doi:10.1038/nature09796

A decade's perspective on DNA sequencing technology

Elaine R. Mardis[1]

The decade since the Human Genome Project ended has witnessed a remarkable sequencing technology explosion that has permitted a multitude of questions about the genome to be asked and answered, at unprecedented speed and resolution. Here I present examples of how the resulting information has both enhanced our knowledge and expanded the impact of the genome on biomedical research. New sequencing technologies have also introduced exciting new areas of biological endeavour. The continuing upward trajectory of sequencing technology development is enabling clinical applications that are aimed at improving medical diagnosis and treatment.

The sequencing of the Human Reference Genome, announced ten years ago, provided a roadmap that is the foundation for modern biomedical research. This monumental accomplishment was enabled by developments in DNA sequencing technology that allowed data production to far exceed the original description of Sanger sequencing[1]. Moving forward in the genomic era in which we now find ourselves, new (or 'next generation') DNA sequencing technology is enabling revolutionary advances in our understanding of health and disease. In essence, sequencing technology is the engine that powers the car that allows us to navigate the human genome roadmap. As that engine becomes ever more powerful, so will the questions we can ask and answer about the geography of our genetic landscape.

Of course, a car with only an engine is unworkable; as such, DNA sequencing technology provides an integral part of a larger system, one with multiple components that must be properly matched in order to achieve high throughput and efficiency. It has essentially never been as 'easy' as simply buying sequencing instruments, plugging them in, and generating data. We need the raw materials, such as fuel (DNA), sparks to ignite the fuel (reagents), mechanical parts to translate fuel and ignition into movement (robotics) and direction (bioinformatics), all working in a carefully engineered balance, and a driver (genome centre) to steer the automobile quickly and efficiently to the desired destination (biological understanding). By inference, as this 'engine' has achieved ever increasing horsepower, the supporting components have evolved to match its output with corresponding levels of performance, and new or completely revised components have been added as required.

In 2001, the technology that sequenced the human genome was based on capillary electrophoresis of individual fluorescently labelled Sanger sequencing reaction products. Each instrument could detect 500–600 bases from each of 96 reactions in around ten hours, with 24-hour unattended operation producing 115 kbp (thousand base pairs) per day. Because of the increased scale required for the Human Genome Project, genome centres had developed a robust, highly automated and inexpensive preparatory process to feed their capillary sequencers. Once the data were produced, mature analysis software was applied to analyse the sequencing reads (each a ~500-bp sequence of A, C, G, T), then to assemble reads that shared sequence identity, reproducing that region of the genome. After assembly, each genomic region was further analysed to identify genes, repeat elements and other features. As the 'drivers' of these sequencing pipelines, genome centres could dial up capacity by increasing the amount of hardware used in the preparatory and sequencing processes, because sequence production, not sequence analysis, was rate limiting.

As I will describe, the ensuing ten years has been marked by dramatic improvements in sequencing technology that have catapulted sequencing to the forefront of biological experimentation and have revolutionized the way that we approach genome-wide questions. One consequence of this revolution has been the coincident revitalization of bioinformatics, predominantly in development efforts aimed at data analysis and interpretation. Taken together, these unprecedented sequencing and analysis capabilities have inspired new areas of enquiry, have solved major questions about the regulation, variability and diaspora of the human genome, and have introduced a genomic era in medical enquiry and (ultimately) practice that will bring about the originally envisioned impact of the Human Genome Project.

Massively parallel sequencing

The first five years following the Human Genome Project provided further definition and annotation of the human genome sequence by comparative genomics; the sequencing of several model organism genomes—such as mouse[2], rat[3], chicken[4], dog[5], chimpanzee[6], rhesus macaque[7], duckbill platypus[8] and cow[9]—provided information about highly conserved genomic elements that are likely to be functional owing to their conservation. These genomes were largely produced by conventional methods, including Sanger-based capillary sequencing. Starting in 2005, a variety of new 'engines' for DNA sequencing that were radically different from the capillary sequencers used to sequence the human and model organism genomes became available from several different manufacturers (Fig. 1). These new engines were 'turbo-charged' by several orders of magnitude compared to their predecessors, because the basic mechanisms for data generation had changed radically, producing far more sequence reads per instrument run and at a significantly lower expense. The availability of multiple commercially available instruments alone represented a paradigm shift from the previous decade, where a single capillary instrument produced by Applied Biosystems dominated the market. Many of these innovative approaches were initially developed with National Institutes of Health (NIH) funding through the 'Technology development for the $1,000 genome' program (http://www.genome.gov/11008124#al-4) introduced during Francis Collins' directorship at the National Human Genome Research Institute (NHGRI).

