Briefings in Bioinformatics, 20(5), 2019, 1795–1811

doi: 10.1093/bib/bby051
Advance Access Publication Date: 14 June 2019
Review article

Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application

Gaye Lightbody, Valeriia Haberland, Fiona Browne, Laura Taggart, Huiru Zheng, Eileen Parkes and Jaine K. Blayney

Corresponding author: Gaye Lightbody, School of Computing, Jordanstown Campus, Ulster University, Shore Road, Newtownabbey BT37 0QB, Co. Antrim, UK. E-mail: g.lightbody@ulster.ac.uk

Abstract

There has been an exponential growth in the performance and output of sequencing technologies (omics data) with full genome sequencing now producing gigabases of reads on a daily basis. These data may hold the promise of personalized medicine, leading to routinely available sequencing tests that can guide patient treatment decisions. In the era of high-throughput sequencing (HTS), computational considerations, data governance and clinical translation are the greatest rate-limiting steps. To ensure that the analysis, management and interpretation of such extensive omics data is exploited to its full potential, key factors, including sample sourcing, technology selection and computational expertise and resources, need to be considered, leading to an integrated set of high-performance tools and systems. This article provides an up-to-date overview of the evolution of HTS and the accompanying tools, infrastructure and data management approaches that are emerging in this space, which, if used within in a multidisciplinary context, may ultimately facilitate the development of personalized medicine.

Key words: high-throughput sequencing; personalized medicine; clinical translation; translational research; high-performance computing; grid computing; cloud computing

Gaye Lightbody is a Lecturer in Computing Science. Her research career has spanned areas of high-performance hardware architectures for signal and data processing systems, biosignal analysis and artificial intelligence.
Valeriia Haberland is a Senior Research Associate (PhD in Computer Science). Her research interests included high-performance computing and automated data matching. Currently, she is interested in causal inference and text mining.
Fiona Browne is a Lecturer in Computing Science. Her research interests focus on artificial intelligence and information fusion in multidisciplinary fields including Bioinformatics.
Laura Taggart is a Senior Scientist with a focus on genomics in the context of stratifying cancer patients with the aim of improving therapeutic outcomes for patients as well as a current interest in the development of liquid biopsies.
Huiru Zheng is a Professor of Computer Science. Her research interests include integrative data analytics in the field of systems biology and systems medicine.
Eileen Parkes is an Academic Clinical Lecturer in Medical Oncology. Her PhD led her from DNA damage to immunobiology, and she continues exploring this rich interface in translational and clinical research.
Jaine K. Blayney is a Lecturer in Translational Bioinformatics with research interests in data integration through compositional statistics approaches, and the development of disease-agnostic bioinformatics and biostatistics techniques to inform biomarker discovery and validation.

Submitted: 30 January 2018; **Received (in revised form):** 1 May 2018

Introduction

Over the past decade, there have been exponential advances in our capacity to sequence a human genome. As recently as 2016, it would have taken over a day [1]. Now, using current technology [2], it is possible to process a genome sequence within an hour [3, 4]. The development of high-throughput sequencing (HTS) technologies has been central to achieving this, with massively parallel sequencing offering larger throughput than the conventional Sanger sequencing [5] approach.

While advances have been made across all aspects of the sequencing workflow, the focus on platform development has made a significant contribution to driving down machine size and HTS costs while facilitating performance gains. In addition, this has been enhanced by reductions in both the cost of computational power and size, as expected through Moore's law [6]. However, since 2007 [7], the reduction in the sequencing cost per genome has surpassed Moore's law; thus, we are now in the era of the sub-$1000 genome. An extensive review of the past 10 years of HTS can be found in [8] along with additional technological solutions in [9, 10] and more recently in [11].

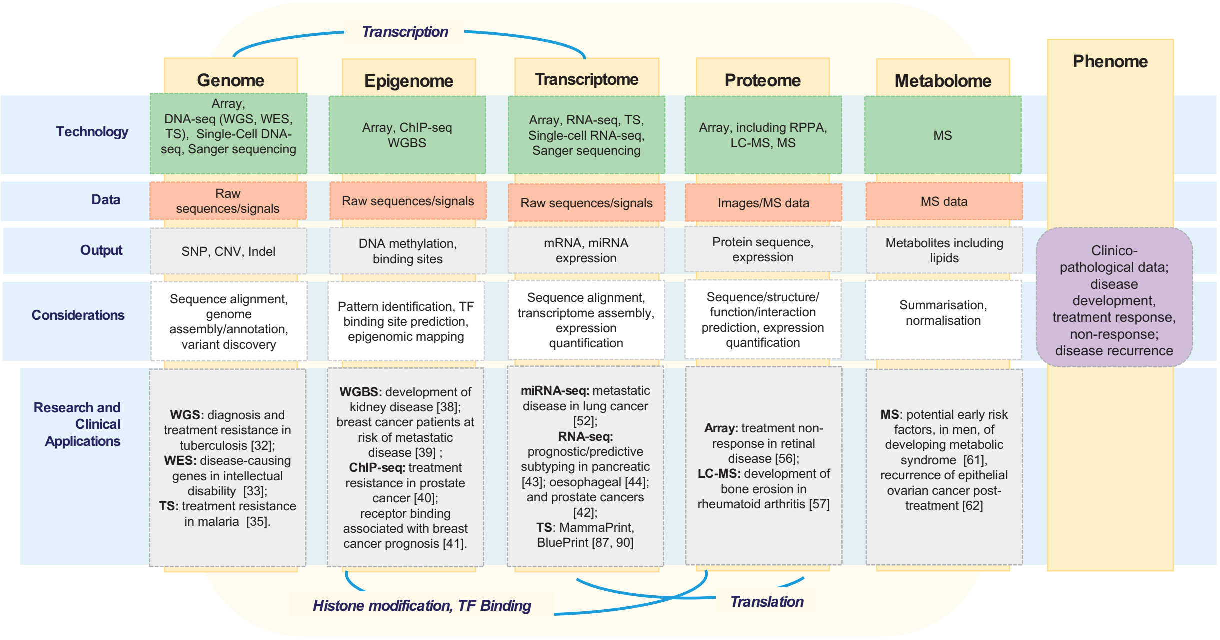
The decreasing costs of HTS have brought it within the reach of smaller laboratories, facilitating the generation of high-dimensional in-house data sets, with typical HTS devices producing over 100 gigabases (Gb) of reads in 24 h [12]. As with other examples of 'Big Data', the steps involved in the design, preprocessing, normalization and downstream analysis of HTS data are significant. Furthermore, there are substantial challenges presented, including sample collection and quality control, selection of HTS technology, to the integration of data sets across platforms and technologies. HTS data therefore present its own set of *in silico* and computational challenges, leading to a 'Data Deluge' [13] in which the emphasis has moved from data generation to the ability to store, access, share and analyse the data effectively. As reported by Sboner *et al.* [14], these additional elements contribute towards a more realistic assessment of the true cost of HTS use. In addition, there are also data governance and patient privacy implications, particularly resulting from the speed of change brought about by the application of HTS in clinical workflows [15, 16].

Considering these intersecting challenges within the biomedical domain, particularly with regard to clinical (and commercial) translation, HTS can be considered from the perspectives of four key stakeholders: biologists, clinicians and patients alongside bioinformaticians/computer scientists. Against this background, we consider common HTS bottlenecks that can be encountered at different workflow stages. We then present potential *in silico* and computational solutions, extending on the review in [17], and examine further rate-limiting issues that may in turn be raised. We therefore conclude with a discussion on the future role of HTS in facilitating biomedical research and its potential translation to clinical decision-making tools.

HTS: from biomedical research to clinical application

In the biomedical domain, HTS can be used to characterize biological markers (biomarkers), including genes and proteins, often derived from human tissue or blood, to understand disease development and progression and/or predict treatment response or patient survival [18]. Biomarkers can be classified into three categories: diagnostic (presence or absence of disease), predictive (how a patient responds to treatment) and prognostic (how long a patient survives post-intervention) [18].

Markers and drivers of disease development, progression and treatment response can be detected at the deoxyribonucleic acid (DNA), ribonucleic acid (RNA) or protein levels with a range of HTS techniques (Figure 1). We consider both biomarker and



	Genome	Epigenome	Transcriptome	Proteome	Metabolome	Phenome
Technology	Array, DNA-seq (WGS, WES, TS), Single-Cell DNA-seq, Sanger sequencing	Array, ChIP-seq WGBS	Array, RNA-seq, TS, Single-cell RNA-seq, Sanger sequencing	Array, including RPPA, LC-MS, MS	MS	
Data	Raw sequences/signals	Raw sequences/signals	Raw sequences/signals	Images/MS data	MS data	
Output	SNP, CNV, Indel	DNA methylation, binding sites	mRNA, miRNA expression	Protein sequence, expression	Metabolites including lipids	Clinico-pathological data; disease development, treatment response, non-response; disease recurrence
Considerations	Sequence alignment, genome assembly/annotation, variant discovery	Pattern identification, TF binding site prediction, epigenomic mapping	Sequence alignment, transcriptome assembly, expression quantification	Sequence/structure/function/interaction prediction, expression quantification	Summarisation, normalisation	
Research and Clinical Applications	**WGS:** diagnosis and treatment resistance in tuberculosis [32]; **WES:** disease-causing genes in intellectual disability [33]; **TS:** treatment resistance in malaria [35].	**WGBS:** development of kidney disease [38]; breast cancer patients at risk of metastatic disease [39] ; **ChIP-seq:** treatment resistance in prostate cancer [40]; receptor binding associated with breast cancer prognosis [41].	**miRNA-seq:** metastatic disease in lung cancer [52]; **RNA-seq:** prognostic/predictive subtyping in pancreatic [43]; oesophageal [44]; and prostate cancers [42]; **TS**: MammaPrint, BluePrint [87, 90]	**Array:** treatment non-response in retinal disease [56]; **LC-MS:** development of bone erosion in rheumatoid arthritis [57]	**MS**: potential early risk factors, in men, of developing metabolic syndrome [61], recurrence of epithelial ovarian cancer post-treatment [62]	

Figure 1. Summary schema of omics levels with associated technologies, data types, outputs, analytical considerations and research and clinical applications. Five levels or components, genome, epigenome, transcriptome, proteome and metabolome are presented, all of which can be considered with respect to the phenome (common patient characteristics). Key associations between omics levels are also represented, including transcription (between the genome and transcriptome), histone modification and TF-binding (connecting the epigenome with the proteome) and translation (from the transcriptome to the proteome).
Source: Adapted from: [19–21] and [22].

HTS applications at five key omics levels, genome, epigenome, transcriptome, proteome and metabolome. These levels are connected via genetic data transfer processes, including transcription, translation, binding and protein modification [23, 24]. As shown in Figure 1, each of these omics levels can be considered with respect to patient characteristics, such as risk of disease or response to a treatment, i.e. phenotypes.

At the genome (DNA) level, alterations in genes are analysed, e.g. single-nucleotide polymorphisms (SNPs), indels, copy number variations (CNVs) and fusion genes [25] (Figure 1). SNPs (equivalent to 'typos' in the genome), indels (insertion or deletion of bases in a sequence) and CNVs (where portions of the genome are repeated) have been linked to disease susceptibility. SNPs in the *VCAM-1* and *ARFGEF2* genes, a deletion in the *CTFR* gene and CNVs in the *HLA* gene were found to be associated with cases of sickle cell anaemia [26], cystic fibrosis (CF) [27] and rheumatoid arthritis [28], respectively. Fusion genes, when two individual genes form a hybrid gene, can also be associated with disease development, as with *TMPRSS2-ERG* in prostate cancer [29]. Microarrays [30], and more recently, DNA-sequencing (DNA-seq), comprising whole-genome sequencing (WGS), whole-exome sequencing (WES) and targeted sequencing (TS) have been used to study these alterations. WGS enables the interrogation of alterations in both the coding and non-coding regions of the genome [31] and has been used to identify multiple SNPs relating to the diagnosis of tuberculosis and treatment resistance [32]. WES is limited to coding regions, approximately 1% of the genome [31]. A more affordable option than WGS, WES may omit potentially informative gene regulation regions, though its use was well founded in a study of intellectual disability, as three novel disease-causing candidate genes were identified [33]. TS focuses on specific regions of the genome and is useful when prior information is known about the disease [34], e.g. in a study to understand resistance to first-line antimalarial therapy, TS identified six novel resistance-causing mutations [35].

