# The application of next-generation sequencing technologies to drug discovery and development

**Peter M. Woollard[1], Nalini A.L. Mehta[2], Jessica J. Vamathevan[1], Stephanie Van Horn[3], Bhushan K. Bonde[1] and David J. Dow[2]**

[1] Computational Biology, Drug Discovery, GlaxoSmithKline Research and Development, Gunnels Wood Road, Stevenage, SG1 2NY, UK
[2] Molecular and Cellular Technologies, Platform Technology and Science, GlaxoSmithKline Research and Development, Gunnels Wood Road, Stevenage, SG1 2NY, UK
[3] Molecular and Cellular Technologies, Platform Technology and Science, GlaxoSmithKline Research and Development, 1250 South Collegeville Road, Collegeville, PA 19426, USA

**Next-generation sequencing (NGS) technologies represent a paradigm shift in sequencing capability. The technology has already been extensively applied to biological research, resulting in significant and remarkable insights into the molecular biology of cells. In this review, we focus on current and potential applications of the technology as applied to the drug discovery and development process. Early applications have focused on the oncology and infectious disease therapeutic areas, with emerging use in biopharmaceutical development and vaccine production in evidence. Although this technology has great potential, significant challenges remain, particularly around the storage, transfer and analysis of the substantial data sets generated.**

## Introduction

The past 60 years have seen a remarkable increase in knowledge around the human genome and genetic code, from the discovery of the structure of DNA, to the invention of DNA sequencing, culminating in the publication of the human genome sequence in 2001 [1–4].

The advent of the human genome project helped to foster the development of faster and cheaper DNA sequencing. Sanger sequencing [2], also now known as 'first generation sequencing', has dominated the past few decades [5], with estimated costs of up to US$3 billion to sequence the human genome (http://www.genome.gov/11006943). The need to conduct large-scale sequencing projects more economically has led to the rapid development of a variety of next-generation sequencing (NGS) technologies (Table 1). With NGS technology, a human genome can now be sequenced for only tens of thousands of dollars, with some recent claims as low as US$5000 (Five Thousand Bucks for Your Genome, http://www.technologyreview.com/biomedicine/21466/MIT).

*Corresponding authors:* Woollard, P.M. (Peter.M.Woollard@gsk.com), Dow, D.J. (David.J.Dow@gsk.com)

Sequencing technologies are developing at a rapid pace and some commentators have likened the rate of decrease in cost per base as a 'genetics Moore's Law' [6]. The pending arrival of third-generation sequencing promises the ability to sequence single DNA molecules in real time at an evermore decreasing cost, bringing researchers closer to sequencing a human genome for US$1000 (NHGRI Seeks Next Generation of Sequencing Technologies, http://www.genome.gov/12513210).

Pharmaceutical companies are embracing NGS technologies and there are many potential impacts throughout the drug discovery process (Table 2). Currently, R&D groups use a variety of high-throughput technologies, including transcriptomics using microarrays, genome-wide association studies (GWAS), metabolomic modelling, Yeast 2 Hybrid (Y2H) assays, proteomics, high-throughput chemistry screening and *in silico* techniques [7,8]. NGS has the potential to augment or complement these existing technologies [9–12] (Table 2).

This review focuses on the various applications of NGS to drug discovery and development rather than details of the platforms themselves, for which the reader is directed to reviews in this area [5,13]. The first section outlines NGS technologies and current


1359-6446/06/$ - see front matter  doi:10.1016/j.drudis.2011.03.006

**TABLE 1**

**Glossary of terms**

| *Term* | Definition | Example |
|---|---|---|
| **Base** | Nucleotide base found in DNA or RNA | A (adenine), C (cytosine), G (guanine) or T (thymine) |
| **ChIP-Seq** | Sequencing DNA footprints to which proteins (e.g. histones or transcription factors) are bound | Generating whole-genome maps of transcription factor binding sites |
| **Copy number variation (CNV)** | Duplication or deletion of a regions of genomic DNA, often involving genes | CYP2D6 ultra-rapid metaboliser allele |
| **Deep sequencing** | Sequencing of specific regions of DNA at high coverage | Detection of low-level viral drug resistance mutations at extremely low frequencies |
| **Epigenetics** | Heritable changes independent of DNA sequence | Histone modifications associated with gene transcription status, for example, methylation on histone H3 at position K4 is found in transcriptionally active genes |
| **Exome sequencing** | Enrichment and sequencing of the coding regions of genes (exons) | Exome sequencing to identify mutations in rare mendelian conditions |
| **RNA-Seq** | Sequencing of the transcriptome, which contains all transcripts within the cell, by first converting the RNA to complementary DNA (cDNA) | Comparison of the transcriptome in diseased versus normal tissues |
| **Sanger sequencing** | Pioneered by Nobel Laureate Fred Sanger; a method of determining the sequence of a DNA molecule using di-deoxy terminator nucleotides | The technology behind the sequencing of the human genome |
| **SNP** | A single nucleotide difference between DNA samples | Can affect protein function (e.g. ADME genes) or can act as a genetic marker in association studies |
| **Transcript** | An RNA molecule that has been formed by copying a section of genomic DNA; can contain gene sequences | A gene is transcribed into mRNA and is then translated into a peptide before becoming a functional protein |

applications. We then review the use of the technology in different parts of the drug discovery and development process. Finally, we discuss the challenges of using NGS followed by a summary and future perspectives.