Since the introduction of these platforms, the past five years have been marked by fierce competition between their manufacturers to greatly

[1]The Genome Center at Washington University School of Medicine, Department of Genetics, Washington University School of Medicine, St Louis, Missouri 63108, USA.

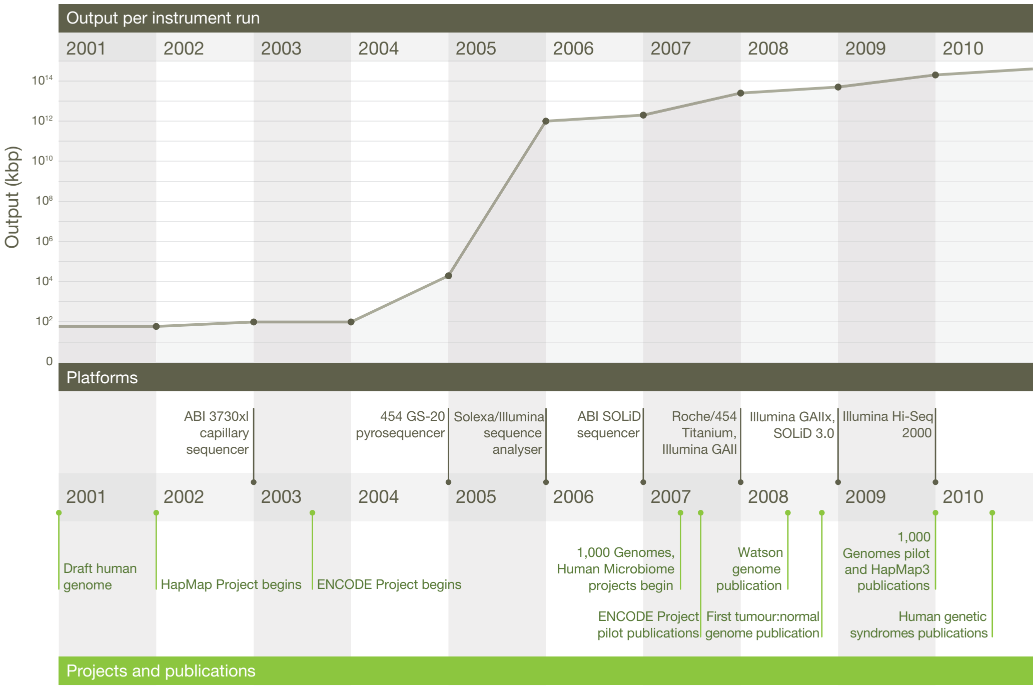

Figure 1 | Changes in instrument capacity over the past decade, and the timing of major sequencing projects. Top, increasing scale of data output per run plotted on a logarithmic scale. Middle, timeline representing major milestones in massively parallel sequencing platform introduction and instrument revisions. Bottom, the timing of several projects and milestones described in the text.

increase the amount of sequence output per run, to increase read lengths (the number of nucleotides per sequence read), to lower costs, and to improve base-calling accuracy for their instruments. Much like a buyer of a new car, genome centres have taken each new 'massively parallel' (see below) instrument for a 'test drive' to explore their performance and to understand their strengths and weaknesses in terms of data quality, associated production processes, and actual cost to operate. (This last includes personnel, consumables, additional instrumentation to operate, amortized equipment rate, and informatics infrastructure.) Their collective efforts to vet each new technology and report their findings by scientific presentations, press releases, word of mouth, and peer-reviewed manuscripts have effectively fuelled the rivalry and competition among the instrument vendors and have resulted, not surprisingly, in both winners and losers in this commercial sector.