Epigenomics encompasses the chemical modification, through internal or external factors, of DNA, which in turn can repress the corresponding gene expression, leading to disease or treatment resistance (Figure 1) [36]. Both microarray and sequencing technologies can be used to quantify DNA methylation status. Bisulphite conversion is necessary for both older (microarray) and newer sequencing-based technologies, facilitating the detection of methylated cytosines (one of four DNA component bases), though is a harsh process that can affect the quality of DNA for downstream analysis [37]. Whole-genome bisulphite sequencing (WGBS) was used to identify methylation of the *IFITM3* gene as a candidate in the development of kidney disease [38]. Legendre *et al.* [39] used WGBS to develop a blood-based methylation patterns that could be used to stratify breast cancer patients into metastatic disease risk groups. Chromatin immunoprecipitation sequencing (ChIP-seq) allows for the precise characterization of transcription factor (TF)-binding sites (location at which a protein binds to DNA to initiate transcription) and patterns of histone (a DNA packaging protein) modification, both of which can affect gene expression. Using this technology, advances in understanding the impact of the epigenome on the development of metastatic disease in patients with early prostate cancer were made [40]. Within oestrogen receptor-positive (ER+) breast cancer, the use of ChIP-Seq helped to identify the prognostic role of the gene *FOXA1* in facilitating ER-binding [41].

The transcriptome encompasses all RNA found in the cell (Figure 1). Messenger RNA (mRNA) is the most commonly studied form of RNA. The transcriptome, capturing the downstream signals from the genome and epigenome, has been used for molecular subtyping and studying drug response [42–45], applying both microarray and HTS technologies. For example, using microarray technology, four breast cancer subtypes associated with patient response to chemotherapy were defined based on a set of RNA patterns (PAM50) [46]. Other RNA types such non-coding RNAs and microRNA (miRNA) have also been described [47]. In particular, miRNA has been shown to be important in disease development and progression through gene regulatory functionality [48]. miRNAs have been associated with relapsing–remitting multiple sclerosis [49] and dormancy of the human immunodeficiency virus type 1 (HIV-1) in patients treated with antiretroviral therapy [50]. RNA can be studied through both microarray and RNA-Sequencing (RNA-Seq) with RNA-Seq also allowing for the discovery of additional modifications, e.g. fusion genes, similar to the genome level [51]. RNA-Seq has also been extensively used, often within a multi-omics or integrative context. This has resulted in the characterization of novel molecular subgroups associated with treatment response and/or survival in multiple cancer studies, including pancreatic [43], oesophageal [44], prostate [42] and cholangiocarcinoma [45]. microRNA sequencing (miRNA-Seq) has also been used in the identification of miRNAs that were significantly associated with remote metastatic disease in lung adenocarcinoma [52].

All the previous elements (genome, epigenome and transcriptome) contribute to the proteome, the set of proteins that comprise an organism (Figure 1) [53]. The sequence, structure and expression of proteins are encoded by the genome but can be altered at the transcriptional level with the potential for changes being introduced at translation [53]. In comparison with other omics levels, it is relatively poorly characterized [54]. Array-based methods, including reverse phase protein array (RPPA) and mass spectrometry (MS) technologies can be applied at this level [55]. Using an array-based technology, Velez *et al.* [56] identified protein targets for a tailored treatment of a patient with inflammatory disease of the retina, reversing sight loss. In addition to array-based methods, MS or liquid chromatography (LC)-MS can be used to study the sequence and structure of proteins, each having a unique weight (mass) fingerprint that can be used to identify their presence in a sample [55]. Liao *et al.* [57] used LC-MS to identify candidate proteins in samples obtained from rheumatoid arthritis patients with none erosion.

Metabolomics is the study of the chemical fingerprints that cellular processes leave behind (Figure 1), i.e. metabolites, which are small molecules, such as amino acids or lipids, resulting from the breakdown of proteins through protein–protein interactions [19]. Similar to proteins, metabolites are identified and studied by MS generating metabolite profiles. The study of metabolites is a well-established and important element in drug discovery, particularly the understanding of the metabolism of a drug and potential associated toxicities [58]. Lipidomics, the study of lipid levels, such as cholesterol and triglyceride, in blood and tissue is a fast-emerging sub-field within metabolomics [59, 60]. Using MS, Sales *et al.* [61] characterised a 'lipotype' in men that corresponded to a potential risk of developing metabolic syndrome. While, Ke *et al.* [62] discovered that in epithelial ovarian cancer, patients post-surgery, who had recurred were found to have high levels of lipid and amino acid metabolism.

Research applications of HTS

The research community's reliance on microarray technology is now being replaced by a welcoming endorsement of sequencing technologies. This trend can be seen in the work of the flagship The Cancer Genome Atlas (TCGA) consortium [63]. In 2008, the first TCGA publication in glioblastoma used Sanger sequencing and array-based technologies to analyse patient samples at the genomic, epigenomic and transcriptomic levels [64]. In 2017, WES, RNA-Seq and miRNA-Seq, in addition to array-based SNP, methylation and protein analysis were used in a study of uterine carcinosarcoma [65]. As predicted [66], sequencing, particularly, RNA-Seq technology is rapidly replacing microarray-based approaches, because of its technical superiority [67] and ability to derive novel biological insights [68]. Moreover, using data from TCGA, pan-cancer studies analysing data from 10 000 solid tumours identified the impact of important biologies such as impaired DNA damage response [69] and comprehensive immune biology across cancers [70]. Indeed, this subsequent improvement in understanding the driving biologies and potential vulnerabilities of cancers demonstrate the importance of HTS in advancing our understanding of disease.

Clinical applications and limitations of HTS

HTS has also been applied within clinical trial contexts including in the development of early cancer detection assays or tests and selection of new treatments for patients not responding to standard regimes [71, 72]. Three current clinical trials use TS at the genome level (Figure 1). The first, the STRIVE Study, is using sequencing to detect and analyse circulating cell-free nucleic acids, present in blood samples taken from patients who had undergone a screening mammogram to improve early detection of breast cancer [73]. Another trial, NCI-Match, has enrolled cancer patients (with solid tumours or lymphomas) that had received treatment, yet had progressed, to help determine drug repurposing options, thereby improving outcomes for cancer patients [74]. Another exploratory study, the Michigan Oncology Sequencing Project (Mi-ONCOSEQ) uses a multi-sequencing approach to stratify clinical trial-eligible patients, with metastatic or refractory cancers. Mi-ONCOSEQ also considers the bioethical issues surrounding genomic testing and results disclosure to patients and clinicians [75].

Reaching the clinical trial stage does not always result in success. Despite identifying a drug–target mutation in nearly half of patients enrolled, the MOSCATO trial [76] reported the ability to deliver this therapy in less than one quarter of patients, of whom 11% responded. The large numbers screened for a limited clinical response is an important shortcoming of current HTS approaches. Despite this, HTS approaches have already resulted in improved outcomes. Sequencing the genome of one exceptional responder in a failed clinical trial of everolimus in bladder cancer, an inhibitor of the gene *mTOR*, identified a mutation of a key *mTOR* regulator, the *TSC1* gene [77]. Further sequencing discovered this mutation in 8% of bladder cancers. Initiatives are now ongoing to sequence exceptional responders in clinical trials to identify other, currently unknown, targetable mutations [78], demonstrating the prospective potent impact of HTS.

Although HTS-based clinical trials may not always fulfil their original potential, crucially, platform and diagnostic acceptance of HTS by regulatory bodies has been forthcoming. In 2013, the Illumina MiSeqDx was the first HTS platform to be approved as an *in vitro* diagnostic tool by the Food and Drug Administration (FDA), alongside two Illumina diagnostic assays, the CF Clinical Sequencing and CF 139-Variant assays, both of which target the region around the *CFTR* gene at the genomic level, for screening and diagnosis purposes [79–81]. Later, in 2016, the FDA published draft guidance for the development of further HTS-based assays for rare inherited diseases [82]. Then, in 2017, the FDA approved a further three HTS-based *in vitro* diagnostic tests, including FoundationOne's companion diagnostic, F1CDx [83], Memorial Sloan Kettering Cancer Center's MSK-IMPACT [84] and Thermo Fisher Scientific's Oncomine Dx Target Test [85, 86]. Both F1CDx and MSK-IMPACT can detect sequence modifications in various cancers to identify patients who may benefit from a number of targeted therapies [83, 84]. Similarly, Thermo Fisher Scientific's Oncomine Dx Target Test also quantifies genomic sequence changes in tumours to guide treatment for non-small cell lung cancer [86].

Translating research into clinical applications

However, regulatory approval does not equate to a global clinical acceptance and uptake. A number of breast cancer predictive transcriptome-based tests were derived in the pre-HTS era, such as PAM50 (Prosigna, NanoString Technologies, United States) [46, 87] and MammaPrint (MammaPrint BluePrint, Agendia BV, The Netherlands) [88, 89], both of which were later developed into commercial tests, the latter using RNA-Seq. Both were approved by both the FDA for use in the United States and in the European Economic Area through the Conformité Européene (CE) mark [82, 90–92] and included in the updated clinical decision-making guidelines from the European Group on Tumor Markers [93]. However, the National Comprehensive Cancer Network [94], while acknowledging other tests were available, only referred to the possible use of the OncotypeDx assay (Genomic Health, CA, USA) [95], which was developed using an older, targeted, RNA-quantification technology, reverse transcription polymerase chain reaction. Understandably, there is still a sense of caution regarding the use of HTS in a clinical context [96, 97], with an argument that further randomized trials are required to demonstrate the effectiveness of approved tests.

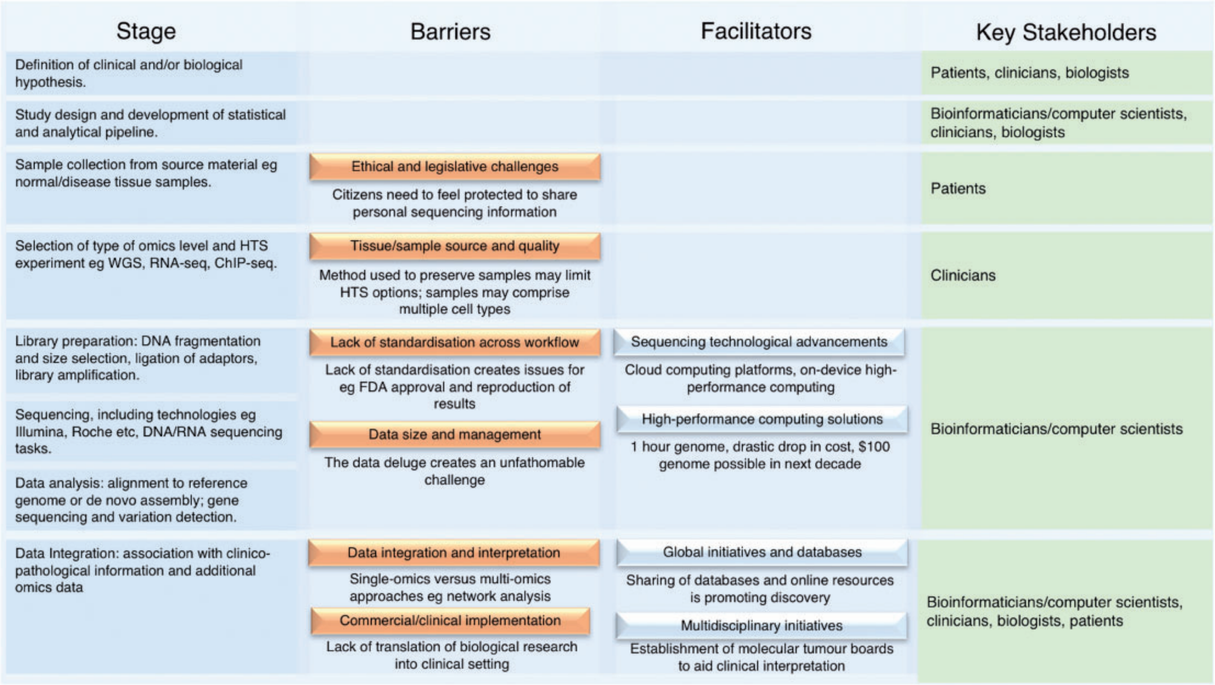
In choosing an HTS technology, users need to consider not only the biological hypothesis being tested but also sample collection and quality control issues, together with downstream computational and analytical overheads associated with a chosen platform. Whether working at the research or clinical translation level, a multidisciplinary approach is required at each HTS stage, bringing together clinicians, biologists and bioinformaticians to ensure ultimate patient benefit.

