard PCR and are suitable for targeting small to medium sized regions of interest. Several approaches have been demonstrated including gene collector (Fredriksson et al. 2007) gene selector (Dahl et al. 2005, 2007) and connector inversion probes (Akhras et al. 2007) but all are essentially based on padlock and molecular inversion probes (Krishnakumar et al. 2008; Li et al. 2009). The basic principle is that large panels of target molecules can be selected and circularised in a single reaction using specially designed probes containing universal sequences. The reaction is then subjected to

Table 2 Advantages and disadvantages of different chemical targeting approaches

Method	Advantages	Disadvantages
PCR	High sensitivity, specificity, reproducibility and uniformity	High cost, low throughput, and cannot be used for large regions or a very large number of genes
Circularisation	Low cost (if many samples), easy to use, high sensitivity and specificity	Uniformity and sensitivity depends on design of probes. Cannot be used for a very large number of genes
Hybrid capture	Medium cost, easy to use, high sensitivity and specificity. Can target large sections of DNA and large numbers of genes	Uniformity and sensitivity depends on design of probes. Array design may be rather inflexible

Source: Mamanova et al. (2010)

exonuclease digestion which degrades all the unwanted DNA but leaves the targets untouched since, being circles, they have no ends. The target sequences can then be amplified using the universal sequences to generate suitable target material for sequencing.

The final method of targeting is hybridisation, which is based on the same principle as DNA microarray technology. Oligonucleotide probes are used to pull-down sequences of interest from whole, fragmented genomic DNA. Unwanted DNA is then washed off, and the captured material eluted and prepared for sequencing. The capture capacity ranges from a few Mb up to the entire exome (Hodges et al. 2007; Porreca et al. 2007) using two general methodologies: conventional solid state arrays (on-array capture) and paramagnetic beads (in-solution capture) (Albert et al. 2007; Chou et al. 2010; Gnirke et al. 2009). A number of custom hybridisation platforms are available including Agilent, Roche Nimblegen and Illumina.

The method of choice is dependent on application; in particular target size, type of target and sample number, required performance, ease of use and costs (Albert et al. 2007; Mamanova et al. 2010). Ideally the targeting method should allow enrichment of multiple different loci independent of their size, sequence composition or spatial distribution, and should be amenable to automation so that it can match the sequencing capacity. However, the current approaches have their own biases (see Table 2), which relate both to the types of sequences that they are able to capture and their ease of use. Key issues, which apply to all methods with varying degrees are uniformity, efficiency of coverage and off-target capture. Whilst these methods are continually being improved, it may ultimately be more cost-effective to sequence a whole genome, computationally masking regions of the genome that are irrelevant to a particular clinical question, and target analysis only to regions with proven clinical significance.

Performance metrics

Rather than review the current performance of each platform (which can be found on each of the manufacturer's websites and at http://www.molecularecologist.com/next-gen-fieldguide/), we outline some of the factors that affect performance and influence the utility of all whole genome sequencing technologies.

Analytical accuracy, systematic errors and quality of base calls

In addition to amplification errors (which will be eliminated by single molecule sequencing), all sequencing methodologies suffer from both random and systematic errors. The raw accuracy of the sequencing process and quality of base calling are critically important factors, particularly for clinical diagnostics. A quality score representing of the probability that the base is called correctly is assigned to each base (These are generally given on a logarithmic scale so that Q10 would be 10% probability of miscall Q20 is 1% probability, Q30 is 0.1% probability etc). Factors affecting the quality score include signal intensity, background noise in the reaction itself or generated by the instrument and crosstalk between clusters. Errors can include overcalls and undercalls (insertions or deletions of bases from the sequence) as well as miscalls (incorrect base assigned) (Albert et al. 2007; Brockman et al. 2008). Different sequencing technologies are prone to different systematic errors, which influence their utility for different applications; for example, accurately sequencing homopolymeric regions can be difficult using pyrosequencing due to intermediate fluorescence signal intensities resulting from the incorporation of n identical nucleotides.

Sanger sequencing has a low (but non-zero) error rate of around 10^{-4} to 10^{-5} for single calls (one error per 1,000–10,000 bases), but the accuracy for detecting heterozygous variants is much more difficult to assess and is almost certainly context dependent to some extent. When it comes to detection of a low level variant, for example mosaic or somatic mutation, the limit of detection in terms of minor allele representation is only around 20% for Sanger sequencing. Current NGS platforms have a somewhat higher raw error rate of around 10^{-2} to 10^{-3} (depending on read length), but this is easily offset by increasing read depth (i.e. consensus accuracy). In fact the desired accuracy can essentially be determined by altering the read depth appropriately. In contrast to Sanger sequencing detection of low level variants is basically limited by the raw error rate and would be substantially below 0.1%.