This so-called 'massively parallel' sequencing technology differs significantly from Sanger capillary sequencing. Although each instrument is distinctly different in its specifics, as detailed in several reviews[10,11] (see also Table 1), all massively parallel devices share certain attributes, as follows. First, the initial preparatory steps are fewer and simpler to perform than for Sanger sequencing. Instead of a bacterial cloning step followed by DNA isolation, massively parallel sequencing begins with the production of a library formed by ligating platform-specific synthetic DNAs (adapters) onto the ends of the fragment population to be sequenced. Second, all platforms require the library fragments to be amplified on a solid surface (either a glass slide or a microbead) by a polymerase-mediated reaction that produces many copies of each single library fragment. Amplification is needed so that the ensuing sequencing reactions produce sufficient signal for detection by the instrument's optical system. However, this step also provides a source of sequencing error that is perpetuated through the downstream processes, because polymerases are never 100% accurate. Third, these instruments perform sequencing reactions as an orchestrated series of repeating steps that are performed and detected automatically. The specifics of the DNA sequencing reaction are different for each platform, emphasizing the amazing range of innovation in chemistry, molecular biology and engineering required to produce sequence information from hundreds

Table 1 | Sequencing platform comparison

	Roche/454	Life Technologies SOLiD	Illumina Hi-Seq 2000	Pacific Biosciences RS
Library amplification method	emPCR* on bead surface	emPCR* on bead surface	Enzymatic amplification on glass surface	NA (single molecule detection)
Sequencing method	Polymerase-mediated incorporation of unlabelled nucleotides	Ligase-mediated addition of 2-base encoded fluorescent oligonucleotides	Polymerase- mediated incorporation of end-blocked fluorescent nucleotides	Polymerase-mediated incorporation of terminal phosphate labelled fluorescent nucleotides
Detection method	Light emitted from secondary reactions initiated by release of PPi	Fluorescent emission from ligated dye-labelled oligonucleotides	Fluorescent emission from incorporated dye-labelled nucleotides	Real time detection of fluorescent dye in polymerase active site during incorporation
Post incorporation method	NA (unlabelled nucleotides are added in base-specific fashion, followed by detection)	Chemical cleavage removes fluorescent dye and 3′ end of oligonucleotide	Chemical cleavage of fluorescent dye and 3′ blocking group	NA (fluorescent dyes are removed as part of PPi release on nucleotide incorporation)
Error model	Substitution errors rare, insertion/ deletion errors at homopolymers	End of read substitution errors	End of read substitution errors	Random insertion/deletion errors
Read length (fragment/paired end)	400 bp/variable length mate pairs	75 bp/50+25 bp	150 bp/100+100 bp	>1,000 bp

Comparison of commercially available next generation platforms (Roche/454, Life Technologies and Illumina) and a single molecule platform (Pacific Biosciences), illustrating the similarities and differences in these technologies, according to several metrics. NA, not applicable; PPi, pyrophosphate.

* emPCR (emulsion PCR) is a bulk amplification process whereby library fragments are combined with beads and PCR reactants in an oil emulsion that allows *en masse* amplification of millions of bead-DNA combinations in a single tube.

of thousands to hundreds of millions of DNA molecules simultaneously. For example, the Roche/454 instrument detects each polymerase-catalysed nucleotide incorporation event by a downstream series of reactions that produce light ('pyrosequencing'), initiated by the pyrophosphate molecules released on nucleotide incorporation. The Life Technologies SOLiD uses a unique DNA ligase-mediated process that, through multiple rounds of template-directed ligation, sequences each nucleotide twice. The Illumina sequencer incorporates fluorescently labelled nucleotides that are chemically blocked such that only one nucleotide incorporation event occurs per fragment population per sequencing cycle. Regardless of the details, massively parallel sequencing reactions are distinguished by the fact that they occur in a nucleotide-by-nucleotide stepwise fashion, rather than by discrete separation and detection (in a 96-at-a-time fashion) of already produced Sanger sequencing reaction products on a capillary instrument. The fourth shared feature of these systems is the ability to obtain sequence information from both the ends of the DNA fragments comprising the sequencing library. Depending on the instrument system and the library construction approach used, one can either sequence at both ends of linear fragments ('paired end sequencing') or from both ends of previously circularized fragments ('mate pair sequencing').