HTS platforms, pipelines and challenges

Against this heterogeneous background of regulatory approval and clinical acceptance, we examine additional barriers to and facilitators of HTS application to personalized medicine. We consider the key initial challenges, including sample collection [98] and quality [99], choice of platform [100], library preparation [101] and sequencing and data analysis [100] (Figure 2). We also highlight key stakeholders at each level.

Sample collection

Patient tissue forms the backbone of personalized medicine research. Samples for analysis may originate from formalin-fixed, paraffin-embedded (FFPE) or fresh-frozen samples. With FFPE, sample quality can be compromised by RNA degradation, leading to HTS library construction failure [98] (Figure 2). Microarray platforms have been developed to reliably quantify transcription from FFPE samples [102]. Although results with respect to RNA-Seq have been promising [103], some suggest that the bottleneck of RNA degradation currently restricts the use of HTS to DNA-seq, e.g. TS or WES [104]. With regard to the latter, the limited concordance between a WES study of fresh-frozen and FFPE melanoma samples raises concerns [105]. Where there is



Stage	Barriers	Facilitators	Key Stakeholders
Definition of clinical and/or biological hypothesis.			Patients, clinicians, biologists
Study design and development of statistical and analytical pipeline.			Bioinformaticians/computer scientists, clinicians, biologists
Sample collection from source material eg normal/disease tissue samples.	Ethical and legislative challenges Citizens need to feel protected to share personal sequencing information		Patients
Selection of type of omics level and HTS experiment eg WGS, RNA-seq, ChIP-seq.	Tissue/sample source and quality Method used to preserve samples may limit HTS options; samples may comprise multiple cell types		Clinicians
Library preparation: DNA fragmentation and size selection, ligation of adaptors, library amplification.	Lack of standardisation across workflow Lack of standardisation creates issues for eg FDA approval and reproduction of results	Sequencing technological advancements Cloud computing platforms, on-device high-performance computing	Bioinformaticians/computer scientists
Sequencing, including technologies eg Illumina, Roche etc, DNA/RNA sequencing tasks.	Data size and management The data deluge creates an unfathomable challenge	High-performance computing solutions 1 hour genome, drastic drop in cost, $100 genome possible in next decade	
Data analysis: alignment to reference genome or de novo assembly; gene sequencing and variation detection.			
Data Integration: association with clinico-pathological information and additional omics data	Data integration and interpretation Single-omics versus multi-omics approaches eg network analysis Commercial/clinical implementation Lack of translation of biological research into clinical setting	Global initiatives and databases Sharing of databases and online resources is promoting discovery Multidisciplinary initiatives Establishment of molecular tumour boards to aid clinical interpretation	Bioinformaticians/computer scientists, clinicians, biologists, patients

Figure 2. Overview of stages, barriers, facilitators and stakeholders in HTS pipelines from hypothesis setting to clinical interpretation. Eight common stages involved within a generic HTS pipeline/workflow are presented, set against factors acting as barriers to, or facilitators of, progress towards commercial/clinical translation and key stakeholders.

prior knowledge of a disease, a TS approach, focused on selected genes or regions, can be more appropriate, maintaining resolution, with increasing efficiency and affordability [106]. With the development of FFPE-tailored pre-processing pipelines alongside refinement in the underlying technologies, it is expected that HTS accuracy will potentially improve, enabling clinical uptake [105, 107]. An alternative approach to adjusting the technology would be to switch to fresh-frozen tissue (or adopt a combined strategy). Such a move would involve input from clinicians, including surgeons and pathologists, particularly in biobanks. This would represent a much more efficient alternative to FFPE, with less technical limitations and could facilitate faster clinical decision-making, though it can present considerable storage and maintenance implications [108].

Sample heterogeneity

Once a sample has been taken from tissue, its composition can be affected by heterogeneity, e.g. in tumour samples, signals may originate from multiple cell types including stroma and immune compartments [99] (Figure 2). This composition varies across samples and has implications for biomarker development, with the potential to confound results. At a bioinformatics level, *in silico* optimisation and/or gene list-based approaches have been applied to separate out signals (termed deconvolution) into their respective cell types [99, 109–112]. Once stratified into separate cell-type components, standard downstream analyses can follow. Experimental (biological) alternatives, namely, cell-specific HTS technologies, are also being used. Single-cell RNA-Seq (scRNA-Seq) has been successful in predicting treatment response in lung adenocarcinoma [113], glioblastoma [114] and melanoma [115]. The processing particularly of scRNA-Seq data requires special consideration. Standard methods, as used with 'bulk' or multi-cell data, are not always appropriate [116–118]. While scRNA-Seq may appear to be a viable alternative to *in silico* approaches, it has been suggested that cell-sorting or cell isolation experimental methods may in turn alter gene expression levels [119].

Platform choice

While sample type considerations may impact on platform choice, an overall assessment of an HTS platform's abilities, relative strengths and weaknesses, from biological, clinical and bioinformatics perspectives, will facilitate the appropriate application of the resultant data [100] (Figure 2). Recent platform examples include the Illumina® MiSeq [120], Ion PGM™ (Personal Genome Machine) [121], the PacBio RS II [122] and Qiagen Gene Reader (Sequencing-By-Synthesis) [123]. Last year, the NovaSeq Series from Illumina exceeded existing performance measures guaranteeing an average sequencing time of 1 h per genome [2]. Genomics England has had a partnership with Ilumina since 2014 [124] and has more recently in 2018 extended its partnerships to include Edico Genomics. This new alliance offers a high-performance DRAGEN Bio-IT Platform [4, 125] that reports performance greater than the 2017 NovaSeq solution. An extensive review of the past 10 years of HTS can be found in [8] along with additional technological solutions in [9, 10] and more recently in [11].

Library preparation

Once a suitable platform has been selected, library preparation, the conversion of nucleic acid materials derived from tissue, etc., into a form suitable for sequencing input, is the next key but potentially a challenging step [101] with biological and bioinformatics implications (Figure 2). Amplification of libraries by polymerase chain reaction (PCR) is prone to introducing bias; although PCR-free methods exist, these too are not challenge-free [101]. Library preparation methods are crucial when only small amounts of DNA be obtained from clinical samples. Sundaram *et al.* [126] compared seven library preparation methods for ChIP-Seq analysis of HeLa cell lines (a preclinical model of cervical cancer) against a PCR-free library preparation approach. This study concluded that there was an inverse correlation between the number cycles of amplification and performance.

Sequencing

There are different HTS approaches depending on the choice of platform, each of which uses bespoke protocols. As such, the output from data from different HTS workflows/platforms can vary [127]. This lack of standardization can present a challenge when comparing the quality and accuracy of output (Figure 2). Although primarily a bioinformatics issue, both biologists and clinicians need to be aware of how different protocols can impact results. Within a clinical diagnostic context, accuracy, reproducibility and standardization of HTS results can be improved through focusing on the development of reference standards [128].

Regardless of technology applied, the initial analysis or base-calling (whereby bases are assigned to peaks) is usually performed using platform-associated proprietary software. Alignment to a reference genome, or alternatively *de novo* assembly, is next performed. Novel methods in both sequence alignment and assembly are routinely proposed and published [129], such as the cloud computing-based CloudBurst and Rainbow [130, 131]. Additionally, enabling technologies such as Hadoop MapReduce can be used to implement algorithms, including RMAP and Bowtie [132] (covered in the 'HPC Solutions' section).

Data analysis and interpretation

Post-alignment, the appropriate analysis of data is central to an HTS project [133] (Figure 2). As the size and complexity of HTS data increase, the development of new analytical methods is required, optimization for speed and memory usage being key [9]. Given the relative youth of HTS, the lack of consensus between HTS analytical methodologies is not surprising [128, 134]. Regardless of hypothesis, platform, library preparation, sequencing protocol or downstream analytical algorithm, it is clear that HTS usage will demand extensive use of resources, both technical and human. The recruitment of skilled bioinformaticians, who can develop and manage the most appropriate tools and work within a multidisciplinary context, is crucial. Therefore, training, and standardization of training, in the use of HTS technologies is also key, as recognized by the NGS Trainer Consortium [135, 136].

Analytical/computational challenges

HTS data sets are both high-dimensional and complex in structure. Integrating such data with other data sets, platforms or technologies, to obtain a complete disease profile, is therefore both algorithmically and computationally challenging. A comprehensive review of meta-omics (integration of independent data sets at the same omics level) and poly-omics (integration of different omics types) algorithmic approaches is presented in Ma and Zhang [22]. Poly-omics projects such as TCGA have applied consensus-based methods to detect connecting patterns between different omics levels, e.g. Cluster of Cluster Assignments (COCA) [137] in breast cancer [138] and iCluster [139] in application to prostate cancer [42] and hepatocellular carcinoma [140]. Alternatively, network-based approaches [141–143] to data analysis have the potential to integrate data from disparate sources, while providing clinically relevant results. Multidisciplinary initiatives such as molecular tumour boards [144, 145], which bring together bioinformaticians, biologists and clinicians, can also help address the issue of translating complex data to be relevant to clinical care providers and patients.

The associated algorithmic approaches can require significant computational power. The resources offered by high-performance computing (HPC) can thus be exploited by bioinformaticians/computer scientists. There is now a major focus on the development of computing tools [146], platforms [147–149], data governance and infrastructure guidelines. A range of HPC solutions to support HTS is examined in the next section.

HPC solutions

HPC can be achieved by using both hardware and software to partition tasks into groups of discrete and independent computations allowing them to be scheduled in parallel, with the seamless integration of results. There are a number of possible HPC solutions that can be tailored to meet computational demands. A short introduction is provided on these distinct HPC areas: cluster [150], graphics processing units (GPUs), cloud computing platforms [151] and field-programmable gate arrays (FPGAs) (Figure 3), together with example solutions in the HTS domain. Each approach differs in terms of technology, cost, performance, scalability and ease of implementation.

Commodity clusters

Commodity clusters (Figure 3, Supplementary Table S1) have attained popularity within bioinformatics, because of their relatively low cost and scalability [152, 153]. They consist of regular desktops, with central processing units (CPUs) (for handling computations) or networked with servers (larger versions of desktops), [154] linked together to form a distributed computer system. This type of infrastructure enables parallel computing to be undertaken in (small) laboratories using low-cost hardware and standard software. However, technical experience is required in-house for the set-up; interconnection of desktops, set-up of the operating system and configuration of parallel programming software.

Open-source software frameworks such as Apache Hadoop [155] can support the scheduling of parallel operations, along with computational load and fault management. Hadoop [156] uses the MapReduce parallel programming framework, as popularized by Google, to facilitate the processing on data sets within the cluster infrastructure.

Kawalia *et al.* [157] describe a WES workflow, which incorporates MapReduce-like components for parallel calculations on clusters, enabling a 'catch-up' between data production and data processing and analysis (Table 1). MapReduce concepts have also been implemented in many other parallel solutions (Table 1) such as the Genome Analysis Toolkit (GATK) [146], a platform used for DNA- and RNA-Seq analysis in TCGA [42] and the 100 000 Genomes Project [158].