Read depth, genome coverage and uniformity

The read depth or depth of coverage refers simply to the number of times a base is sequenced in a single run of the machine. The required read depth varies depending upon the specific application and level of certainty required for the result. However, coverage of the genome is non-uniform due to factors such as repetitive elements, non-uniform targeting and variable GC content, which affects both amplification and sequencing efficiency (Dohm et al. 2008). For diagnostic purposes it is necessary to increase overall coverage to ensure the regions with least coverage meet the desired standard; any that do not should be failed. This can be very costly in terms of capacity, particularly where coverage is very variable. The theoretical depth required to detect a heterozygous variation with particular probability of success can be calculated. For re-sequencing

applications mapping the reads is guided by a reference sequence and theoretically requires much lower coverage (8–12x), than assembling genomes de novo (25–70x) (Schuster 2008). However, in practice, a depth of coverage of around 20× at each base is required for confident variant calling (Bentley et al. 2008).

Read length (number of bases per read)

Read length is an important factor in certain applications, such as sequencing through repetitive regions, identifying genomic rearrangements and getting short range haplotype information. In addition, longer reads make alignment to a reference sequence substantially easier by reducing the number of potential matches. Current second generation NGS platforms achieve reads length of 35–400 bases (Metzker 2010; ten Bosch and Grody 2008), but this is rapidly improving. It is anticipated that many the third generation platforms will have substantially longer read lengths. For example Oxford Nanopore Technologies, Pacific Biosciences and Life Technologies claim that their new platforms will have read lengths in excess of 1 kb and claims beyond this are not infrequent.

Sample multiplexing

Factors such as run capacity, sample multiplexing, run time and cost all have a major impact on the suitability of a particular platform for a particular application or laboratory. These factors vary substantially between machines, applications and chosen sequencing protocol. Whilst NGS has significantly reduced the per-base cost of sequencing, cost per test savings will only be realised if the capacity of the instrument is effectively used. In many cases the format of the experiment does not require the full capacity of a run for a single sample so methods for analysing multiple samples in a single run are important. Several methods are available to achieve such sample multiplexing (ten Bosch and Grody 2008). For targeted sequencing, it is possible to mix multiple different tests so that results from each specific test relate to only one individual patient. However, this is only effective if the reads are mapped to specific regions of interest and in many cases it is preferable to use the whole genome as a reference as this guards against non-specific targeting. In addition, many sequencing platforms allow physical separation between samples, for example by dividing the flow cell in to a number of channels. Finally, DNA 'barcode tags' can be added to the ends of DNA fragments during initial preparation. These are sequenced along with the fragment during the run and serve to identify the source of each sequence read during analysis (Binladen et al. 2007; Meyer et al. 2007).

Reagent and instrument cost

Reagent cost for sequencing has plummeted over the last decade, from a cost of around $500/Mb for Sanger sequencing reagents, to less than $0.50/Mb for reagents on the newest NGS platforms (Wetterstrand 2011). However, the sequencing machines themselves are often fairly expensive ranging from US$ 0.2–1 million. With ever increasing capacity the output from sequencing runs becomes greater and greater. The cost of handling and storing all this data should not be treated lightly—ultimately this is likely to be a more significant cost than generating the data.

Data analysis and interpretation

Massively parallel sequencing generates an enormous volume of data, the analysis of which requires substantial computational power, purpose-built bioinformatics tools and accurate databases of genomic variation to aid interpretation. The informatics pipeline for human genome resequencing using NGS technology can broadly be divided into three analytical steps:

1. *Primary analysis*: *base calling*—converting light signal intensities into a sequence of nucleotides. This is generally performed automatically by software on the sequencing machine itself and each call is associated with a raw quality score.
2. *Secondary analysis*: *alignment and variant calling*—mapping DNA reads to an annotated reference sequence and determining the extent of variation from the reference.[2] Because it is often not possible to unambiguously align a read to a unique position in the reference genome, particularly allowing for variation between the reference and the sample genome, a mapping quality score may be used to measure the likelihood that a read is mapped correctly. Numerous algorithms and software packages have been developed for this process (Flicek and Birney 2009; Magi et al. 2010) which is becoming increasingly automated. Various dedicated software packages have also been developed specifically for cancer genome assembly and variant calling, which take into account factors such as genetic heterogeneity in the sample (Ding et al. 2010; Magi et al. 2010). In the final stage of the alignment phase, sequence data are annotated with

[2] The process of alignment is substantially harder for de novo genome assembly, as there is no reference sequence against which DNA reads can be compared and mapped. Therefore specialist genome assembly methodologies and software have been developed, which are no longer directly relevant to human genome sequencing.

structural and functional biological information and visualised through a graphical interface or genome browser.