Paired end sequencing libraries are the standard means by which human genomes are sequenced, because they are straightforward to make and require a small amount of DNA. Mate pair libraries, by contrast, are quite DNA expensive owing to the low yield of circularization of large DNA molecules (a yield that diminishes proportionally with increasing length of the DNA molecules used). However, mate pair libraries provide valuable information about larger structural events because they sample DNA sequence over a larger distance (1.5–20 kilobases, kb) than do paired end libraries (~300–500 bp). The benefit of obtaining sequence data from both ends of library fragments in human genome sequencing is obvious when one considers the highly repetitive nature of the genome. Explicitly, aligning at least one end read of a pair uniquely onto the reference sequence provides sufficient certainty that the read pair is uniquely mapped to its locus of origin. Conversely, aligning short, single end or 'fragment' reads to the genome results in a higher proportion of non-unique placements—reads that cannot be used for variant discovery. As described later, the other value of paired end data lies in its use for discovering structural variation in the genome.

Although massively parallel platforms have significantly affected our ability to study the human genome and to better understand its variability in a multitude of contexts, these technologies have required profound changes to the data analysis pipelines that previously had been so straightforward. In particular, the new sequencing engines have introduced data analysis challenges owing to the massive scale of the data to be analysed, the significant decrease in the read length, and in the dramatically different error profiles of each read type, when compared to those of capillary data. These new challenges have resulted in a revitalization of the bioinformatics-based pursuit of sequence data analysis at all levels. This renaissance can be attributed to the attractiveness of the analysis challenges introduced by large data sets and to the fact that an increasing number of compelling biological questions are now approachable with only a few experiments' worth of data, owing to the greatly increased scale and significantly lower cost of massively parallel sequencing. But whereas the data generation is straightforward, often a corresponding analytical approach to the derivation of answers is not. This fact has forged alliances between experimentalists and computational biologists as never before in genomic science, and emphasizes both the enhanced capabilities and analytical difficulties brought about by massively parallel sequencing, in contrast to the technology initially used to chart the human genome.

Defining our genomic roadmap

Variations in our sequence roadmap

Once the Human Reference Genome was in-hand[12], the efforts of the International Human Genome Project teams turned to completing all regions of the genome to high quality[13] on a chromosome-by-chromosome basis. Subsequently, many of these same laboratories began efforts to identify the positions of common single-nucleotide polymorphisms (SNPs) known to exist in the genome. The international SNP discovery efforts were known as the 'HapMap' projects, because they aimed to map the haplotype diversity in the human genome. These projects again required a concerted effort across many laboratories[14–17] to characterize common SNP variation (present at 5% or greater allele frequency for the population studied) in multiple human populations.

The HapMap efforts culminated in the identification of more than 8 million common SNP positions genome-wide, most of which were generated by Sanger-based capillary methods. These were added to the dbSNP public database at NCBI (the National Center for Biotechnology Information), and so represented an important reference addendum that further emphasized the intricate genomic roadmap of individuals from various ethnic backgrounds. In addition, many HapMap SNP positions were used to increase the density of common variant sites on commercial SNP genotyping arrays, producing a research tool with which human geneticists began to evaluate large case-control cohorts for common complex disease studies. These genome-wide association studies were designed to test the 'common variants common disease' hypothesis, identifying specific loci that were associated with the occurrence of a common disease (that is, predominant in cases but not controls). Although these approaches have succeeded to various extents in identifying disease-associated SNPs[18], the likely contribution that rare or 'private' SNPs make to disease susceptibility is now being investigated by combining massively parallel sequencing with case-control cohorts. One such study has shown this approach to be particularly effective in identifying rare variants found only in the genomes of affected individuals (cases) that explain the biology of the disease (in this case, extreme obesity)[19].

Beyond SNPs, multiple efforts have explored the breadth of human genome diversity, demonstrating that as individuals, our roadmaps differ in many ways beyond single nucleotide differences[20,21]. For example, there are a myriad of small- and large-scale differences in genomes, including focused insertion and deletion events (indels) genome-wide that add one to several nucleotides per event, amplification or deletion events that result in increased or decreased numbers of copies of specific genome segments (copy number polymorphisms), changes in the orientation of genome segments (inversions), and novel genome content (insertions). Although most such large-scale variations originally were observed using microarray-based methods[22,23], several groups have demonstrated the ability to use information from paired end or mate pair data sets towards high-resolution characterization of all classes of structural variation in the human genome[24–26]. Structural variant discovery is achieved by examining the separation and orientation of aligned read pairs on the reference. Namely, if groups of read pairs are identified with interpair distances that are further apart or closer together than expected based on the size of inserts used in constructing the library, or with the forward and reverse reads aligned either in an unexpected orientation, or onto different chromosomes, these provide evidence for structural variation relative to the reference genome.