## What is next-generation sequencing technology?

The development of NGS platforms represents a great advance in technology. In comparison to Sanger sequencing, the NGS platforms are able to produce orders of magnitude more sequence data through massively parallel processes, which result in substantial quantities of data at a low cost per base (Box 1).

NGS can be performed on several commercial platforms, including Roche 454 (http://www.454.com/), Illumina (http://www.illumina.com/) and the SOLiD platform from Life Technologies (http://www.lifetechnologies.com). Broadly speaking, the processes followed by these platforms are similar and include: template preparation (genomic or cDNA) by shearing to create fragment libraries, massive parallel clonal amplification of individual DNA molecules and then sequencing to generate short reads. Finally, an informatic alignment of the short reads is performed to reconstruct the starting template sequence (Fig. 1) [5]. Although the readout from the machine is DNA sequence, these platforms afford the opportunity to conduct multiple types of experiment (Table 2).

Third-generation platforms under development include zero-mode waveguides, semiconductor and nanopore sequencing technologies. These platforms promise even larger and faster data generation, although are a few years away from robustly achieving this ([14,15], http://www.iontorrent.com).

## Target identification and validation

Identifying potential therapeutic drug targets and validating their suitability can be a complex process involving many different experimental platforms. NGS can be used in the very early stages of target identification to provide detailed genomics data in the same way as traditional tools, such as microarrays. RNA-Seq can be used to perform differential gene expression studies of diseased versus normal tissue; this identifies genes and pathways that might be important in disease pathology, which can inform the selection of new targets for intervention [9]. Data from NGS of whole human exomes are being successfully applied to identify mutations in genes underlying rare Mendelian disorders, which could also inform target identification [16]. In addition, using NGS has enabled the resolution of genetic linkage studies, a finding that has potential in identifying new drug targets from complex trait genetics studies [16,17].

NGS has also proved to be a useful tool in the characterisation of therapeutically relevant mutations in mice [18], which can often contribute important information to target validation. Through exome or targeted region capture sequencing, underlying mutations can be rapidly isolated without the requirement for lengthy genetic mapping studies [18].

## Ultra-high throughput screening

NGS has also found application in high throughput compound screening using a method called 'encoded library technology' (ELT) [19]. Compounds with covalently bound short oligonucleotide labels are built up iteratively through many rounds of synthesis. Hits are identified by sequencing the oligonucleotide tag

**TABLE 2**

**Broad applications of NGS to drug discovery**

| *Applications* | Pros of NGS | Cons of NGS | Alternatives | Refs |
|---|---|---|---|---|
| **Mutation detection: personalised medicine** | Can sequence large genome regions to identify efficacy markers | Initial setup and running cost for NGS | Large-scale Sanger sequencing technology | [64] |
| **ChIP-Seq: target identification and/or validation and compound profiling for epigenetics** | Enables study of epigenetic targets at the whole-genome level | Many possible algorithms for data analysis and complex data interpretation | ChIP-on-chip assay using microarray-based technology | [44] |
| **CNV: target identification, personalised medicine, for example, cancer** | Uncovers all types of CNV; no a priori assumptions about location of CNVs required | Large and complex rearrangements might not be detected | Comparative genomic hybridisation | [35,65, 66] |
| **Exome sequencing: target identification and/or drug resistance studies, biomarker discovery** | Identify rare variants, using deep sequence coverage | Sequence variation in non-coding regions and introns not detected | Large-scale Sanger sequencing technology | [16] |
| **RNA-Seq: target identification and/or validation by studying differential gene or miRNA expression between normal and diseased tissue** | Detects alternative splicing and low expression transcripts; has large dynamic range | Bias during library preparation can result in over-representation of transcript 3′ ends | Microarray-based technology | [11,12, 67] |
| **HITS-CLIP: study of RNA–protein interactions** | Enables study of RNA–protein interactions | A relatively new application; not many studies to date | Microarray-based technology | [68] |
| **Ribosome profiling: target identification by measuring protein translation rates using sequencing to identifying ribosomal footprints** | Potential to enable analysis of the whole cell proteome by sequencing | A relatively new application, not many studies to date | Conventional proteomics technologies, for example, mass spectrometry | [69] |
| **Small RNA sequencing (e.g. miRNA): biomarker discovery** | Assay and quantify all small RNAs present | Data analysis complex owing to the presence of isomiRs | Fluorescently labelled PCR techniques | [22,40] |
| **Bacterial genome sequencing: target identification, validation and diagnostics to identify new strains and mechanisms of drug resistance** | Small genomes mean many strains can be sequenced per run | Short read alignment can result in gaps in coverage owing to repeat sequence | Large-scale Sanger sequencing technology | [45,47, 48,70] |

attached to the compound, which contains distinct sequences relating to which chemical groups make up the compound. Researchers are thus able to screen very large chemical libraries (800 million compounds) against targets by affinity selection [19].