GPU computing

GPUs (Figure 3, Supplementary Table S2) are card-based devices, which can be slotted into the graphics port of a laptop or desktop. One GPU card can comprise hundreds of computational

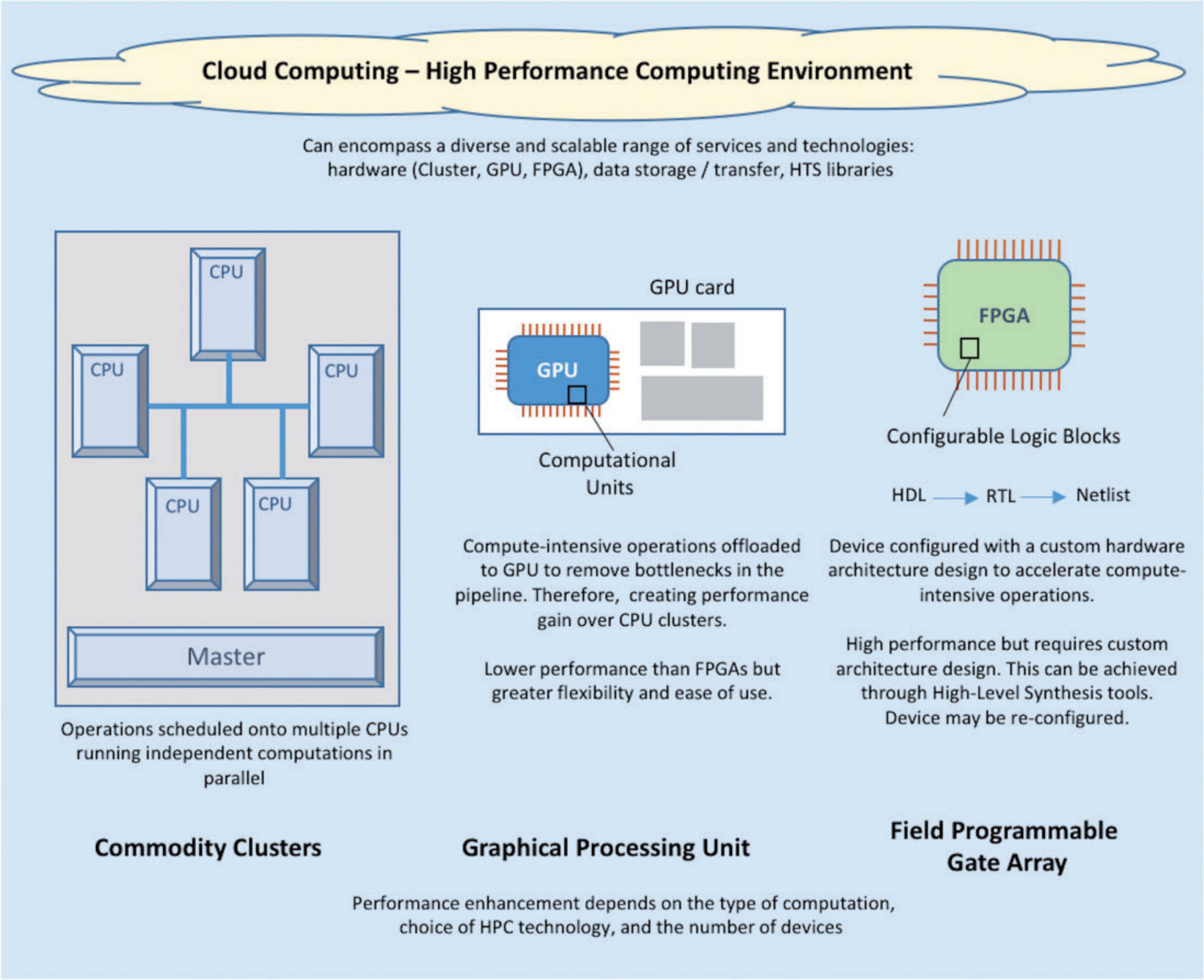

Cloud Computing – High Performance Computing Environment
Can encompass a diverse and scalable range of services and technologies:
hardware (Cluster, GPU, FPGA), data storage / transfer, HTS libraries
CPU
CPU
CPU
CPU
CPU
Master
Operations scheduled onto multiple CPUs running independent computations in parallel
Commodity Clusters
GPU card
GPU
Computational Units
Compute-intensive operations offloaded to GPU to remove bottlenecks in the pipeline. Therefore, creating performance gain over CPU clusters.
Lower performance than FPGAs but greater flexibility and ease of use.
Graphical Processing Unit
FPGA
Configurable Logic Blocks
HDL
RTL
Netlist
Device configured with a custom hardware architecture design to accelerate compute-intensive operations.
High performance but requires custom architecture design. This can be achieved through High-Level Synthesis tools. Device may be re-configured.
Field Programmable Gate Array
Performance enhancement depends on the type of computation, choice of HPC technology, and the number of devices

Figure 3. Overview of the difference options for high HPC. This is an illustration of commodity clusters, GPUs, FPGAs and cloud solutions. It highlights differences in performance, flexibility and level of custom design.
Note: HDL, hardware description language; RTL, register-transfer level.

units, in comparison with a CPU, offering increased scalability and processing performance [154, 181, 182]. Considering the price to performance ratio, parallel GPUs are potentially a more affordable and efficient option when compared with multiple, sequential CPUs [181–183].

Owing to the low cost and high-throughput processing capabilities, the GPU solution would be suitable for use in small research groups/laboratories. Code developed to run on CPUs cannot be ported to GPUs because of differences in architecture design. Therefore, computational expertise is a must. Also, data transfer between CPU and GPU memories [184] can create computational bottlenecks, limiting the potential for performance gain. Furthermore, modern GPUs have a complex architecture, which is vendor-dependent, e.g. Advanced Micro Devices Inc (AMD)\ATI Technologies Inc (ATI) or NVIDIA. Compute Unified Device Architecture (CUDA) [185] offered by NVIDIA is the most used platform and model for GPU parallel programming.

A large number of CUDA-compatible HTS data processing and analysis tools have been developed in the past for use with RNA-seq [163] and DNA-seq, e.g. Cushaw [186], BarraCUDA [187], SOAP3 [188], CUDASW++ [189] and SeqNFind [183], with a focus on sequence alignment using GPUs [186, 187] or CPUs and GPUs combined [189] (Table 1).

Cloud computing

Cluster and GPU-based solutions can be implemented in-house. Cloud computing (Figure 3, Supplementary Table S3) refers to the use of off-site (remote) computers or servers for storage and processing, accessed by a user across a network connection. A major advantage of cloud solutions is that they provide adaptable storage and performance, without the necessity to deploy and maintain internal resources [147, 148], thereby providing scalable solutions to individual researchers through to large-scale clinical labs.

At the start of the 'Big Data' era, cloud computing was dominated by the use of Hadoop-based clusters. Since then, there has been a significant growth in the services provided by cloud vendors, offering data and project management tools that facilitate collaborations, regulate access to shared data and enable visualization and analysis of that data. Commercial options provide powerful solutions; however, organizations can develop their own private clouds using open-source facilities. These in-house servers may be regarded as more suitable solutions for sensitive data (e.g. patient information). Public clouds can be a viable option if sensitive data are encrypted, anonymized or used at a sufficiently abstract level omitting sensitive details [190].

Table 1. Example HTS applications using cluster, GPU, cloud and FPGA HPC solutions

HPC solution	HTS personalized medicine applications
Cluster	Exome analysis workflow: [157]
	GATK [146] used by TCGA [42] and the 100 000 Genomes Project [158]
	Sequence alignment BLAST [159]
	Dimensionality reduction, Self-Organizing Maps (SOM) [160]
GPU	Process-intensive tasks such as RNA-seq alignment [161] and assembly [162].
	Review of GPUs applied to RNA-Seq on cancer [163] such as parallel construction of Fuzzy C-Means clustering algorithm [164]
	Read mapping [165]
	Error correction [166]
Cloud	HTS read mapping algorithms such as CloudBurst [130], CloVR [149] and the Crossbow [167]
	Tailored bioinformatics platforms: BIOVIA ScienceCloud [147], DNAnexus [148], BaseSpace Sequence Hub [168] and Seven Bridges [169]
	Key projects have used public and private clouds, namely, International Cancer Genome Consortium and 100 000 Genomes Project
FPGA	Survey of FPGAs used in computational biology contexts: [170]
	General overview: [171, 172]
	Alignment algorithms: [173, 174].
	Basic Local Alignment Search Tool (BLAST) FPGA accelerators: [175–177]
	Short read mapping: [174]
	Genome sequencing: MapReduce framework with acceleration on FPGA [178]
	Large-scale protein sequence alignment: [173]
	Complexity analysis of sequence tracts algorithm for low-complexity regions (LCRs) in protein sequences: [179]
	DRAGEN (Dynamic Read Analysis for Genomics) Processor: [180]

The major players in commercial cloud provision, Amazon Web Services (AWS) Elastic Compute Cloud [191], Google Genomics [192] and Microsoft Azure [193], guarantee data security, with scalability and speed. High-profile HTS studies such as the 100 000 Genome Sequencing project have used private clouds while partnering with private companies, including AWS [194] and UK Cloud [195].

While commercial cloud solutions provide user-friendly interfaces with extensive toolkits, there are inherent disadvantages, including a lack of flexibility [196]. Open-source alternatives include platforms and pipelines, such as the alignment tool CloudBurst [130], a platform that combines virtual machine and cloud technologies, and CloVR [149] and the automated pipeline, Crossbow [167] (Table 1). However, open-source solutions arguably require more investment from the user, including system installation and management and the implementation of data analysis pipelines [196], all requiring substantial technical skills [149, 197].

FPGA-based platforms

FPGA devices (Figure 3, Supplementary Table S4) are programmable integrated circuits, which consist of an array of configurable logic blocks each comprising local memory and computational units. The FPGA's strength lies in its ability to reconfigure the dedicated hardware resources to meet the specific design needs of the implemented algorithms. FPGAs can yield great performance gains over GPUs for highly regular parallel operations. However, they are significantly more difficult to program, although this process has been simplified through recent high-level synthesis tools [198–200].

Furthermore, as with GPUs, FPGAs need to be part of a larger HPC environment for controlling which operations are sent to the device. However, vendors such as Intel have been developing hybrid CPU-FPGA Programmable Acceleration Cards [201] providing support for an acceleration stack of software, firmware and tools to assist this process. Recently, FPGAs have also found application in cloud platforms, such as Microsoft Azure [202] and Amazon AWS [203] providing additional flexibility and performance. Development would still need to be undertaken by bioinformaticians/computer scientists with computational skills in hardware design; however, the tools and solutions are evolving to make FPGA acceleration a more accessible option [199–203].

FPGAs have been used in computational biology settings [170], though to a lesser extent than cluster, GPU and cloud-based options with respect to HTS (Table 1). The most high-profile example involves Edico Genome, developers of the FPGA-powered DRAGEN Bio-IT Platform [4] and their partnership with Genomics England [125]. FPGAs are central to enabling this work, offering acceleration on sequencing pipeline computational bottlenecks, e.g. alignment and mapping. Owing to its high level of parallelism, DRAGEN can process a 'whole human genome at 30x coverage in about 20 minutes, compared to 20-30 hours using a CPU-based system' [4].

Each technology discussed offers advantages in their own right, providing performance gains dependent on the approaches taken. However, these solutions differ in terms of scalability, flexibility, cost and computational expertise for implementation. These solutions do not necessarily need to be taken individually, and the combination of clusters, GPUs, FPGAs and cloud-based workflows offers great promise to provide tailored genomic analysis solutions.

Data management and governance

While technological and bioinformatics developments have paved the way for the generation of HTS data on smaller machines within reduced time frames and limited budgets, new challenges have arisen. Governance comes to the fore when considering the storage, sharing and privacy of the resultant data generated.

Data size

While HTS data production costs are falling, the associated storage costs are reducing at a much slower rate [5]. Obtaining the actual sequence is only one part of a more complex overhead. Data storage, transmission, navigation and searches and the associated data processing resource and tools must also be considered [14, 204].

Management of large genomic data sets is discussed by Batley and Edwards [205]. Although data volume reduces from terabytes/gigabytes at the raw sequence stage, to gigabytes/megabytes once stored in text sequence format, there are further challenges in terms of data searchability and accessibility. Using standard sequence comparison algorithms is time-consuming; furthermore, tools such as BLAST are computationally intensive [160, 177, 206].

Compression techniques offer another effective storage solution [207–209], often comparing sequences against reference genomes [204, 210]. In Brandon *et al.* [204] resultant differences were encoded using entropy-based methods such as Huffman. Through such techniques, a 345-fold compression rate was achieved, in one example reducing a 56 MB sequence down to 167 KB.

The graphical representation and interpretation of data is also an important factor, particularly as data sets increase in size and diversity, leading to the development of visualization tools [211, 212].

Data security

Genetic information can provide the ultimate insight into the health of private individuals. As such it needs to be treated with the greatest levels of confidence, security and ethical standards. Once such data become a component of a computer infrastructure, high-level cyber security measures need to be used, namely, encryption, authentication and authorization [213]. Furthermore, before inclusion in a publicly available HTS repository, donor anonymity must be safeguarded [214, 215].

The US Presidential Commission for the Study of Bioethical Issues recommended that there needs to be 'strong baseline protections while promoting data access and sharing' [216]. Such sharing should be with the goal of progressing biological knowledge for public benefit.

Recognizing the translational challenges posed by data repositories [217], bodies such as the Electronic MEdical Records and GEnomics (eMERGE) Consortium [218] have contributed towards developing good practice guidelines and standards in the governance of genomic data (Table 2). In sharing or publishing data, ensuring the anonymity of patients is routinely achieved through de-identification. In certain research areas, e.g. the study of rare diseases, there is a risk of traceability through publication of associated information such as age, ethnicity and gender [214]. The security and storage of such data can be further protected by considering it as protected health information (PHI).