The 1,000 Genomes Project is adding resolution and new information to our understanding of genome diversity across all levels of variation[27,28]. It combines the scale and cost of massively parallel human genome re-sequencing and analysis with many of the populations already studied in the HapMap projects as well as newly consented populations. To date, the Roche/454, Life Technologies SOLiD and Illumina platforms have been used for data production in this project. When the 1,000 Genomes Project is completed in 2012, information will be in hand regarding SNP, indel and structural variation for more than 2,000 individuals, and will be available through public databases to further enrich the detail of our human genome roadmap. It goes without saying that such a feat would not have been possible without the availability of massively parallel sequencing, but most will not fully appreciate the multitude of algorithmic and bioinformatic innovations required to fully mine this rich data set to its fullest.

Another area of biological enquiry that has been aided by sequencing technology is that of identifying genes affected by mutation in cancer tissue genomes. When my centre and others initially began sequencing candidate gene lists in specific cancer samples in the early 2000s, our approaches consisted of designing polymerase chain reaction (PCR) primer sets specific to the genes we thought would be mutated, amplifying those from the genomic DNA of the tumour, and sequencing with capillary-based instruments[29–31]. The bioinformatics-based tools for identifying these mutations were largely modified from existing tools that were originally used to sequence the human genome. Although important discoveries were made during these early years in cancer genomics, the approaches were slow and expensive, and our enquiries were limited to genes whose mutations 'made sense' in terms of what already was known about tumour cell biology. With massively parallel sequencing, emerged the ability to pool the reaction products of thousands of PCRs and sequence them all at once, thereby reducing the cost of sequencing and dramatically increasing the rapidity with which the data could be obtained. By aligning the resulting sequence reads to the Human Genome Reference, and modifying existing algorithms for identifying mutations, mutation discovery was facilitated as well.

Shortly after these approaches were developed and published, my centre used Illumina massively parallel sequencing to sequence the complete tumour and normal genomes from a patient diagnosed with acute myeloid leukaemia (AML), and then developed methods to comparatively analyse these two genomes, identifying tumour-unique (somatic) alterations in the process[32]. This effort required that we develop entirely new bioinformatics-based methods to do the following: (1) ensure that we had completely sequenced the genomes to the depth and breadth needed to then identify and compare the millions of single-nucleotide differences identified in both the tumour and normal genomes (human genomes typically have about one difference per 1,000 bases when we compare any human genome data set to the reference genome sequence), and (2) ultimately, sort out the handful (typically 3,000–10,000) that are somatic, or unique to the tumour genome. Although an expensive endeavour at the time (we estimate the combination of data production and of novel bioinformatics tools development totalled $1.6 million for the first tumour/normal pair), we and others have subsequently sequenced and analysed hundreds of human genomes using primarily Illumina and Life Technologies platforms, as the cost per genome has plummeted and the sequence data output per instrument run has increased by 100-fold. Moreover, additional algorithmic developments have combined with paired end data to reveal somatic structural variations, both for focused (small numbers of inserted or deleted nucleotides) and large events (such as inter-chromosomal translocations) and to improve our ability to find point mutations. With the newest massively parallel instruments from Illumina, the data production for each tumour and normal pair can be completed in about 8 days on a single instrument at a fully loaded cost per pair of around $30,000. Although our analysis methods continue to be refined, the comprehensive data analysis required to characterize these paired genomes remains the most expensive and difficult aspect of whole genome re-sequencing by massively parallel methods[33,34].

The difficulties in the data analysis mentioned above are due to many factors, including the size and complexity of the human genome, the ever-changing read length and accuracy of next-generation sequencing data, and the computational demands needed to compute the full range of variant detection with the highest possible accuracy. In spite of refinements to these analytical methods, it is still an important and necessary step to perform orthogonal validation of discovered variants before reporting them, which further adds to the cost and time required for whole genome comparative methods.