## Biomarkers

With the challenge issued by the US Food and Drug Administration's (FDA) Critical Path Initiative to make better use of genomic technologies, biomarkers are set to have an ever more important role in the drug development process ([20], http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative). NGS could have a central role in the discovery of new genomic biomarkers, owing to the many different types of experiment that can be performed on a single machine. From a single sample, it could be possible to generate complementary data sets from genome DNA sequencing, miRNA, both transcriptome sequencing and transcriptome quantification to epigenetic changes of DNA methylation, histone post-translational modifications and even protein translation (ribosome profiling) (Table 2). The challenge will be around developing data analysis tools that could simultaneously analyse across these vast data sets, looking for biomarker signatures.

A further interesting application of the technology is 'personalised' biomarkers, which have recently been described in the field of oncology [21]. In this study, a patient's tumour was analysed to generate a tumour-specific biomarker. Genomic rearrangements were identified specific to the tumour, which were not present in the patient's normal somatic tissue. Digital PCR assays were then designed across rearrangement breakpoints to provide a sensitive tumour-specific biomarker which was successfully used to monitor residual disease following treatment [21].

Small RNAs could provide a further application of NGS in biomarker discovery [22]. MicroRNAs (miRNAS) are implicated in the control of protein translation and are present in blood plasma. NGS could be used to assay tissues or blood plasma to generate whole-genome miRNA profiles, which could then be mined for biomarker signatures.

## Personalised medicine and pharmacogenetics

The study of the influence of genes and genetic variation on the response to medicines in terms of both efficacy and side effects is called pharmacogenetics and the tailoring of treatment based on an individual's genetic make up is often referred to as personalised medicine or targeted therapy. Ultimately, the aim is that a patient

**BOX 1**

### Comparing Sanger and next generation sequencing

- The initial preparation of the DNA sample is more labour intensive for NGS than for Sanger, but the amount of sequence data obtained per sample is substantially more.
- The number of sequencing reads from a single instrument per run is of the order of thousands with Sanger, but millions to billions with NGS; for example, a bacterial genome can be sequenced in a single run in days using NGS, versus months using Sanger sequencing.
- Read lengths from Sanger sequencing are up to 900 bp, but in NGS vary from 30 to 500 bp depending on the platform [5].
- DNA sequencing costs have been driven down by NGS [5], base pair per dollar costs show a consistent 19-months doubling time reduction for Sanger sequencing. For NGS, the equivalent figure is approximately 5-months doubling time cost reduction [55].
- NGS can detect somatic mutations at $\leq$1% frequency, whereas Sanger sequencing has significantly less sensitivity.
- The greater versatility of NGS is illustrated in generating whole-genome data sets, such as miRNA and ChIP-Seq; Sanger sequencing lacks this capability.

would be prescribed a medicine for a disease with the best personal profile: efficacious, the right dose, with the least risk of adverse effects [23].

## Genetic variation

Based upon small- or medium-scale candidate gene studies, or genome-wide scans involving genotyping of single nucleotide polymorphisms (SNPs), many examples of genetic associations with medicine response have been reported to date [24,25]. A recent study of patients with bipolar disorder showed more differential benefit for valproate depending on their genotype of XBP1-116C/G [26] and similarly for the *DRD4* gene in smoking cessation [27]. A frequent proposed use of pharmacogenetics is to stratify patient populations in clinical trials into responders and non-responders according to their genetic profiles, which could lead to smaller and less expensive subsequent clinical trials [8]. By contrast, others note that pharmacogenetics has not translated significantly into drug discovery, development or clinical practice [28]. NGS promises to facilitate this area of research by uncovering all of the common and rare genetic variation in human populations and, indeed, the 1000 Genomes Project has made great progress to date towards this goal [29]. With a comprehensive genetic map of all human variation produced by NGS, researchers will be able to perform more detailed experiments to detect genetic variation underlying the response to medicines.

## Clinical diagnostics

Array-based panels for detecting variants of genes involved in absorption, distribution, metabolism, and excretion (ADME) of pharmaceutical compounds are already available [30], thus some personal adverse drug reactions (ADRs) can be predicted, which can guide drug selection. NGS has a large potential to be used in clinical diagnostics [13,31]; however, one study reported that using NGS found only 60–70% of the variants identified by Sanger sequencing [31]. The reasons were mainly due to short read lengths, low read depth coverage or homopolymers; the authors noted that the first two issues are currently being resolved.

A project is underway to develop a platform for analysing clinical tissue samples using NGS (CLC Bio press release, September 8, 2010; http://www.clcbio.com/index.php?id=1590). The application areas include molecular diagnostics research and re-analysis of preclinical trials where drugs have failed despite relatively high rates of efficacy. NGS has also recently been used in clinical gene therapy studies to determine cell fate [32].