Despite the computational benefits of cloud-based solutions, the security of data and subsequent analysis held within such frameworks are still considered bottlenecks [224]. Cloud-based providers have responded, through the development of in-built facilities, such as encryption, auditing, data backup and recovery, to comply with data governance and management regulations as required, e.g. by the Health Insurance Portability and Accountability Act (HIPAA) of 1996 [225]. Examples include AWS DNAnexus [148] and the hybrid offering from Microsoft Azure [226]. HTS-tailored alternatives such as BC Platforms can be implemented in-house, targeting security- and cost-conscious end users [227].

Genetic databases can be shared successfully and at a global scale. One such example, GenBank, is a generic sequence database (nucleotide sequences and their protein translations) established and coordinated by the National Center for

Table 2. US and EU organizations established for the protection of health and personal data

Legislation	Date	Description
Health Insurance Portability and Accountability Act (HIPPA)	1996	HIPPA safeguards individuals' PHI [219]. Its privacy rules set guidelines on how health data can be disseminated through suitable de-identification. Two standards (Safe Harbor and Expert Determination) may be used for the de-identification process [220]
Health Information Technology for Economic and Clinical Health Act (HITECH)	2009	HIPAA was later supplemented by the Health Information Technology for Economic and Clinical Health Act (HITECH)
Genetic Information Nondiscrimination Act of 2008 (GINA)	2008	A US Federal Law that prohibits discrimination in health insurance and employment as a result of genetic information [221] Note however that GINA does not provide complete coverage, e.g. it does not prohibit health insurers from using genetic information in determining insurance premiums
Patient Protection and Affordable Care Act (ACA) of 2010	2010	Makes it illegal for health insurers to raise premiums or remove cover for those with pre-existing conditions
Directive 95/46/EC of the European Parliament and Council of the European Union (EU)	1995	This directive covers the protection of individuals with regard to the processing of personal data and on the free movement of such data. (Official Journal of the European Union L 281: 0031–0050.)
Directive (EU) 2016/680 Regulation (EU) 2016/679/	With effect from 2018	Directive 95/46/EC will be repealed and replaced by the regulation and directive on the protection of natural persons with regard to the processing of personal data—General Data Protection Regulation (GDPR) [222, 223]

Biotechnology (part of the National Institutes of Health in the United States) [228]. The initiative is part of the International Nucleotide Sequence Database Collaboration (INSDC), comprising members from DNA DataBank of Japan (DDBJ) and the European Nucleotide Archive (ENA) [229–231].

The INSDC collections, comprising submissions from both small- and large-scale independent laboratories, are freely available. INSDC adds to its database daily, with a GenBank public update every 2 months. The expansion of GenBank's database has doubled approximately every 18 months [232], highlighting the growth, supporting infrastructure and acceptance, in sharing sequencing data. The European Bioinformatics Institute (EBI) repository, ArrayExpress [233], also allows researchers to upload their HTS data sets for public distribution. In depositing data, standardization is required, e.g. Minimum Information About a Microarray Experiment (MIAME) and Minimum Information about a high-throughput nucleotide SEQuencing Experiment (MINSEQE) guidelines [234].

Ethical issues surrounding genetic data

There are multiple ethical challenges when handling HTS data. Apart from the obvious examples, e.g. a data breach, there are other more subtle, unanticipated incidences. A key case is that of Henrietta Lacks. Henrietta died from cervical cancer in 1951; yet, the cell line derived from her tumour (HeLa) is still replicating and has become a pivotal resource as a preclinical model [235]. The ethical conflict in this example arose from the lack of consideration for the family of Lacks and indeed lack of consent with regard to the publication of the results from sequencing of the cell line. An uploaded sequenced HeLa sample was retracted from the ENA because of privacy concerns in 2013 [236, 237]. This highlighted the lack of clarity and legislation surrounding ownership over donated samples and the potential impact for family members.

Furthermore, if we consider the process of genetic discovery, what is the line between research and clinical diagnosis [238]? This presents many quandaries for retrospective research projects in particular. If a cancer patient, who had agreed to donate material from their tumour for a research project, was found to possess a particular gene mutation, do researchers have a responsibility to inform the patient and/or the patient's family [75]? If a compound had not yet been approved for treating this particular mutation, this new knowledge could not be used to advance the health of the patient.

If we take it on ourselves to sequence our DNA, could this impact insurance? Table 2 provides a summary of current legislation for the United States and European Union (EU). Both the US Health Information Technology and Clinical Health Act (HITECH) and the EU Directive 2016/680 Regulation 2016/679 provide safeguarding of individuals health data and how it is handled and transmitted. As part of the 2010 US Patient Protection and Affordable Care Act (ACA) cancer risk assessment, via genetic testing, was promoted as a preventive measure under the assurance that no person would be negatively impacted by changes in cost or provision in their insurance cover [239]. While the legislation was not fully comprehensive of all conditions, the 'good-will' of preventive medicine, based on personalized risk, was present. However, with current changes in US legislation and the development of the 'Preserving Employee Wellness Programs Act' [240], concerns have been raised regarding the protection of employees' rights [241]. The full implications are unclear, but it does appear that employees will be given fewer options in terms of privacy, contradicting the legislation as set out by the 2008 Genetic Information Nondiscrimination Act (GINA). If employers are empowered to this extent, there is a risk that the public will lose confidence in, and acceptance of, genetic testing, impacting negatively on the uptake in preventative screening.

Discussion

We have provided a broad overview of the facilitators and barriers associated with the widespread adoption of HTS in personalized medicine (Figure 2). Technological advances have been a key driver in offering affordable and efficient access to sequencing solutions, with the 1 h genome sequence now a reality [3] and Illumina forecasting a $100 cost per genome within 3–10 years [242]. Illumina have also been developing chip-based sequencing incorporating their DNA- and RNA-Seq technologies into a semiconductor device with the resulting product launched in 2017 [243].

In terms of computational power to perform analysis, technology is at a significant stage. Cloud platforms offer scalability, security and computational performance [148, 191, 193, 227, 244]. Meanwhile, advances beyond the cloud also continue with visions for silicon chip-based and mobile solutions [243, 245] with an eye towards real-time processing of HTS data.

A multidisciplinary approach to technological development and translational research is required to promote HTS within personalized medicine. However, barriers must be acknowledged; the phenomenal production rate of sequencing data has the potential to overwhelm current computing infrastructures and bioinformatics resources [5, 14, 246].

Addressing heterogeneity in output through standardization is crucial when considering HTS data integration with healthcare informatics structures. In particular, to consider assimilation, electronic healthcare records, raw HTS data and associated ontologies, must be normalized [247]. This challenge has been recognized with a call for replicable and auditable workflows [171]. This must be supported by an investment in informatics infrastructure, with a focus on storage and software development [248]. Patient consent highlights the need for a goodwill 'buy-in' by the general public, in terms of data-sharing, alongside a closer patient involvement [215]. This can only be achieved if there is confidence in privacy assurances.

A disconnect between HTS data production and the analytics required to facilitate biological understanding still exists. Li *et al.* [249] acknowledge that 'integrative analysis of this rich clinical, pathological, molecular and imaging data represents one of the greatest bottlenecks in biomarker discovery research in cancer and other diseases'. This may be addressed by larger studies such as the 100 000 Genome Project, which aims to sequence the genomes of 100 000 patients enabling downstream integration of results with associated clinico-pathological data [158] or the PatientsLikeMe project, which is collaborating with governmental and pharmaceutical companies [250].

Such large-scale projects will depend on the standardization of data management and analysis, if HTS-produced biomarkers are to be translated into the clinic for patient diagnosis and treatment stratification. Also, avoiding the 'Winner's Curse' can be achieved through the use of appropriate study design, robust statistical methods and validation [251, 252]. Standardized, replicable pipelines, from sequencing to downstream analysis, are therefore now required, such as the FDA/HUPO Proteomics Standards Initiative-established Sequencing Quality Control (SEQC) project [253]. While we are still playing bioinformatics catch-up with the HTS wave, new small-scale real-time sequencing solutions are coming on-stream [3, 4, 32]. It is essential that we apply the lessons learned from the previous computational and governance challenges to keep pace with new HTS developments.

Conclusion

To ensure its place within the personalized medicine arsenal, first and foremost, the computational resources required for HTS processing must be accessible in terms of costs, skills and efficiency. Standardization in HTS processing and analytical pipelines will facilitate validation and ensure replication of results, within clinically relevant time frames. This, in turn, alongside multidisciplinary collaboration, will enable its full integration into patient care and treatment, through the provision of new diagnostic, predictive and prognostic tests.

Key Points

- An overview on sequencing technologies and their role in personalized medicine.
- Identification of current bottlenecks in the translation of 'omic' data to personalized medicine.
- Up-to-date review on current computational technologies, infrastructure and future solutions to handling and analysing of sequencing data in real time.
- The changing required in clinical governance in the face of rapid adoption of sequencing technologies into clinical workflows.
- This paper provides a review of high-throughput sequencing in the context of biomedical research to clinical use with a focus on applications, pipelines, processes and technologies along with challenges.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

References

1. Miller NA, Farrow EG, Gibson M. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Med* 2015;**7**:100.
2. Illumina. NovaSeq series—the next era in sequencing starts now. Illumina, 2018. https://www.illumina.com/systems/sequencing-platforms/novaseq.html
3. Fikes B. New machines can sequence human genome in one hour, Illumina announces. *The San Diego Union-Tribune*. 2017. http://www.sandiegouniontribune.com/business/biotech/sd-me-illumina-novaseq-20170109-story.html
4. Edico Genome. *DRAGEN Bio-IT platform*. Edico Genome, 2018. http://edicogenome.com/dragen-bioit-platform/
5. Baker M. Next-generation sequencing: adjusting to data overload. *Nat Methods* 2010;**7**(7):495–9.
6. Schaller RR. Moore's law: past, present and future. *IEEE Spectr* 1997;**34**(6):52–9.
7. Wetterstrand K. *DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP)*. National Human Genome Research Institute, 2017. https://www.genome.gov/sequencingcostsdata/
8. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**(6):333–51.
9. Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet* 2010;**11**(1):31–46.
10. Loman NJ, Misra RV, Dallman TJ, *et al*. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012;**30**(5):434–9.
11. Mardis ER. DNA sequencing technologies: 2006–2016. *Nat Protoc* 2017;**12**(2):213–18.
12. Naccache SN, Federman S, Veeraraghavan N, *et al*. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res* 2014;**24**(7):1180–92.
13. Anderson C. Data deluge. *Clin OMICS* 2017;**4**(1):26–9.
14. Sboner A, Mu X, Greenbaum D, *et al*. The real cost of sequencing: higher than you think! *Genome Biol* 2011;**12**(8):125.
15. Leipzig J. A review of bioinformatic pipeline frameworks. *Brief Bioinform* 2016;**18**:bbw020.
16. Muir P, Li S, Lou S, *et al*. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol* 2016;**17**:53.
17. Lightbody G, Browne F, Zheng H, *et al*. The role of high performance, grid and cloud computing in high-throughput sequencing. In: *The 2016 IEEE International Conference on Bioinformatics & Biomedicine, At Shenzhen, China*. IEEE, 2016, 890–5. https://ieeexplore.ieee.org/document/7822643/
18. NCI. *Definition of personalized medicine—National Cancer Institute Dictionary of Cancer Terms*. NCI, 2017. https://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=561717
19. Beger RD, Dunn W, Schmidt MA, *et al*. Metabolomics enables precision medicine: 'a white paper, community perspective'. *Metabolomics* 2016;**12**(10):149.
20. Tourneau CL, Kamal M, Tsimberidou AM, *et al*. Treatment algorithms based on tumor molecular profiling: the essence of precision medicine trials. *J Natl Cancer Inst* 2016;**108**(4):djv362.
21. Ritchie MD, Holzinger ER, Li R, *et al*. Methods of integrating data to uncover genotype–phenotype interactions. *Nat Rev Genet* 2015;**16**(2):85–97.
22. Ma T, Zhang A. Omics Informatics: From Scattered Individual Software Tools to Integrated Workflow Management Systems. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**(4):926–46.
23. Alberts B, Johnson A, Lewis J, *et al*. *Molecular Biology of the Cell*. 4th edition. New York: Garland Science; 2002. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21054/
24. Gibney ER, Nolan CM. Epigenetics and gene expression. *Heredity* 2010;**105**(1):4–13.
25. Haraksingh RR, Snyder MP. Impacts of variation in the human genome on gene regulation. *J Mol Biol* 2013;**425**(21):3970–7.
26. Dworkis DA, Klings ES, Solovieff N, *et al*. Severe sickle cell anemia is associated with increased plasma levels of TNF-R1 and VCAM-1. *Am J Hematol* 2011;**86**(2):220–3.
27. White MB, Amos J, Hsu JM, *et al*. A frame-shift mutation in the cystic fibrosis gene. *Nature* 1990;**344**(6267):665–7.
28. Craddock N, Hurles ME, Cardin N, *et al*. Genome-wide association study of CNVs in 16, 000 cases of eight common diseases and 3, 000 shared controls. *Nature* 2010;**464**:713–20.
29. Tomlins SA, Rhodes DR, Perner S, *et al*. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 2005;**310**(5748):644–8.
30. Pollack JR, Perou CM, Alizadeh AA, *et al*. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* 1999;**23**(1):41–6.