3. *Tertiary analysis*: *interpretation*—analysing and filtering variants to assess their inheritance, uniqueness, and likely functional impact (Kuhlenbñumer et al. 2010). This process requires comparison against databases of genomic variation (Kuntzer et al. 2010) (including both normal and pathogenic variants) and algorithms for evaluating the likely pathogenicity of a particular mutation [e.g. by assessing haploinsufficiency (Huang et al. 2010) of large deletions or loss of function variants (MacArthur and Tyler-Smith 2010), predicting the effect of amino acid substitutions caused by non-synonymous coding variants (Ng and Henikoff 2006)]. Although many software packages already exist for this process, accurate interpretation of the effect of genomic variants in an individual is still in its infancy, and more purpose-build packages will need to be developed to allow clinical diagnostic use. Meaningful clinical interpretation is likely to remain a major challenge for the foreseeable future.

A vast array of software packages are now available, both commercial and open source (i.e. freely available). Whilst open source options may offer a more flexible final solution, building a pipeline involves linking multiple software units, each performing a specific task, and remains the domain of dedicated bioinformaticians. The scope and applicability of integrated commercial packages is now enabling laboratory scientists to directly analyse their own data although it should be noted that significant time and effort will be necessary in order to acquire an appropriate level of understanding of the processes and how they affect the final results. A useful reference for available packages can be found at: http://seqanswers.com/forums/showthread.php?t=43.

Service providers

A number of national and international centres offer both data production and analysis services. In the UK the MRC has funded the establishment of four regional sequencing hubs which are primarily intended to support small and medium sized research projects, and the Wellcome Trust Sanger Institute continues to undertake large-scale sequencing research projects. In addition to research facilities in numerous countries international providers include Complete Genomics, a US company established in 2005 with the specific aim of providing a comprehensive human DNA service for pharmaceutical and academic research, and BGI (formally Beijing Genomics Institute), the first citizen-managed, non-profit research institution in China with probably the largest sequencing capacity in the world. It is unclear what effect these integrated service providers will have on the future of human whole genome sequencing, and currently most research and diagnostic laboratories both produce and analyse their own data.

Conclusion

The era of affordable genome resequencing is almost upon us, opening opportunities for both research and medical diagnostics. Exciting clinical applications of NGS and human genomes include

- Multi-gene diagnostic panels (Morgan et al. 2010)
- Achieving a molecular diagnosis for rare genetic diseases (Lupski et al. 2010; Ng et al. 2010a, b; Worthey et al. 2011; Vissers et al. 2010)
- Tissue matching and HLA-typing (Bentley et al. 2009; Gabriel et al. 2009; Lind et al. 2010)
- Non-invasive prenatal diagnosis (Chiu et al. 2008; Fan et al. 2008; Lo et al. 2010)
- Quantifying the burden of disease from solid tumours (Leary et al. 2010; McBride et al. 2010) and
- Cancer genome profiling leading to stratified treatment regimens (Campbell et al. 2010; Diamandis et al. 2010; Stratton et al. 2009)

RNA sequencing (RNA-seq) and chromatin immunoprecipitation (ChIP) sequencing can also be used to study gene expression and for detection of somatic mutations, gene fusions, and other non-mutational events, an understanding of which can have an impact on management of diseases such as cancer (Cowin et al. 2010; Robison 2010).

However, numerous barriers to clinical translation still exist, including: validation of the technology; standardisation of the analysis pipeline; integration of information from the numerous databases of genomic variation; building a robust evidence base to allow clinical interpretation of novel variants; developing a service delivery infrastructure that can capitalise upon the high-throughput advantages of new sequencing technologies; providing an appropriately skilled health care workforce to deal with genomic medicine; and addressing the numerous ethical, legal and social implications of sequencing, storing and accessing whole genomes. These issues will need to be addressed before human whole genome resequencing can be used routinely in the clinic.