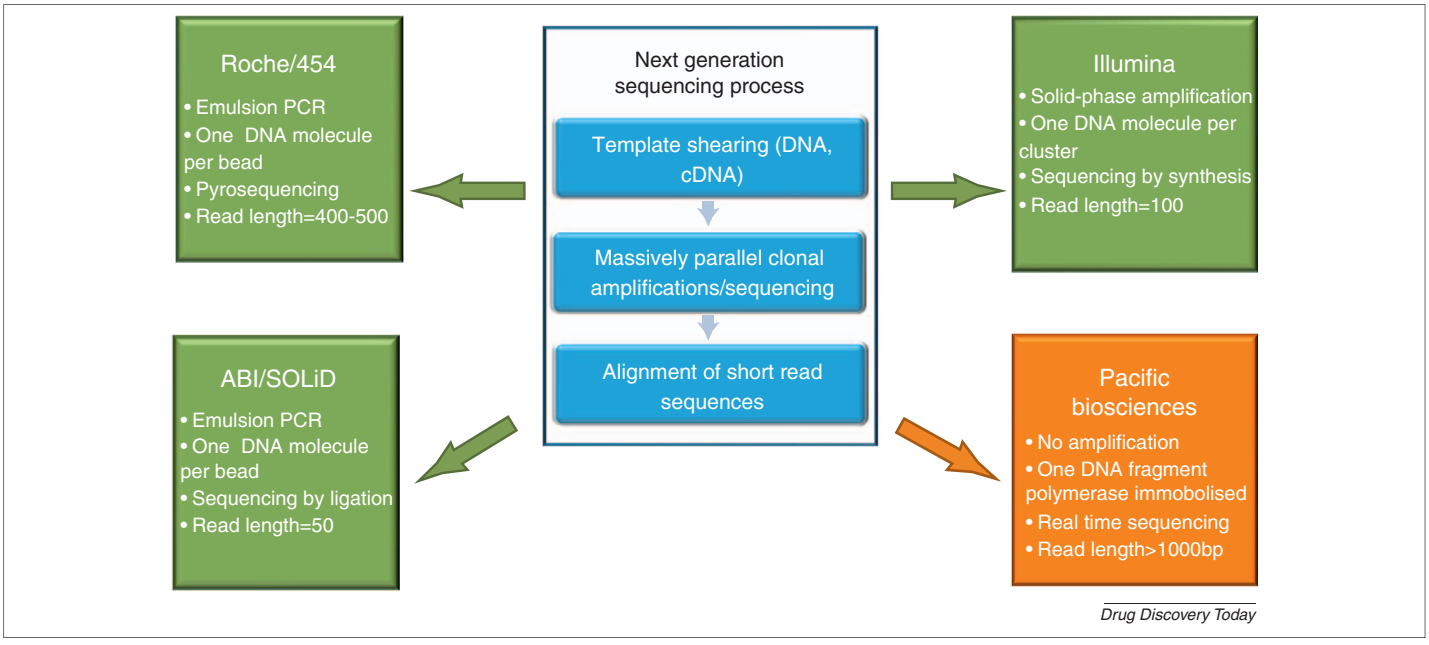


**FIGURE 1**

What is NGS? The three most used platforms in green are depicted with one of the emerging third generation platforms [5].

## Applications to oncology

NGS is proving a powerful tool in characterising tumour cells at the genomic, transcriptomic and epigenetic levels. It can identify both genomic structural rearrangements, such as deletion, duplication and inversion events, and single base changes; it is also very sensitive and can detect the presence of low-level somatic mutations against a germline background [33]. These properties make the method ideally suited to the study of tumour cells, which by their nature contain many such mutational events. Cataloguing all mutations, copy number aberrations and somatic rearrangements in an entire cancer genome at base pair resolution can now be performed in a matter of weeks [21]. Researchers are already reporting the complete genome sequence of several tumour types, such as breast and colorectal cancer, along with potential biomarkers [21]. The identification and cataloguing of tumour mutations can implicate genes or pathways in the tumorigenesis process, suggesting new potential drug targets, or identifying tumour-specific mutations forming the basis for targeted therapies.

## Resequencing

NGS has already been applied to numerous resequencing studies, which have led to whole-genome sequencing of complete normal and cancer genomes [34,35]. This now enables rapid identification of patient-specific rearrangements in solid tumours [36]. Personalised biomarkers, as noted previously, are now being developed to detect the presence of tumour-specific genomic rearrangements in plasma samples from patients [21].

Chemotherapy response and resistance have both been studied using NGS in several cancers, including ovarian cancer and tongue adenocarcinoma [37,38]. Correlating observed gene mutations and amplifications with known cancer pathways can provide putative oncogenesis mechanisms. Tumour sequencing by NGS has the potential to aid clinical decision making, disease understanding and to guide the choice of drug treatment [38]. Following treatment, NGS has also been used to analyse DNA in serum from patients with breast cancer tumours to detect residual disease [39].

## RNA sequencing

NGS is also finding application in the study of short RNAs. A comprehensive study of miRNA in acute myeloid leukaemia was performed using NGS, with novel findings of differentially expressed miRNAs [40]. Transcriptome sequencing using Illumina and 454 technologies has also been found to be a powerful tool for detecting novel gene fusions in cancer cell lines and tissues [41].