31. Meienberg J, Bruggmann R, Oexle K, *et al*. Clinical sequencing: is WGS the better WES? *Hum Genet* 2016;**135**(3):359–62.
32. Votintseva AA, Bradley P, Pankhurst L, *et al*. Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. *J Clin Microbiol* 2017;**55**:1285–98.
33. de Ligt J, Willemsen MH, van Bon BW, *et al*. Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* 2012;**367**(20):1921–9.
34. Lionel AC, Costain G, Monfared N, *et al*. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet Med* 2018;**20**(4):435.
35. Rao PN, Uplekar S, Kayal S, *et al*. A method for amplicon deep sequencing of drug resistance genes in plasmodium falciparum clinical isolates from India. *J Clin Microbiol* 2016;**54**(6):1500–11.
36. Bohacek J, Mansuy IM. Epigenetic inheritance of disease and disease risk. *Neuropsychopharmacology* 2013;**38**(1):220–36.
37. Jorda M, Peinado MA. Methods for DNA methylation analysis and applications in colon cancer. *Mutat Res* 2010;**693**: 84–93.
38. Rackham OJ, Langley SR, Oates T, *et al*. A Bayesian approach for analysis of whole-genome bisulfite sequencing data identifies disease-associated changes in DNA methylation. *Genetics* 2017;**205**(4):1443–58.
39. Legendre C, Gooden GC, Johnson K, *et al*. Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer. *Clin Epigenetics* 2015;**7**:100.
40. Tan PY, Chang CW, Chng KR, *et al*. Integration of regulatory networks by NKX3-1 promotes androgen-dependent prostate cancer survival. *Mol Cell Biol* 2012;**32**(2):399–414.
41. Ross-Innes CS, Stark R, Teschendorff AE, *et al*. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* 2012;**481**(7381):389–93.
42. Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* 2015;**163**:1011–25.
43. Raphael BJ, Hruban RH, Aguirre AJ, *et al*. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* 2017;**32**:185–203.e13.
44. Kim J, Bowlby R, Mungall AJ, *et al*. Integrated genomic characterization of oesophageal carcinoma. *Nature* 2017;**541**: 169–74.
45. Farshidfar F, Zheng S, Gingras MC, *et al*. Integrative genomic analysis of cholangiocarcinoma identifies distinct IDH-mutant molecular profiles. *Cell Rep* 2017;**18**(11):2780–94.
46. Parker JS, Mullins M, Cheang MCU, *et al*. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 2009;**27**:1160–7.
47. Frith MC, Pheasant M, Mattick JS. The amazing complexity of the human transcriptome. *Eur J Hum Genet* 2005;**13**(8): 894–7.
48. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;**116**(2):281–97.
49. Keller A, Leidinger P, Lange J, *et al*. Multiple sclerosis: microRNA expression profiles accurately differentiate patients with relapsing-remitting disease from healthy controls. *PLoS One* 2009;**4**(10):e7440.
50. Huang J, Wang F, Argyris E, *et al*. Cellular microRNAs contribute to HIV-1 latency in resting primary CD4+T lymphocytes. *Nat Med* 2007;**13**(10):1241–7.
51. Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**(1):57–63.
52. Daugaard I, Venø MT, Yan Y, *et al*. Small RNA sequencing reveals metastasis-related microRNAs in lung adenocarcinoma. *Oncotarget* 2017;**8**:27047–61.
53. Banks RE, Dunn MJ, Hochstrasser DF, *et al*. Proteomics: new perspectives, new biomedical opportunities. *Lancet* 2000; **356**(9243):1749–56.
54. Oprea TI, Bologa CG, Brunak S, *et al*. Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov* 2018;**17**(5):317–32.
55. Becnel LB, McKenna NJ. Minireview: progress and challenges in proteomics data management, sharing, and integration. *Mol Endocrinol* 2012;**26**(10):1660–74.
56. Velez G, Roybal CN, Colgan D, *et al*. Personalized proteomics for the diagnosis and treatment of idiopathic inflammatory disease. *JAMA Ophthalmol* 2016;**134**(4):444–8.
57. Liao H, Wu J, Kuhn E, *et al*. Use of mass spectrometry to identify protein biomarkers of disease severity in the synovial fluid and serum of patients with rheumatoid arthritis. *Arthritis Rheum* 2004;**50**(12):3792–803.
58. Obach RS. Pharmacologically active drug metabolites: impact on drug discovery and pharmacotherapy. *Pharmacol Rev* 2013;**65**(2):578–640.
59. Quehenberger O, Dennis EA. The human plasma lipidome. *N Engl J Med* 2011;**365**(19):1812–23.
60. Acevedo A, Duran C, Ciucci S, *et al*. LIPEA: lipid pathway enrichment analysis. *bioRxiv* 2018. doi: 10.1101/274969.
61. Sales S, Graessler J, Ciucci S, *et al*. Gender, contraceptives and individual metabolic predisposition shape a healthy plasma lipidome. *Sci Rep* 2016;**6**(1):27710.
62. Ke C, Li A, Hou Y, *et al*. Metabolic phenotyping for monitoring ovarian cancer patients. *Sci Rep* 2016;**6**(1):23334.
63. TCGA. The Cancer Genome Atlas. 2018. https://cancergenome.nih.gov/
64. McLendon R, Friedman A, Bigner D, *et al*. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;**455**(7216):1061–8.
65. Cherniack AD, Shen H, Walter V, *et al*. Integrated molecular characterization of uterine carcinosarcoma. *Cancer Cell* 2017; **31**(3):411–23.
66. Mutz KO, Heilkenbrinker A, Lönne M, *et al*. Transcriptome analysis using next-generation sequencing. *Curr Opin Biotechnol* 2013;**24**(1):22–30.
67. Zhao S, Fung-Leung WP, Bittner A, *et al*. Comparison of RNA-seq and microarray in transcriptome profiling of activated T cells. *PLoS One* 2014;**9**(1):e78644.
68. Zhang W, Yu Y, Hertwig F, *et al*. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol* 2015;**16**:133.
69. Knijnenburg TA, Wang L, Zimmermann MT, *et al*. Genomic and molecular landscape of DNA damage repair deficiency across The Cancer Genome Atlas. *Cell Rep* 2018;**23**(1):239–54.e6.
70. Thorsson V, Gibbs DL, Brown SD, *et al*. The immune landscape of cancer. *Immunity* 2018;**48**(4):812–30.e14.
71. Aravanis AM, Lee M, Klausner RD. Next-generation sequencing of circulating tumor DNA for early cancer detection. *Cell* 2017;**168**(4):571–4.
72. Abrams J, Conley B, Mooney M, *et al*. National Cancer Institute's Precision Medicine Initiatives for the new National Clinical Trials Network. *Am Soc Clin Oncol Educ Book* 2014;**34**:71–6.
73. ClinicalTrials.gov. Identifier NCT03085888. The STRIVE Study: breast cancer screening. Bethesda: U.S. National Library of Medicine, 2017.

74. Barroilhet L, Matulonis U. The NCI-MATCH trial and precision medicine in gynecologic cancers. *Gynecol Oncol* 2018; **148**(3):585–90.

75. Roychowdhury S, Iyer MK, Robinson DR, *et al*. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Transl Med* 2011;**3**:111ra121.

76. Massard C, Michiels S, Ferté C, *et al*. High-throughput genomics and clinical outcome in hard-to-treat advanced cancers: results of the MOSCATO 01 trial. *Cancer Discov* 2017;**7**(6): 586–95.

77. Iyer G, Hanrahan AJ, Milowsky MI, *et al*. Genome sequencing identifies a basis for everolimus sensitivity. *Science* 2012; **338**(6104):221.

78. Chau NG, Lorch JH. Exceptional responders inspire change: lessons for drug development from the bedside to the bench and back. *Oncologist* 2015;**20**(7):699–701.

79. Collins FS, Hamburg MA. First FDA authorization for next-generation sequencer. *N Engl J Med* 2013;**369**(25):2369–71.

80. Sosnay PR, Siklosi KR, Van Goor F, *et al*. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat Genet* 2013;**45**:1160–7.

81. Hughes EE, Stevens CF, Saavedra-Matiz CA, *et al*. Clinical sensitivity of cystic fibrosis mutation panels in a diverse population. *Hum Mutat* 2016;**37**(2):201–8.

82. US Food and Drug Administration. Use of standards in FDA regulatory oversight of Next Generation Sequencing (NGS)—based In Vitro Diagnostics (IVDs) used for diagnosing germline diseases (draft guidance). US Food and Drug Administration, 2016. https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/Guidance Documents/UCM509838.pdf

83. Foundation Medicine. FoundationOne CDx. Foundation Medicine. 2017. https://www.foundationmedicine.com/genomic-testing/foundation-one-cdx

84. Memorial Sloan Kettering Cancer Center. MSK researchers develop targeted test for mutations in both rare and common cancers. Memorial Sloan Kettering Cancer Center, 2018. https://www.mskcc.org/msk-impact

85. US Food and Drug Administration. FDA Fact Sheet—CDRH's approach to tumor profiling next generation sequencing tests. US Food and Drug Administration, 2018. https://www.fda.gov/downloads/medicaldevices/productsandmedical procedures/invitrodiagnostics/ucm584603.pdf

86. Thermo Fisher Scientific. Oncomine Dx target test. Thermo Fisher Scientific, 2018. https://www.thermofisher.com/uk/en/home/clinical/diagnostic-testing/condition-disease-diagnostics/oncology-diagnostics/oncomine-dx-target-test.html

87. Wallden B, Storhoff J, Nielsen T, *et al*. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med Genomics* 2015;**8**:54.

88. Saghatchian M, Mook S, Pruneri G, *et al*. Additional prognostic value of the 70-gene signature (MammaPrint®) among breast cancer patients with 4-9 positive lymph nodes. *Breast* 2013;**22**(5):682–690.

89. van de Vijver MJ, He YD, van't Veer LJ, *et al*. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002;**347**(25):1999–2009.

90. US Food and Drug Administration. 510(k) substantial equivalence determination decision summary assay and instrument combinationtemplate: Prosigna. US Food and Drug Administration, 2017. https://www.accessdata.fda.gov/cdrh_docs/reviews/K130010.pdf

91. Agendia. Agendia announces CE mark for NGS-Based MammaPrint®BluePrint®Kit enhancing access to personalized treatment for breast cancer patients in Europe. Agendia, 2018. http://www.agendia.com/agendia-announces-ce-mark-for-ngs-based-mammaprint-blueprint-kit/

92. NanoString Technologies. NanoString Technologies obtains CE mark for PAM50-based test for breast cancer. NanoString Technologies, 2012. http://investors.nanostring.com/static-files/8de61464-0e6b-482a-b5c6-1665c9a8e90c

93. Duffy MJ, Harbeck N, Nap M, *et al*. Clinical use of biomarkers in breast cancer: updated guidelines from the European Group on Tumor Markers (EGTM). *Eur J Cancer* 2017;**75**:284–98.

94. NCCN. *National Comprehensive Cancer Network—NCCB clinical practice guidelines in oncology*. NCCN, 2018. https://www.nccn.org/

95. Paik S, Shak S, Tang G, *et al*. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004;**351**(27):2817–26.

96. Prasad V. Perspective: the precision-oncology illusion. *Nature* 2016;**537**(7619):S63.

97. Prasad V. Why the US Centers for Medicare and Medicaid Services (CMS) should have required a randomized trial of Foundation Medicine (F1CDx) before paying for it. *Ann Oncol* 2018;**29**(2):298–300.

98. Zhang P, Lehmann BD, Shyr Y, *et al*. The utilization of formalin fixed-paraffin-embedded specimens in high throughput genomic studies. *Int J Genomics* 2017;**2017**:1–9.

99. Shen-Orr SS, Tibshirani R, Khatri P, *et al*. Cell type–specific gene expression differences in complex tissues. *Nat Methods* 2010;**7**(4):287–9.

100. Henson J, Tischler G, Ning Z. Next-generation sequencing and large genome assemblies. *Pharmacogenomics* 2012;**13**(8): 901–15.

101. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp Cell Res* 2014;**322**:12–20.

102. Kennedy RD, Bylesjo M, Kerr P, *et al*. Development and independent validation of a prognostic assay for stage II colon cancer using formalin-fixed paraffin-embedded tissue. *J Clin Oncol* 2011;**29**:4620–6.

103. Graw S, Meier R, Minn K, *et al*. Robust gene expression and mutation analyses of RNA-sequencing of formalin-fixed diagnostic tumor samples. *Sci Rep* 2015;**5**(1):12335.

104. Menon R, Deng M, Boehm D, *et al*. Exome enrichment and SOLiD sequencing of formalin fixed paraffin embedded (FFPE) prostate cancer tissue. *Int J Mol Sci* 2012;**13**(7):8933–42.

105. De Paoli-Iseppi R, Johansson PA, Menzies AM, *et al*. Comparison of whole-exome sequencing of matched fresh and formalin fixed paraffin embedded melanoma tumours: implications for clinical decision making. *Pathology* 2016; **48**(3):261–6.

106. Lu J, Getz G, Miska EA, *et al*. MicroRNA expression profiles classify human cancers. *Nature* 2005;**435**(7043):834–8.

107. Wagle N, Berger MF, Davis MJ, *et al*. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer Discov* 2012;**2**(1):82–93.

108. Arreaza G, Qiu P, Pang L, *et al*. Pre-Analytical Considerations for Successful Next-Generation Sequencing (NGS): challenges and opportunities for Formalin-Fixed and Paraffin-Embedded tumor tissue (FFPE) samples. *Int J Mol Sci* 2016; **17**(9):1579.

109. Gong T, Hartmann N, Kohane IS, *et al*. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS One* 2011;**6**(11):e27156.

110. Moffitt RA, Marayati R, Flate EL, *et al*. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat Genet* 2015;**47**(10): 1168–78.
111. Yoshihara K, Shahmoradgoli M, Martinez E, *et al*. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 2013;**4**:1–29.
112. Li Y, Xie X. A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. *BMC Bioinforma* 2013;**14(Suppl 5)**: S11.
113. Kim KT, Lee HW, Lee HO, *et al*. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol* 2015; **16**:127.
114. Patel AP, Tirosh I, Trombetta JJ, *et al*. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 2014;**344**(6190):1396–401.
115. Tirosh I, Izar B, Prakadan SM, *et al*. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 2016;**352**(6282):189–96.
116. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 2015;**16**(3):133–45.
117. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol* 2016;**17**:63.
118. Yuan GC, Cai L, Elowitz M, *et al*. Challenges and emerging directions in single-cell analysis. *Genome Biol* 2017;**18**(1):84.
119. Feezor RJ, Baker HV, Mindrinos M, *et al*. Whole blood and leukocyte RNA isolation for gene expression analyses. *Physiol Genomics* 2004;**19**(3):247–54.
120. Illumina. MiSeq gene & small genome sequencer. Illumina, 2016. http://www.illumina.com/systems/miseq.html
121. Thermo Fisher Scientific. Ion PGM system for next-generation sequencing. Thermo Fisher Scientific, 2018. https://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-pgm-system-for-next-generation-sequencing.html
122. PacBio. PacBio RS II. Pacific Biosciences, 2017. http://www.pacb.com/products-and-services/pacbio-systems/rsii/
123. Qiagen. *GeneRead Sequencing (NGS)*. QIAGEN, 2018. https://www.qiagen.com/de/resources/technologies/ngs/
124. Genomics England. *UK to become world number one in DNA testing with plan to revolutionise fight against cancer and rare diseases*. Genomics England, 2014. https://www.genomicsengland.co.uk/uk-to-become-world-number-one-in-dna-testing-with-plan-to-revolutionise-fight-against-cancer-and-rare-diseases/
125. Genomics England. Genomics England adopts Edico Genome's DRAGEN Bio-IT Platform. 2018. https://www.genomicsengland.co.uk/genomics-england-adopts-edico-genomes-dragen-bio-it-platform/
126. Sundaram AY, Hughes T, Biondi S, *et al*. A comparative study of ChIP-seq sequencing library preparation methods. *BMC Genomics* 2016;**17**(1):816.
127. Quail MA, Smith M, Coupland P, *et al*. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012;**13**(1):341.
128. Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. *Nat Rev Genet* 2017;**18**(8): 473–84.
129. Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. *Nat Methods* 2010;**7**:479.
130. Schatz MC. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* 2009;**25**(11):1363–9.
131. Zhao S, Prenger K, Smith L, *et al*. Rainbow: a tool for large-scale whole-genome sequencing data analysis using cloud computing. *BMC Genomics* 2013;**14**:425.
132. Smith AD, Chung WY, Hodges E, *et al*. Updates to the RMAP short-read mapping software. *Bioinformatics* 2009;**25**(21):2841–2.
133. McPherson JD. Next-generation gap. *Nat Methods* 2009; **6(Suppl 11)**:S2–5.
134. van Dijk EL, Auger H, Jaszczyszyn Y, *et al*. Ten years of next-generation sequencing technology. *Trends Genet* 2014;**30**(9): 418–26.
135. Schiffthaler B, Kostadima M, Delhomme N, *et al*. Training in high-throughput sequencing: common guidelines to enable material sharing, dissemination, and reusability. *PLoS Comput Biol* 2016;**12**(6):e1004937.
136. HTS Teacher's Consortium. HTS training material repository. 2016. http://bioinformatics.upsc.se/htmr
137. Hoadley KA, Yau C, Wolf DM, *et al*. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 2014;**158**(4):929–44.
138. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;**490**:61–70.
139. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 2009;**25**(22):2906–12.
140. Ally A, Balasundaram M, Carlsen R, *et al*. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* 2017;**169**:1327–41.e23.
141. Ciucci S, Ge Y, Durán C, *et al*. Enlightening discriminative network functional modules behind principal component analysis separation in differential-omic science studies. *Sci Rep* 2017;**7**:43946.
142. Kuperstein I, Grieco L, Cohen DPA, *et al*. The shortest path is not the one you know: application of biological network resources in precision oncology research. *Mutagenesis* 2015; **30**(2):191–204.
143. Zhang W, Chien J, Yong J, *et al*. Network-based machine learning and graph theory algorithms for precision oncology. *NPJ Precis Oncol* 2017;**1**:25.
144. Burkard ME, Deming DA, Parsons BM, *et al*. Implementation and clinical utility of an integrated academic-community regional molecular tumor board. *JCO Precis Oncol* 2017;(1): 1–10.
145. Gupta A, Ayub M, Miller C, *et al*. 1628O Development of the Manchester Cancer Research Centre Molecular Tumour Board for matching patients to clinical trials based on tumour and ctDNA genetic profiling. *Ann Oncol* 2017;**28**: mdx390.
146. McKenna A, Hanna M, Banks E, *et al*. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**: 1297–303.
147. ScienceCloud. *A secure cloud solution*. ScienceCloud, 2017. https://www.sciencecloud.com/
148. DNAnexus. *DNAnexus*. DNAnexus, 2017. https://www.dnanexus.com/
149. Angiuoli SV, Matalka M, Gussman A, *et al*. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics* 2011;**12**:356.

150. Mushtaq H, Al-Ars Z. Cluster-based apache spark implementation of the GATK DNA analysis pipeline. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Washington, DC: IEEE, 2015, 1471–7. doi: 10.1109/BIBM.2015.7359893.
151. Wiewiórka MS, Messina A, Pacholewska A, *et al*. SparkSeq: fast, scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision. *Bioinformatics* 2014;**30**(18):2652–3.
152. Anderson TE, Culler DE, Patterson DA. Case for NOW (Networks of Workstations). *IEEE Micro* 1995;**15**(1):54–64.
153. Barak A, La'adan O. The MOSIX multicomputer operating system for high performance cluster computing. *Futur Gener Comput Syst* 1998;**13**(4–5):361–72.
154. Blayney J, Haberland V, Lightbody G, *et al*. Biomarker discovery, high performance and cloud computing: a comprehensive review. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2015, 1514–19.
155. The Apache Software Foundation. *Welcome to Apache*^{TM} *Hadoop*$^{®}$*!* The Apache Software Foundation, 2014. http://hadoop.apache.org/
156. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. In: *OSDI'04 Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation*, Vol. 51. 2004, 107–13.
157. Kawalia A, Motameny S, Wonczak S, *et al*. Leveraging the power of high performance computing for next generation sequencing data analysis: tricks and twists from a high throughput exome workflow. *PLoS One* 2015;**10**(5): e0126321.
158. Genomics England. The 100,000 genomes project. Genomics England, 2017. https://www.genomicsengland.co.uk/the-100000-genomes-project/
159. Yang X, Liu Y, Yuan C, *et al*. Parallelization of BLAST with MapReduce for long sequence alignment. In: *2011 Fourth International Symposium on Parallel Architectures, Algorithms and Programming (PAAP)*. Tianjin, China: IEEE, 2011, 241–6. https://ieeexplore.ieee.org/document/6128510/
160. Sul SJ, Tovchigrechko A. Parallelizing BLAST and SOM algorithms with MapReduce-MPI library. In: *2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW)*. 2011, 481–9.
161. Sundfeld D, Havgaard JH, Gorodkin J, *et al*. CUDA-Sankoff: using GPU to accelerate the pairwise structural RNA alignment. In: *2017 25th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*. 2017, 295–302.
162. Li D, Liu CM, Luo R, *et al*. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;**31**(10):1674–6.
163. Rahman MA, Muniyandi RC. Review of GPU implementation to process of RNA sequence on cancer. *Inform Med Unlocked* 2018;**10**:17–26.
164. Rowińska Z, Gocławski J. Cuda based fuzzy C-means acceleration for the segmentation of images with fungus grown in foam matrices. *Image Process Commun* 2012;**17**: 191–200.
165. Aji AM, Zhang L, Feng W. GPU-RMAP: accelerating short-read mapping on graphics processors. In: *2010 IEEE 13th International Conference on Computational Science and Engineering (CSE)*. 2010, 168–75.
166. Shi H, Schmidt B, Liu W, *et al*. A parallel algorithm for error correction in high-throughput short-read data on CUDA-enabled graphics hardware. *J Comput Biol* 2010;**17**(4):603–15.
167. Langmead B, Schatz MC, Lin J, *et al*. Searching for SNPs with cloud computing. *Genome Biol* 2009;**10**(11):R134.
168. Illumina. *BaseSpace Sequence Hub*. Illumina, 2017.
169. SevenBridges. *Actionable informatics for biomedical research*. Seven Bridges Genomics, 2018.
170. Ramdas T, Egan G. A survey of FPGAs for acceleration of high performance computing and their application to computational molecular biology. In: *TENCON 2005 2005 IEEE Region 10*. Melbourne, Qld., Australia: IEEE, 2005. doi: 10.1109/TENCON.2005.300963.
171. Chrysos G, Sotiriades E, Rousopoulos C, *et al*. Opportunities from the use of FPGAs as platforms for bioinformatics algorithms. In: *2012 IEEE 12th International Conference on Conference: Bioinformatics & Bioengineering (BIBE)*. 2012, 559–65.
172. Schmidt B, Hildebrandt A. Next-generation sequencing: big data meets high performance computing. *Drug Discov Today* 2017;**22**(4):712–17.
173. Dydel S, Bała P. Large scale protein sequence alignment using FPGA reprogrammable logic devices. In: J Becker, M Platzner, S Vernalde (eds). *Field Programmable Logic and Application. FPL 2004, Lecture Notes in Computer Science*, Vol. **3203**. Springer, Berlin, Heidelberg, 2004, 23–32.
174. Tan G, Zhang C, Tang W, *et al*. Accelerating irregular computation in massive short reads mapping on FPGA co-processor. *IEEE Trans Parallel Distrib Syst* 2016;**27**(5): 1253–64.
175. Sotiriades E, Dollas A. A general reconfigurable architecture for the BLAST algorithm. *J VLSI Signal Process Syst Signal Image Video Technol* 2007;**48**(3):189–208.
176. Segundo EJGN, Nedjah N, de Macedo Mourelle L. A scalable parallel reconfigurable hardware architecture for DNA matching. *Integr VLSI J* 2013;**46**(3):240–6.
177. Guo X, Wang H, Devabhaktuni V. A systolic array-based FPGA parallel architecture for the BLAST algorithm. *ISRN Bioinforma* 2012;**2012**:1–11.
178. Wang C, Li X, Zhou X, *et al*. Genome sequencing using mapreduce on FPGA with multiple hardware accelerators. In: *Conference: Proceedings of the ACM/SIGDA international symposium on Field programmable gate arrays*. ACM, 2013, 266.
179. Papadopoulos A, Kirmitzoglou I, Promponas VJ, Theocharides T. FPGA-based hardware acceleration for local complexity analysis of massive genomic data. *Integr VLSI J* 2013;**46**(3):230–9.
180. Goyal A, Kwon HJ, Lee K, *et al*. Ultra-fast next generation human genome sequencing data processing using DRAGEN Bio-IT processor for precision medicine. *Open J Genet* 2017; **7**(01):9–19. doi: 10.4236/ojgen.2017.71002.
181. Melanakos J. Parallel computing on a personal computer. *Biomed Comput Rev* 2008; http://www.bcr.org/content/parallel-computing-personal-computer
182. Fan Z, Qiu F, Kaufman A, *et al*. GPU cluster for high performance computing. In: *SC'04: Proceedings of the 2004 ACM/IEEE Conference on Supercomputing*. Pittsburgh, PA: IEEE, 2004, 47. doi: 10.1109/SC.2004.26.
183. Carr DA, Paszko C, Kolva D. *SeqNFind*$^{®}$*: a GPU accelerated sequence analysis toolset facilitates bioinformatics*. Nature Methods | Application Notes, 2011, 1–4. https://www.nature.com/app_notes/nmeth/2011/110908/full/an8037.html
184. Fujii Y, Azumi T, Nishio N, *et al*. Data transfer matters for GPU computing. In: *2013 International Conference on Parallel and Distributed Systems (ICPADS)*. 2013, 275–82.
185. NVIDIA. *CUDA GPUs*. NVIDIA Developer, 2017. https://developer.nvidia.com/cuda-gpus