Acknowledgments The content of this paper forms part of a PHG Foundation Report on the implications of whole genome sequencing for health (downloadable at www.phgfoundation.org). The authors wish to thank the project team at the PHG Foundation for their invaluable feedback on this work. The PHG Foundation is the

working name of the Foundation for Genomics and Population Health, a charitable company registered in England and Wales, charity No. 1118664 company No. 5823194.

References

Akhras MS, Unemo M, Thiyagarajan S, Nyren P, Davis RW, Fire AZ, Pourmand N (2007) Connector inversion probe technology: a powerful one-primer multiplex DNA amplification system for numerous scientific applications. PLoS One 2:e915

Aksimentiev A, Heng JB, Timp G, Schulten K (2004) Microscopic kinetics of DNA translocation through synthetic nanopores. Biophys J 87:2086–2097

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4:903–905

Astier Y, Braha O, Bayley H (2006) Toward single molecule DNA sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5′-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. J Am Chem Soc 128:1705–1710

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira CR, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Catenazzi CE, Chang S, Neil CR, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott FW, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling NB, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris PD, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva RA, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna SJ, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59

Bentley G, Higuchi R, Hoglund B, Goodridge D, Sayer D, Trachtenberg EA, Erlich HA (2009) High-resolution, high-throughput HLA genotyping by next-generation sequencing. Tissue Antigens 74:393–403

Binladen J, Gilbert MT, Bollback JP, Panitz F, Bendixen C, Nielsen R, Willerslev E (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS One 2:e197

Bowers J, Mitchell J, Beer E, Buzby PR, Causey M, Efcavitch JW, Jarosz M, Krzymanska-Olejnik E, Kung L, Lipson D, Lowman GM, Marappan S, McInerney P, Platt A, Roy A, Siddiqi SM, Steinmann K, Thompson JF (2009) Virtual terminator nucleotides for next-generation DNA sequencing. Nat Methods 6: 593–595

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di VM, Garaj S, Hibbs A, Huang X, Jovanovich SB, Krstic PS, Lindsay S, Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss JA (2008) The potential and challenges of nanopore sequencing. Nat Biotechnol 26:1146–1153

Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res 18:763–770

Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, Stebbings LA, Morsberger LA, Latimer C, McLaren S, Lin ML, McBride DJ, Varela I, Nik-Zainal SA, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Griffin CA, Burton J, Swerdlow H, Quail MA, Stratton MR, Iacobuzio-Donahue C, Futreal PA (2010) The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature 467:1109–1113

Chiu RWK, Chan KCA, Gao Y, Lau VYM, Zheng W, Leung TY, Foo CHF, Xie B, Tsui NBY, Lun FMF, Zee BCY, Lau TK, Cantor CR, Lo YMD (2008) Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci USA 105:20458–20463

Chou LS, Liu CS, Boese B, Zhang X, Mao R (2010) DNA sequence capture and enrichment by microarray followed by next-generation sequencing for targeted resequencing: neurofibromatosis type 1 gene as a model. Clin Chem 56:62–72

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 4:265–270

Cowin PA, Anglesio M, Etemadmoghadam D, Bowtell DDL (2010) Profiling the cancer genome. Annu Rev Genomics Hum Genet 11:133–159

Dahl F, Gullberg M, Stenberg J, Landegren U, Nilsson M (2005) Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. Nucl Acids Res 33:e71

Dahl F, Stenberg J, Fredriksson S, Welch K, Zhang M, Nilsson M, Bicknell D, Bodmer WF, Davis RW, Ji H (2007) Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc Natl Acad Sci USA 104:9387–9392

Diamandis M, White NMA, Yousef GM (2010) Personalized medicine: marking a new epoch in cancer patient management. Mol Cancer Res 8:1175–1187

Ding L, Wendl MC, Koboldt DC, Mardis ER (2010) Analysis of next generation genomic data in cancer: accomplishments and challenges. Hum Mol Genet 19(R2):R188–R196

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36:e105

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden

D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science 323: 133–138

Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR (2008) Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci USA 105: 16266–16271

Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. Nat Methods 6:S6–S12

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GR, Ruffier M, Schuster M, Sobral D, Spudich G, Tang YA, Trevanion S, Vandrovcova J, Vilella AJ, White S, Wilder SP, Zadissa A, Zamora J, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJ, Parker A, Proctor G, Vogel J, Searle SM (2011) Ensembl 2011. Nucleic Acids Res 39:D800–D806