## Epigenetics

Epigenetics is increasingly being studied using NGS, particularly in oncology, as epigenetic regulation has a role in the development and progression of tumours [42]. Histone modification and methylation of DNA are two important epigenetic mechanisms that regulate the transcriptional status of genes. Using ChIP-Seq technology, post-translational modifications of histones and the location of transcription factors can be studied at the whole-genome level [43], whereas methylated DNA immunoprecipitation (meDIP) and bisulphite protocols can be used to study the methylation of DNA itself [44].

## Applications to vaccine development

The high sensitivity of NGS in detecting sequences present at low levels means NGS can be used as a quality-control tool to detect adventitious viruses in vaccines, using a metagenomics approach [45]. Indeed, polio, rubella, measles, mumps and rotavirus vaccines have already been analysed in this way and companies are already offering this approach as a screening service to pharmaceutical and biotech companies [46].

Researchers can use NGS early in the development of vaccines to ensure that any contaminants in the processes, if present, are detected early. There are three main stages of vaccine production where sequencing could prove valuable: cell substrates (the cells used to produce the vaccine); vaccine seeds (the viruses inserted into the cells acting as a master stock for the vaccine); and mammalian media (often used in vaccine production). Transcriptome sequencing can be used to detect active or replicating viruses and also any latent viruses that are present. Finally, using whole-genome sequencing, NGS could have a greater role in future vaccine development processes by helping to decipher why some individuals experience adverse effects and immune responses to a given vaccine. Indeed, NGS could also be used to monitor the whole development of vaccines evaluating their safety, efficacy and differences in host response [45].

## The impact of NGS in infectious diseases

An increasing abundance of microbial genomics data has been generated by NGS technologies. The application of NGS is already widespread in antibacterial and antiviral drug discovery. Examples include resequencing of target genes to understand genetic variation, target discovery by sequencing isogenic-sensitive and -resistant mutant strains, as well as sequencing of SNPs and resistance markers to track antibiotic resistance in epidemiological studies [47]. Much of the current microbial genomics NGS literature stems from academic and private research institutions, with few publications by pharmaceutical companies or affiliates to date. Understanding mutations that cause drug resistance is important in anti-infective research.

## Investigating drug resistance

NGS has been clearly demonstrated to detect low-abundance HIV drug-resistant variants [48]. In a study by Le *et al.*, low-level variants were found in 22 subjects, in comparison with Sanger sequencing, which found variants in only three subjects. Most of these newly found resistance mutations correlated with historical antiretroviral use, which could be of interest to clinicians in planning subsequent antiretroviral regimens for patients who have undergone a range of treatments [48]. Another study found that low levels of CXCR4-using virus increased after treatment with maraviroc, the CCR5 antagonist, demonstrating the effect of drug pressure on the viral structure population [49]. NGS was used in antibacterial research to track the spread of drug resistance, where it was shown that resistance was maintained at low levels (0.4%) for the drug levofloxacin in 1106 *Streptococcus pneumoniae* strains over a period of five years [50].

## Identifying drug targets

NGS has also been successfully used to identify novel drug targets. A breakthrough in TB research owing to an early application of

NGS enabled the characterisation of mutations in whole genomes to identify the unknown target of a diarylquinoline (DARQ) lead compound, R207910 [51]. By comparing whole-genome sequences from both compound-sensitive and -resistant strains of *Mycobacterium tuberculosis* and *Mycobacterium smegmatis* and looking for sequence differences, the drug target of R207910 was identified as the proton pump of ATP synthase. Complementation assays subsequently confirmed the finding. This study demonstrated the utility of NGS in whole-cell *in vitro* screens to identify novel drug targets. The Centre for Disease Control and Prevention led an NGS approach that enabled the identification of a new Ebola virus where real-time PCR assays for the Zaire and Sudan Ebola viruses failed to identify the presence of this virus [52]. This has great implications for designing diagnostic assays for haemorrhagic fever disease in humans, as well as an impact on the ongoing efforts to develop effective anti-virals and vaccines.

## Biopharmaceuticals

NGS is beginning to find application in the development of biopharmaceuticals, particularly in the area of antibody library generation. Several groups have investigated how the massively parallel nature of the technology can also complement antibody library generation.

In one study, NGS was used alongside real-time PCR in an attempt to remodel the traditional phage display pipeline [53]. The objective was to streamline the process to remove rate-limiting steps. Overnight bacterial cultures and transducing counting unit assays were replaced with DNA extractions and quantitative real-time PCR, in conjunction with NGS of phage vector inserts. This resulted in a streamlined protocol that was less labour intensive, significantly faster and cheaper. In addition to the process improvements, the substantial nature of NGS data sets could afford a more accurate assay of antibody library diversity.

Interactome profiling using an NGS readout was used to identify protein interactors of a multifunctional enzyme [10]. The combination of cDNA phage display and NGS to identify captured cDNA inserts identified several interactors, some of which were novel and were subsequently confirmed in functional assays. Intriguingly, it was possible to identify and refine the binding domain with a known interactor, the glycoprotein fibronectin. In comparison with traditional methods involving enzyme-linked immunosorbent assay (ELISA) and Sanger sequencing, this new method proved faster and produced more information. Although initially applied to the study of protein–protein interactions, the method might also be applicable to the study of antibody–antigen binding.