Variations in our functional roadmap

Another benefit of these new engines is that they are allowing biologists to explore to unprecedented depths the specific differences in DNA regulation that define each tissue's biological roadmap. In fact, massively parallel sequencing is permitting us to answer many fundamental questions that were previously too expensive to perform at a genome-wide scale. For example, the changing associations of histones and chromosomal DNA during embryonic development, the exact placement of regulatory DNA-binding proteins on genomic DNA, the genome-wide methylation of chromosomal DNA, and the attendant alterations in gene expression levels associated with such events all can be investigated by combining an appropriate experimental front-end with a massively parallel read-out. The extent to which any or all of these measures change due to specific stimuli, over developmental time or in different tissue types also can be ascertained. One such effort aimed at performing these characterizations is the ENCODE (encyclopedia of DNA elements) Project[35]. This project was started in 2003 and used microarray-based assays in the pilot phase, but has now moved to using massively parallel methods owing to their reduced cost, and increased resolution and speed. Such genome-wide characterization capabilities are somewhat analogous to being able to drive further and faster than ever before, while charting the geography along the routes travelled at an unprecedented level of detail. Although the scope is breathtaking, each type of experiment has a list of shared and unique considerations that the bioinformatics analysis must take into account in order to separate true signal from noise, and to deliver an accurate genomic roadmap. At the beginning of each analytical approach, these experiments all require the alignment of sequence reads onto the Human Reference Genome sequence—effectively an assignment to the chromosomal locus of origin for each DNA or RNA fragment obtained from the experiment performed.

Thus, we are using the reference in the way it was intended, as a guide to help us discover information about the human genome, and its function, regulation and alteration in health and disease. Ultimately, these experiments will enable a transformation of biomedical research and medical practice—a transformation that already has begun[18,36].

Charting new territories

In addition to their significant impact on our understanding of human biology, massively parallel sequencing technologies have enabled new areas of genomic enquiry that also are germane to human health. One such area is termed 'metagenomics'—essentially, this term describes the sequencing-based characterization of DNA or RNA isolated from a mixed organism population sample obtained from its natural habitat. In human biological enquiry, metagenomic studies of the human body seek to characterize the content and complexity of microbial, viral and/or non-human eukaryotic organisms obtained from external (skin) and internal (colon, vagina) surfaces.

Depending on the body site and sampling method, variable amounts of human DNA and RNA can be simultaneously isolated, complicating the subsequent analysis to varying degrees. Although there are metagenomic studies that pre-date the availability of massively parallel sequencing instruments, the field has been transformed by rapid, inexpensive and abundant sequence data production and facile preparatory methods offered by next-generation platforms that are especially suitable for these complex genomic samples. As described earlier, the analytical challenges of these data sets have elicited an enormous amount of interest, not only in developing analysis approaches to obtain the biological information that can be mined from them, but also in applying this information to answer the fundamental questions posed by metagenomic enquiry, namely 'what's there?', 'what roles are they playing?' and 'how does the population change when the environment changes?'[37]. Often, the answers to these questions lie in examining metagenomic sequences by six-frame amino-acid translation, followed by database searching with the resulting massive data set. In this regard, longer sequence reads produce longer amino acid sequences, and hence a higher probability of correctly identifying the population members and/or their metabolic capabilities. Because the Roche/454 platform has longer read lengths than either Illumina or Life Technologies SOLiD, it has been the platform of choice for metagenomics studies.

The search for unknown aetiological agents in human disease has been facilitated by massively parallel sequencing. The incredible depth of sequencing that can be applied to a given sample, such as stool samples from an outbreak of diarrhoea, or DNA isolated from a sarcoid tumour, enables the identification of novel viruses or bacteria whose DNA or RNA will be coincidently isolated and sequenced with that of the human samples[38–40]. Again, the ability to detect and identify the common agent's DNA signature among the many host-derived reads requires a concerted and systematic approach to sequence data analysis. Here, longer read lengths provide more facile detection of non-human sequences and also are more straightforward to assemble, often allowing reconstruction of the aetiological agent's genome.

Finally, personal genomics—the sequencing of an individual's genome for health-related enquiry or for determining genetic predisposition to disease—emerged as an application of massively parallel sequencing once costs began to drop. James Watson, who shared the 1962 Nobel Prize in Physiology or Medicine with Francis Crick and Maurice Wilkins for discovering the chemical structure of DNA, was the first person to have his genome sequenced by massively parallel methods using the Roche/454 platform[41]. Other personal genomics have included solving the causative mutations in the familial Charcot-Marie-Tooth syndrome of James Lupski[42], a noted human geneticist, and in elucidating the mutation responsible for two siblings afflicted with Miller syndrome[43].