Fredriksson S, Baner J, Dahl F, Chu A, Ji H, Welch K, Davis RW (2007) Multiplex amplification of all coding sequences within 10 cancer genes by gene-collector. Nucleic Acids Res 35:e47

Gabriel C, Danzer M, Hackl C, Kopal G, Hufnagl P, Hofer K, Polin H, Stabentheiner S, Pröll J (2009) Rapid high-throughput human leukocyte antigen typing by massively parallel pyrosequencing for high-resolution allele identification. Hum Immunol 70: 960–964

Garaj S, Hubbard W, Reina A, Kong J, Branton D, Golovchenko JA (2010) Graphene as a subnanometre trans-electrode membrane. Nature 467:190–193

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27: 182–189

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z (2008) Single-molecule DNA sequencing of a viral genome. Science 320: 106–109

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet 39:1522–1527

Howorka S, Cheley S, Bayley H (2001) Sequence-specific detection of individual DNA strands using engineered nanopores. Nat Biotechnol 19:636–639

Huang N, Lee I, Marcotte EM, Hurles ME (2010) Characterising and predicting haploinsufficiency in the human genome. PLoS Genet 6:e1001154

Karow J (2010a) Ion torrent systems presents $50,000 electronic sequencer at AGBT. Genomeweb in sequence

Karow J (2010b) Life tech details real-time single-molecule tech at AGBT; combines Qdots with FRET-based detection. Genomeweb in sequence

Kircher M, Kelso J (2010) High-throughput DNA sequencing—concepts and limitations. Bioessays 32:524–536

Krishnakumar S, Zheng J, Wilhelmy J, Faham M, Mindrinos M, Davis R (2008) A comprehensive assay for targeted multiplex amplification of human DNA sequences. Proc Natl Acad Sci USA 105:9296–9301

Krivanek OL, Chisholm MF, Nicolosi V, Pennycook TJ, Corbin GJ, Dellby N, Murfitt MF, Own CS, Szilagyi ZS, Oxley MP, Pantelides ST, Pennycook SJ (2010) Atom-by-atom structural and chemical analysis by annular dark-field electron microscopy. Nature 464:571–574

Kuhlenbñumer G, Hullmann J, Appenzellerm S (2010) Novel genomic techniques open new avenues in the analysis of monogenic disorders. Hum Mutat. doi:10.1002/humu.21400

Kuntzer J, Eggle D, Klostermann S, Burtscher H (2010) Human variation databases. Database. doi:10.1093/database/baq015

Leary RJ, Kinde I, Diehl F, Schmidt K, Clouser C, Duncan C, Antipova A, Lee C, McKernan K, De La Vega FM, Kinzler KW, Vogelstein B, Diaz LA, Velculescu VE (2010) Development of personalized tumor biomarkers using massively parallel sequencing. Sci Transl Med 2:20ra14

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows Wheeler transform. Bioinformatics 25:1754–1760

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858

Li JB, Gao Y, Aach J, Zhang K, Kryukov GV, Xie B, Ahlford A, Yoon JK, Rosenbaum AM, Zaranek AW, LeProust E, Sunyaev SR, Church GM (2009) Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. Genome Res 19: 1606–1615

Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, Slavich L, Walker R, Hsiao T, McLaughlin L, D'Arcy M, Gai X, Goodridge D, Sayer D, Monos D (2010) Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum Immunol 71:1033–1042

Lo YMD, Chan KCA, Sun H, Chen EZ, Jiang P, Lun FMF, Zheng YW, Leung TY, Lau TK, Cantor CR, Chiu RWK (2010) Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci Transl Med 2: 61ra91

Luckey JA, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB, Smith LM (1990) High speed DNA sequencing by capillary electrophoresis. Nucl Acids Res 18:4417–4421

Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DCY, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, Stankiewicz P, Halperin JJ, Yang C, Gehman C, Guo D, Irikat RK, Tom W, Fantin NJ, Muzny DM, Gibbs RA (2010) Whole-genome sequencing in a patient with charcot-marie-tooth neuropathy. N Engl J Med 362:1181–1191

MacArthur DG, Tyler-Smith C (2010) Loss-of-function variants in the genomes of healthy humans. Hum Mol Genet 19:R125–R130

Magi A, Benelli M, Gozzini A, Girolami F, Torricelli F, Brandi M (2010) Bioinformatics for next generation sequencing data. Genes 1:294–307

Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N (2011) What can exome sequencing do for you? J Med Genet 48(9):580–589

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7: 111–118

Mardis ER (2011) A decade's perspective on DNA sequencing technology. Nature 470:198–203

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE,

McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

McBride DJ, Orpana AK, Sotiriou C, Joensuu H, Stephens PJ, Mudie LJ, Hämäläinen E, Stebbings LA, Andersson LC, Flanagan AM, Durbecq V, Ignatiadis M, Kallioniemi O, Heckman CA, Alitalo K, Edgren H, Futreal PA, Stratton MR, Campbell PJ (2010) Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. Genes Chromosom Cancer 49:1062–1069

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46

Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. Nucl Acids Res 35:e97

Morgan JE, Carr IM, Sheridan E, Chu CE, Hayward B, Camm N, Lindsay HA, Mattocks CJ, Markham AF, Bonthron DT, Taylor GR (2010) Genetic diagnosis of familial breast cancer using clonal sequencing. Hum Mutat 31:484–491

Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61–80

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ (2010a) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30–35

Ng SB, Nickerson DA, Bamshad MJ, Shendure J (2010b) Massively parallel sequencing and rare disease. Hum Mol Genet 19:R119–R124

Olasagasti F, Lieberman KR, Benner S, Cherf GM, Dahl JM, Deamer DW, Akeson M (2010) Replication of individual DNA molecules under electronic control using a protein nanopore. Nat Nanotechnol 5(11):798–806

Peng H, Ling XS (2009) Reverse DNA translocation through a solid-state nanopore by magnetic tweezers. Nanotechnology 20:185101

Pettersson E, Lundeberg J, Ahmadian A (2009) Generations of sequencing technologies. Genomics 93:105–111

Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J (2007) Multiplex amplification of large sets of human exons. Nat Methods 4:931–936

Robison K (2010) Application of second-generation sequencing to cancer genomics. Brief Bioinform 11:524–534

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230:1350–1354

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat Methods 5:16–18

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH, Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679

Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458:719–724

Summerer D (2009) Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing. Genomics 94:363–368

Tanaka H, Kawai T (2009) Partial sequencing of a single DNA molecule with a scanning tunnelling microscope. Nat Nanotechnol 4:518–522

Teer JK, Mullikin JC (2010) Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet 19:R145–R151

ten Bosch JR, Grody WW (2008) Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. J Mol Diagn 10:484–492

Tewhey R, Warner JB, Nakano M, Libby B, Medkova M, David PH, Kotsopoulos SK, Samuels ML, Hutchison JB, Larson JW, Topol EJ, Weiner MP, Harismendy O, Olson J, Link DR, Frazer KA (2009) Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol 27:1025–1031

Tsutsui M, Taniguchi M, Yokota K, Kawai T (2010) Identifying single nucleotides by tunnelling current. Nat Nanotechnol 5:286–290

Tsutsui M, Matsubara K, Ohshiro T, Furuhashi M, Taniguchi M, Kawai T (2011) Electrical detection of single methylcytosines in a DNA oligomer. J Am Chem Soc 133:9124–9128

Tucker T, Marra M, Friedman JM (2009) Massively parallel sequencing: the next big thing in genetic medicine. Am J Hum Genet 85:142–154

Varley KE, Mitra RD (2008) Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. Genome Res 18:1844–1850

Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, van Bon BW, Hoischen A, de Vries BB, Brunner HG, Veltman JA (2010) A de novo paradigm for mental retardation. Nat Genet 42:1109–1112

Voelkerding KV, Dames SA, Durtschi JD (2009) Next-generation sequencing: from basic research to diagnostics. Clin Chem 55:641–658

Wetterstrand KA (2011) DNA sequencing costs: data from the NHGRI large-scale genome sequencing programme. Available at www.genome.gov/sequencingcosts. Accessed 1 March 2011

Worthey E, Mayer A, Syverson G, Helbling D, Bonacci B, Decker D, Serpe J, Dasu T, Tschannen M, Veith R, Basehore M, Broeckel U, Tomia-Mitchell A, Arca M, Casper J, Margolis D, Bick D, Hessner M, Routes J, Verbsky J, Jacob H, Dimmock D (2011) Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med 13(3):255–262

Wu HC, Astier Y, Maglia G, Mikhailova E, Bayley H (2007) Protein nanopores with covalently attached molecular adapters. J Am Chem Soc 129:16142–16148