## Challenges in storage, transfer and analysis of NGS data

The size of the data sets produced from an NGS experiment present challenges in the storage, transfer and analysis of data. Due consideration should be given to the installation and access to adequate IT solutions for data storage and transfer, both before the implementation of an NGS platform and for future projections of increasing amounts of data. Richter and Sexton [54] discuss that the initial investment in the instrument is accompanied by an almost equal investment in upgrading the informatics infrastructure of the institution, hiring staff to analyse the data produced by the instrument, and storing the data for future use. In practice, there are significant data transfer issues, including network bandwidth, owing to the large data files. It has been demonstrated that NGS data throughput growth is exceeding computer hardware developments [55].

## Software and cloud computing

Many open source and commercial short read analysis tools are available, some from the platform vendor, each with its own set of strengths and weaknesses. The software choice for a particular project depends on whether the reference genome is known, the biological application and what hardware and software are available [56]. Massive parallelism reduces assembly problems to days or hours [57,58], but this might require specific hardware. For smaller groups who are not in major genome sequencing centres, cloud computing is a potential solution [55,57,59]. Cloud computing provides the ability to rent and perform computation on standard, commodity computer hardware over the Internet. Effectively, this outsourcing of many of the data management and analysis challenges is an attractive solution; and examples are emerging covering various experimental applications and software [60]. Data still need to be transferred to clouds and there are, of course, important considerations, such as data security, to be put into place before implementation. Further data analysis challenges are reviewed elsewhere [61].

## Standardisation

In line with FDA regulations regarding pharmaceutical R%D processes, it would be beneficial to standardise NGS protocols and data. Indeed, a key community effort initially for microarrays and now including NGS was initiated by the FDA: the microarray quality control (MAQC), and the sequencing equivalent, sequencing quality control (SEQC) projects (http://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/ [62]).

The robustness and reproducibility of NGS have also been investigated. A recent study shows that greater than two biological replicates provides only a marginal gain of information [63]. If the technology proves reproducible across the many applications, this could mean that fewer biological replicates are required, reducing the cost per experiment.

## Conclusion

NGS offers many advantages over Sanger sequencing (Box 1), particularly high throughput, lower cost and greater versatility. In the pharmaceutical industry, it facilitates large knowledge-gaining experiments that could not be financially justified or were not even possible five years ago, including metagenomics to compare different disease states or patient variability, genome sequencing of model organisms or ELT compound library screening techniques.

There is no doubt that the use of NGS technology in drug discovery is gaining momentum. There already exist multiple potential applications of the technology (Table 2) across the drug discovery pipeline, from early-stage target identification through to phase IV studies (Fig. 2). It will be interesting to see how much this disruptive technology replaces or augments existing platforms, or indeed reshapes parts of the drug discovery pipeline, eliminating steps or providing new ways of working. Early use has

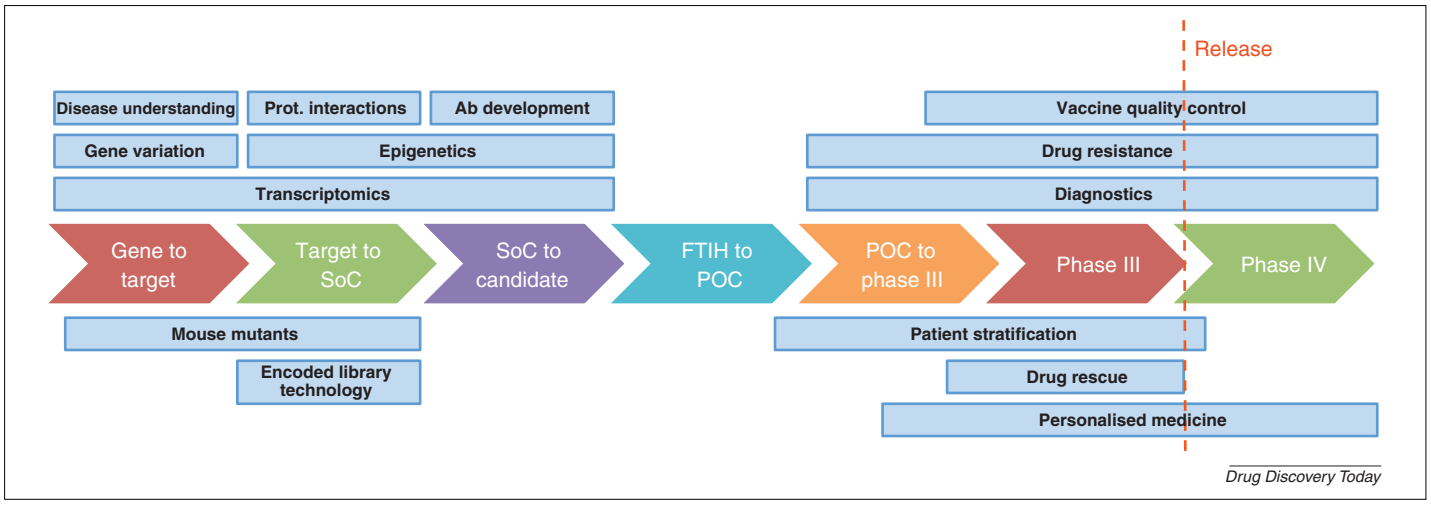


**FIGURE 2**

Summary of NGS applications in drug discovery.

been evident in the oncology and infectious disease therapy areas, with emerging applications in biopharmaceuticals, personalised medicine, biomarker discovery and vaccine development.