Beyond these examples, there are several interesting areas to which massively parallel sequencing might be applied, effectively furthering some of the early roadmap efforts mentioned above, as well as opening interesting areas of enquiry that were previously not possible[36]. These include (1) identifying the genomic differences between chromosomal and mitochondrial DNA derived from different tissues in a single human body, (2) establishing the gene expression profiles and patterns of all developmental and adult human tissues, (3) defining the spectrum of temporal changes in DNA methylation and histone-binding patterns of these same tissues, and (4) identifying the non-coding RNA expression profiles and their variation in human tissues.

By combining massively parallel sequencing with new whole genome DNA amplification approaches, we might also anticipate sequencing the complete genomes of single cells. Perhaps one of the most exciting possibilities addressed by this capability would permit an understanding of the genomic differences between individual tumour cells in a heterogeneous solid tumour type, such as breast cancer.

The future of sequencing

The amazing acceleration in biological enquiry enabled by the current massively parallel instrumentation is clearly just beginning. These instruments will continue to evolve, and new platforms just introduced, or under development, will have a continuing impact on biomedical research for years to come. What can we anticipate about the near-term expansion of applications using these instruments? One obvious area of expansion will be the use of massively parallel sequencing for 'genome-guided' medicine. This would involve using the speed and scope of new sequencing technologies and data analysis for diagnosis: targeted (specific genes) or whole-genome sequencing would be used to characterize the individual patient's disease, and to determine potential treatment modalities based on these data. We already have an example of this genome-guided approach; we have used whole genome sequencing and analysis to diagnose an AML patient thought to have acute promyelocytic leukaemia whose pathological diagnosis did not conform to the diagnosis obtained by cytogenetic assays, and we identified an uncommon chromosomal insertion event with a net result that mimics the common translocation (J. Welch *et al.*, manuscript submitted). Other groups have reported similar studies where massively parallel sequencing and analysis delivered answers to patients that aided their diagnosis and treatment for cancer[44] or identified the mutated gene responsible for causing a rare syndromic disease[43,45–47].

Certainly, one of the complications for certain types of genomic diagnoses using massively parallel instruments will be the time required to generate sequencing data (including preparatory steps and sequencing) and to completely analyse, validate and interpret these data in a medical context. The highest capacity instruments currently available require 8–14 days to produce data. Unfortunately, these run times and the ensuing analysis may not permit the return of information in a suitable time frame, relative to the patient's need for a diagnosis.

However, interesting possible solutions to the temporal limitations of certain diagnostic applications are anticipated from the newest massively parallel systems, at present in various stages of development or in early commercial release. As these systems are capable of delivering sequencing data from single molecules of DNA as they are being sequenced—rather than as a stepwise series of nucleotide addition steps that are analysed after the sequencing instrument has finished—the time for the sequence data generation step is shortened significantly relative to next-generation systems. One such instrument from Pacific Biosciences that is being tested in early access sites monitors each one of an array of individual polymerases while DNA synthesis is occurring, in order to obtain the single molecule sequences in a minimum of 30 minutes. Other instruments in development, including those by Oxford Nanopore or IBM/Roche, use nanopore technology to identify individual DNA nucleotides as the DNA fragment passes through the nanopore, by one of several detection approaches. Although the current capacities of real-time sequencers would not permit whole human genome sequencing in a single run, the near-term application of these instruments could be on focused evaluation of specific human genes or on the genomes of aetiological agents[48] for diagnosis, prognosis or therapeutic prescription.

To be clear, although the data generation steps are relatively quick, they must be properly coupled with equally rapid sample preparation and library construction steps and with refined, time-effective sequence analysis and interpretation appropriate for the clinical setting. In this regard, single molecule systems appear to be capable of delivering read lengths that are more akin to our 2001 era capillary instruments. Indeed, once the error models for each device are well characterized, producing longer reads may return us to the level of facility in data analysis for diagnostic medicine applications that we once enjoyed while sequencing the human genome. Although this would eliminate an important bottleneck in analysis, there are multiple aspects to sort out before diagnostic sequencing becomes commonplace. Nonetheless, the future of genomic medicine via massively parallel sequencing seems imminent. As time progresses, our 'engines' continue to improve in their sophistication and power, further enabling us to explore the human genome roadmap in our continuing journey to improve human health.

1. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74,** 5463–5467 (1977).
2. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
3. Gibbs, R. A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
4. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432,** 695–716 (2004).
5. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–819 (2005).
6. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
7. Gibbs, R. A. *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316,** 222–234 (2007).
8. Warren, W. C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453,** 175–183 (2008).
9. Elsik, C. G. *et al.* The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324,** 522–528 (2009).
10. Mardis, E. R. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. *Genome Med* **1,** 40 (2009).
11. Metzker, M. L. Sequencing technologies - the next generation. *Nature Rev. Genet.* **11,** 31–46 (2010).
12. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
13. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
14. The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
15. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).

16. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
17. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
18. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* doi:10.1038/nature09792 (this issue).
19. Harismendy, O. *et al.* Population sequencing of two endocannabinoid metabolic genes identifies rare and common regulatory variants associated with extreme obesity and metabolite level. *Genome Biol.* **11,** R118 (2010).
    First demonstration that rare sequence variants could be identified using next-generation sequencing in well-phenotyped cases and controls, and that functional significance in the phenotype could be assigned to the suspect variants.
20. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453,** 56–64 (2008).
21. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet.* **37,** 727–732 (2005).
22. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–454 (2006).
23. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).
24. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–426 (2007).
25. Chen, K. *et al.* BreakDancer: an algorithm for high resolution mapping of genomic structural variation. *Nature Methods* **6,** 677–681 (2009).
26. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genet.* **41,** 1061–1067 (2009).
27. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330,** 641–646 (2010).
    Initial structural variation data analysis resulting from 1,000 Genomes Project data, demonstrating the yield of such information from a large-scale project using next-generation sequencing.
28. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
29. Pao, W. *et al.* EGF receptor gene mutations are common in lung cancers from ''never smokers'' and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc. Natl Acad. Sci. USA* **101,** 13306–13311 (2004).
30. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314,** 268–274 (2006).
31. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–1113 (2007).
32. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).
33. Mardis, E. R. Cancer genomics identifies determinants of tumor biology. *Genome Biol.* **11,** 211 (2010).
34. Mardis, E. R. The $1,000 genome, the $100,000 analysis? *Genome Med* **2,** 84 (2010).
35. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
36. Green, E. D., Guyer, M. S. & National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* doi:10.1038/nature09764 (this issue).
37. Nelson, K. E. *et al.* A catalog of reference genomes from the human microbiome. *Science* **328,** 994–999 (2010).
38. Loh, J. *et al.* Detection of novel sequences related to African Swine Fever virus in human serum and sewage. *J. Virol.* **83,** 13019–13025 (2009).
39. Presti, R. M. *et al.* Quaranfil, Johnston Atoll, and Lake Chad viruses are novel members of the family Orthomyxoviridae. *J. Virol.* **83,** 11599–11606 (2009).
40. Finkbeiner, S. R. *et al.* Identification of a novel astrovirus (astrovirus VA1) associated with an outbreak of acute gastroenteritis. *J. Virol.* **83,** 10836–10839 (2009).
41. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
42. Lupski, J. R. *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.* **362,** 1181–1191 (2010).
    Personal genome sequencing used to identify a rare allelic variant that causes Charcot-Marie-Tooth syndrome in the family of J. R. Lupski.
43. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328,** 636–639 (2010).
44. Jones, S. J. *et al.* Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome Biol.* **11,** R82 (2010).
45. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature Genet.* **42,** 30–35 (2010).
46. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genet.* **42,** 790–793 (2010).
    One of the first demonstrations of using exome sequencing to identify a major causative mutation in Kabuki syndrome, using genomic DNA from a small number of unrelated affected individuals.
47. Gilissen, C. *et al.* Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. *Am. J. Hum. Genet.* **87,** 418–423 (2010).
48. Chin, C. S. *et al.* The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* **364,** 33–42 (2011).

Acknowledgements I thank R. Wilson, D. Dooling and G. Weinstock for critical reading of the manuscript. J. McMichael was integral to the creation of Fig. 1.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to the author (emardis@wustl.edu).