186. Liu Y, Schmidt B, Maskell DL. Cushaw: a cuda compatible short read aligner to large genomes based on the Burrows-Wheeler transform. *Bioinformatics* 2012;**28**(14):1830–7.
187. Klus P, Lam S, Lyberg D, *et al.* BarraCUDA—a fast short read sequence aligner using graphics processing units. *BMC Res Notes* 2012;**5**(1):27.
188. Liu CM, Wong T, Wu E, *et al.* SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics* 2012; **28**(6):878–9.
189. Liu Y, Wirawan A, Schmidt B. CUDASW++ 3.0: accelerating Smith-Waterman protein database search by coupling CPU and GPU SIMD instructions. *BMC Bioinformatics* 2013;**14**:117.
190. Abadi DJ. Data management in the cloud: limitations and opportunities. *IEEE Data Engineering Bulletin* 2009;**32**:5–12.
191. AWS. *Amazon elastic compute cloud (EC2)*. AWS, 2017. https://aws.amazon.com/ec2/
192. Google Cloud Platform. Google genomics. Google Cloud Platform, 2017. https://cloud.google.com/genomics/
193. Microsoft Azure. *Microsoft Azure: cloud computing platform and services*. Microsoft Azure, 2017. https://azure.microsoft.com/
194. Granados Moreno P, Joly Y, Knoppers BM. Public–Private Partnerships in Cloud-Computing Services in the Context of Genomic Research. *Frontiers in Medicine* 2017;**4**:3.
195. UK Cloud. *Genomics England selects skyscape to support 100,000 Genomes Project*. UK Cloud, 2015.
196. Kwon T, Yoo WG, Lee WJ, *et al.* Next-generation sequencing data analysis on cloud computing. *Genes Genomics* 2015; **37**(6):489–501.
197. Field D, Tiwari B, Booth T, *et al.* Open software for biologists: from famine to feast. *Nat Biotechnol* 2006;**24**(7):801–3.
198. Woods R, McAllister J, Lightbody G, *et al. FPGA-Based Implementation of Signal Processing Systems*. West Sussex, UK: Wiley, 2017. https://www.wiley.com/en-us/FPGA+based+Implementation+of+Signal+Processing+Systems%2C+2nd+Edition-p-9781119077954
199. Xilinx. *Xilinx: Vivado design suite*. Xilinx, 2018. https://www.xilinx.com/products/design-tools/vivado.html
200. Intel. *Intel FPGA SDK for OpenCL—overview*. Intel, 2018. https://www.altera.com/products/design-software/embedded-software-developers/opencl/overview.html
201. Intel Altera. *Intel® FPGA Acceleration Hub—acceleration stack for Intel INTEL® FPGA Acceleration Hub—Xeon CPU with FPGAs*. Intel Altera, 2018. https://www.altera.com/solutions/acceleration-hub/acceleration-stack.html
202. Fieldman M. *Microsoft goes all in for FPGAs to build out AI cloud | TOP500 supercomputer sites*. Top500, 2016. https://www.top500.org/news/microsoft-goes-all-in-for-fpgas-to-build-out-cloud-based-ai/
203. AWS. *Amazon EC2 F1 instances—run customizable FPGAs in the AWS cloud*. AWS, 2018. https://aws.amazon.com/ec2/instance-types/f1/
204. Brandon MC, Wallace DC, Baldi P. Data structures and compression algorithms for genomic sequence data. *Bioinformatics* 2009;**25**(14):1731–8.
205. Batley J, Edwards D. Genome sequence data: management, storage, and visualization. *Biotechniques* 2009;**46**:333–6.
206. Vouzis PD, Sahinidis NV. GPU-BLAST: using graphics processors to accelerate protein sequence alignment. *Bioinformatics* 2011;**27**(2):182–8.
207. Pinho AJ, Pratas D. MFCompress: a compression tool for fasta and multi-fasta data. *Bioinformatics* 2014;**30**(1):117–18.
208. Qiao D, Yip WK, Lange C. Handling the data management needs of high-throughput sequencing data: speedGene, a compression algorithm for the efficient storage of genetic data. *BMC Bioinformatics* 2012;**13**(1):100.
209. Biji C, Nair A. Benchmark dataset for whole genome sequence compression. *IEEE/ACM Trans Comput Biol Bioinforma* 2016;**14**:1228–36.
210. Hsi-Yang Fritz M, Leinonen R, Cochrane G, *et al.* Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res* 2011;**21**(5):734–40.
211. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;**14**(2): 178–92.
212. Conesa A, Götz S, García-Gómez JM, *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005;**21**(18):3674–6.
213. Datta S, Bettinger K, Snyder M. Secure cloud computing for genomic data. *Nat Biotechnol* 2016;**34**(6):588–91.
214. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. *Nat Rev Genet* 2014;**15**(6):409–21.
215. Erlich Y, Williams JB, Glazer D, *et al.* Redefining genomic privacy: trust and empowerment. *PLoS Biol* 2014;**12**(11): e1001983.
216. Presidential Commission for the Study of Bioethical Issues. Privacy and progress in whole genome sequencing. Presidential Commission for the Study of Bioethical Issues, 2012. http://bioethics.gov/sites/default/files/PrivacyProgress508.pdf
217. McGuire AL, Basford M, Dressler LG, *et al.* Ethical and practical challenges of sharing data from genome-wide association studies: the eMERGE Consortium experience. *Genome Res* 2011;**21**(7):1001–7.
218. NHGRI. *Electronic Medical Records and Genomics (eMERGE) Network*. NHGRI, 2017. https://www.genome.gov/27540473/electronic-medical-records-and-genomics-emerge-network/
219. US Department of Health and Human Services. Health Insurance Portability and Accountability Act of 1996. *US Statut Large* 1996;**110**:1936–2103.
220. Office for Civil Rights. *Guidance Regarding methods for de-identification of protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule*. Office for Civil Rights, 2012. http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
221. US Congress. *H.R.493—110th Congress (2007-2008): genetic information nondiscrimination act of 2008*. US Congress, 2008. https://www.congress.gov/bill/110th-congress/house-bill/00493
222. European Commission. *Reform of EU data protection rules*. European Commission. 2016. http://ec.europa.eu/justice/data-protection/reform/index_en.htm
223. Chassang G. The impact of the EU general data protection regulation on scientific research. *Ecancermedicalscience* 2017; **11**:709.
224. Schatz M, Langmead B, Salzberg S. Cloud computing and the DNA data race. *Nat Biotechnol* 2010;**28**(7):691–3.
225. AWS. *Cloud compliance—Amazon Web Services (AWS) compliance*. AWS, 2018. https://aws.amazon.com/compliance/
226. Microsoft Azure. *Big compute: HPC and batch large-scale cloud computing power on demand*. Microsoft Azure, 2017. https://azure.microsoft.com/en-gb/solutions/big-compute/
227. BC Platforms. *BC platforms—software platforms for next-generation sequencing*. BC Platforms, 2017. http://bcplatforms.com/

228. NCBI. *GenBank home*. NCBI, 2017. https://www.ncbi.nlm.nih.gov/genbank/
229. INSDC. *International nucleotide sequence database collaboration*. INSDC, 2017. http://www.insdc.org/
230. DDBJ. *DNA Data Bank of Japan*. DDBJ, 2017. http://www.ddbj.nig.ac.jp/
231. ENA. *European nucleotide archive*. ENA, 2018. http://www.ebi.ac.uk/ena
232. Benson DA, Cavanaugh M, Clark K, *et al*. GenBank. *Nucleic Acids Res* 2013;**41**(Database issue):D36–D42.
233. EMBL-EBI. *ArrayExpress—functional genomics data*. EMBL-EBI, 2017. https://www.ebi.ac.uk/arrayexpress/
234. Edgar R, Barrett T. NCBI GEO standards and services for microarray data. *Nat Biotechnol* 2006;**24**(12):1471–2.
235. Skloot R. *The immortal life of Henrietta Lacks*. New York, NY: Random House Inc., 2010. p. 384. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2898613/
236. Parry W. Controversial 'HeLa' cells: use restricted under new plan. 2013, www.LiveScience.com.
237. Landry JJM, Pyl PT, Rausch T, *et al*. The genomic and transcriptomic landscape of a HeLa cell line. *G3* 2013;**3**(8):1213–24.
238. Samuels ME, Orr A, Guernsey DL, *et al*. Is gene discovery research or diagnosis? *Genet Med* 2008;**10**(6):385–90.
239. Walcott FL, Dunn BK. Legislation in the genomic era: the affordable care act and genetic testing for cancer risk assessment. *Genet Med* 2015;**17**(12):962–4.
240. US Congress. *Text—H.R.1313—115th Congress (2017-2018): preserving employee wellness programs act*. US Congress, 2017.
241. Sun LH. Employees who decline genetic testing could face penalties under proposed bill. *The Washington Post*, 2017.
242. Herper M. *Illumina promises to sequence human genome for $100—but not quite yet*. Forbes, 2017. https://www.forbes.com/sites/matthewherper/2017/01/09/illumina-promises-to-sequence-human-genome-for-100-but-not-quite-yet/#58262050386d
243. Heger M. *Illumina unveils new high-throughput sequencing instrument at JP Morgan*. GenomeWeb, 2017. https://www.genomeweb.com/sequencing/illumina-unveils-new-high-throughput-sequencing-instrument-jp-morgan
244. AWS. *Architecting for HIPAA security and compliance on Amazon Web Services*. AWS, 2015. https://aws.amazon.com/compliance/hipaa-compliance/
245. Kühnemund M, Wei Q, Darai E, *et al*. Targeted DNA sequencing and in situ mutation analysis using mobile phone microscopy. *Nat Commun* 2017;**8**:13913.
246. Schatz MC, Langmead B. The DNA data deluge: fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze. *IEEE Spectr* 2013;**50**:26–33.
247. Endrullat C, Glökler J, Franke P, *et al*. Standardization and quality management in next-generation sequencing. *Appl Transl Genomics* 2016;**10**:2–9.
248. Shoenbill K, Fost N, Tachinardi U, *et al*. Genetic data and electronic health records: a discussion of ethical, logistical and technological considerations. *J Am Med Inform Assoc* 2014;**21**(1):171–80.
249. Li G, Bankhead P, Dunne PD, *et al*. Embracing an integromic approach to tissue biomarker research in cancer: perspectives and lessons learned. *Brief Bioinform* 2017;**18**:634–46.
250. AstraZeneca. *Research-based BioPharmaceutical Company*. AstraZeneca UK, 2017. https://www.astrazeneca.co.uk/
251. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting Methods. *J Natl Cancer Inst* 2007;**99**(2):147–57.
252. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol* 2005;**23**(29):7332–41.
253. Human Proteome Organisation. *The HUPO proteomics standards initiative*. Human Proteome Organisation, 2018. https://hupo.org/Proteomics-Standards-Initiative