Implementation of the technology requires careful planning, particularly in providing hardware, software and skills to transfer, analyse and store these very large data sets, even if sequencing and some analysis is outsourced. The informatics challenges will grow as the iterations of the technology develop through third- and perhaps even fourth-generation technologies. Key to successful application will be the ability to analyse ever-more increasing and detailed data sets to extract biological meaning, and this will require the development of new and sophisticated algorithms.

With the potential to execute multiple experiment types on one platform and the increasing amounts of human and model organism data being generated, NGS might find application in systems biology [23]. There is also much potential in clinical diagnostics. The drug discovery and development industry has many challenges, including the increasing costs of R&D and the provision of differentiated medicines to satisfy the needs of patients, regulators and payers. These, along with emerging themes such as diagnostics and personalised medicine, will be highly impacted by NGS, with its multiple diverse applications and decreasing cost as the technology continues to advance.

## Conflict of interest

All authors are current employees of GlaxoSmithKline.

## Acknowledgements

We thank Philippe Sanseau, Ganesh Sathe, Israel Gloger and Julie Huxley-Jones for helpful comments and critical reading of the manuscript.

## References

1 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
2 Sanger, F. *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A* 74, 5463–5467
3 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
4 Watson, J.D. and Crick, F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171, 737–738
5 Metzker, M.L. (2010) Sequencing technologies – the next generation. *Nat. Rev. Genet.* 11, 31–46
6 Pettersson, E. *et al.* (2009) Generations of sequencing technologies. *Genomics* 93, 105–111
7 Auffray, C. *et al.* (2009) Systems medicine: the future of medical genomics and healthcare. *Genome Med.* 1, 2
8 Foot, E. *et al.* (2010) Pharmacogenetics – pivotal to the future of the biopharmaceutical industry. *Drug Discov. Today* 15, 325–327
9 Shendure, J. (2008) The beginning of the end for microarrays? *Nat. Methods* 5, 585–587
10 Di, N.R. *et al.* (2010) Rapid interactome profiling by massive sequencing. *Nucleic Acids Res.* 38, e110
11 Matsumura, H. *et al.* (2010) High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. *PLoS ONE* 5, e12010
12 Hashimoto, S. *et al.* (2009) High-resolution analysis of the 5′-end transcriptome using a next generation DNA sequencer. *PLoS ONE* 4, e4108
13 Voelkerding, K.V. *et al.* (2009) Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* 55, 641–658
14 Eid, J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138
15 Derrington, I.M. *et al.* (2010) Nanopore DNA sequencing with MspA. *Proc. Natl. Acad. Sci. U. S. A* 107, 16060–16065
16 Ng, S.B. *et al.* (2010) Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.* 42, 30–35
17 Bowden, D.W. *et al.* (2010) Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in the *ADIPOQ* gene in the IRAS Family Study. *Hum. Mol. Genet.* 19, 4112–4120
18 D'Ascenzo, M. *et al.* (2009) Mutation discovery in the mouse using genetically guided array capture and resequencing. *Mamm. Genome* 20, 424–436
19 Clark, M.A. *et al.* (2009) Design, synthesis and selection of DNA-encoded small-molecule libraries. *Nat. Chem. Biol.* 5, 647–654
20 Coons, S.J. (2009) The FDA's critical path initiative: a brief introduction. *Clin. Ther.* 31, 2572–2573
21 Leary, R.J. *et al.* (2010) Development of personalized tumor biomarkers using massively parallel sequencing. *Sci. Transl. Med.* 2 20ra14
22 Lee, L.W. *et al.* (2010) Complexity of the microRNA repertoire revealed by next generation sequencing. *RNA* 16, 2170–2180

23 Auffray, C. *et al.* (2010) Predictive, preventive, personalized and participatory medicine: back to the future. *Genome Med.* 2, 57
24 Hetherington, S. *et al.* (2002) Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. *Lancet* 359, 1121–1122
25 Takeuchi, F. *et al.* (2009) A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet.* 5, e1000433
26 Kim, B. *et al.* (2009) Preliminary evidence on the association between XBP1-116C/G polymorphism and response to prophylactic treatment with valproate in bipolar disorders. *Psychiatry Res.* 168, 209–212
27 Leventhal, A.M. *et al.* (2010) Dopamine D4 receptor gene variation moderates the efficacy of bupropion for smoking cessation. *Pharmacogenomics J.* DOI: 10.1038/tpj.2010.64.
28 Gervasini, G. *et al.* (2010) Pharmacogenetic testing and therapeutic drug monitoring are complementary tools for optimal individualization of drug therapy. *Eur. J. Clin. Pharmacol.* 66, 755–774
29 Durbin, R.M. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073
30 Cuyas, E. *et al.* (2010) Errors and reproducibility of DNA array-based detection of allelic variants in ADME genes: PHARMAchip. *Pharmacogenomics* 11, 257–266
31 Hoppman-Chaney, N. *et al.* (2010) Evaluation of oligonucleotide sequence capture arrays and comparison of next-generation sequencing platforms for use in molecular diagnostics. *Clin. Chem.* 56, 1297–1306
32 Paruzynski, A. *et al.* (2010) Genome-wide high-throughput integrome analyses by nrLAM-PCR and next-generation sequencing. *Nat. Protoc.* 5, 1379–1395
33 Bueno, R. *et al.* (2010) Second generation sequencing of the mesothelioma tumor genome. *PLoS ONE* 5, e10612
34 Ley, T.J. *et al.* (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456, 66–72
35 Kim, P.M. *et al.* (2008) Analysis of copy number variants and segmental duplications in the human genome: evidence for a change in the process of formation in recent evolutionary history. *Genome Res.* 18, 1865–1874
36 McBride, D.J. *et al.* (2010) Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. *Genes Chromosomes Cancer* 49, 1062–1069
37 Cheng, L. *et al.* (2010) Analysis of chemotherapy response programs in ovarian cancers by the next-generation sequencing technologies. *Gynecol. Oncol.* 117, 159–169
38 Jones, S.J. *et al.* (2010) Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome Biol.* 11, R82
39 Beck, J. *et al.* (2010) Next generation sequencing of serum circulating nucleic acids from patients with invasive ductal breast cancer reveals differences to healthy and nonmalignant controls. *Mol. Cancer Res.* 8, 335–342
40 Ramsingh, G. *et al.* (2010) Complete characterization of the microRNAome in a patient with acute myeloid leukemia. *Blood* 116, 5316–5326
41 Maher, C.A. *et al.* (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature* 458, 97–101
42 Balch, C. *et al.* (2009) Minireview: epigenetic changes in ovarian cancer. *Endocrinology* 150, 4003–4011
43 Neff, T. and Armstrong, S.A. (2009) Chromatin maps, histone modifications and leukemia. *Leukemia* 23, 1243–1251
44 Popp, C. *et al.* (2010) Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* 463, 1101–1105
45 Victoria, J.G. *et al.* (2010) Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. *J. Virol.* 84, 6033–6040
46 Onions, D. and Kolman, J. (2010) Massively parallel sequencing, a new method for detecting adventitious agents. *Biologicals* 38, 377–380
47 Brown, J.R. (2010) Next generation sequencing for antibacterial drug discovery. *Int. Drug Discov.* 5, 18–23
48 Le, T. *et al.* (2009) Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. *PLoS ONE* 4, e6079
49 Archer, J. *et al.* (2009) Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4 (CXCR4)-using HIV-1 with ultra-deep pyrosequencing. *AIDS* 23, 1209–1218
50 Davies, T.A. *et al.* (2006) Infrequent occurrence of single mutations in topoisomerase IV and DNA gyrase genes among US levofloxacin-susceptible clinical isolates of *Streptococcus pneumoniae* from nine institutions (1999–2003). *J. Antimicrob. Chemother.* 57, 437–442
51 Andries, K. *et al.* (2005) A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. *Science* 307, 223–227
52 Towner, J.S. *et al.* (2008) Newly discovered Ebola virus associated with hemorrhagic fever outbreak in Uganda. *PLoS Pathog.* 4, e1000212
53 As-Neto, E. *et al.* (2009) Next-generation phage display: integrating and comparing available molecular tools to enable cost-effective high-throughput analysis. *PLoS ONE* 4, e8338
54 Richter, B.G. and Sexton, D.P. (2009) Managing and analysing next-generation sequence data. *PLoS Comput. Biol.* 5, e1000369
55 Stein, L.D. (2010) The case for cloud computing in genome informatics. *Genome Biol.* 11, 207
56 Palmieri, N. and Schlotterer, C. (2009) Mapping accuracy of short reads from massively parallel sequencing and the implications for quantitative expression profiling. *PLoS ONE* 4, e6323
57 Langmead, B. *et al.* (2009) Searching for SNPs with cloud computing. *Genome Biol.* 10, R134
58 Simpson, J.T. *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19, 1117–1123
59 Baker, M. (2010) Next-generation sequencing: adjusting to data overload. *Nat. Methods* 7, 495–499
60 Langmead, B. *et al.* (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.* 11, R83
61 Fuller, C.W. *et al.* (2009) The challenges of sequencing by synthesis. *Nat. Biotechnol.* 27, 1013–1023
62 Shi, L. *et al.* (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* 28, 827–838
63 Rozowsky, J. *et al.* (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* 27, 66–75
64 Henn, B.M. *et al.* (2010) Fine-scale population structure and the era of next-generation sequencing. *Hum. Mol. Genet.* 19, R221–R226
65 Chiang, D.Y. *et al.* (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* 6, 99–103
66 Wood, H.M. *et al.* (2010) Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Res.* 38, e151
67 Cloonan, N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5, 613–619
68 Chi, S.W. *et al.* (2009) Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature* 460, 479–486
69 Morris, D.R. (2009) Ribosomal footprints on a transcriptome landscape. *Genome Biol.* 10, 215
70 Zagordi, O. *et al.* (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res.* 38, 7400–7409