Full text provided by www.sciencedirect.com
ScienceDirect

# Whole-genome re-sequencing

David R Bentley

DNA sequencing can be used to gain important information on genes, genetic variation and gene function for biological and medical studies. The growing collection of publicly available reference genome sequences will underpin a new era of whole genome re-sequencing, but sequencing costs need to fall and throughput needs to rise by several orders of magnitude. Novel technologies are being developed to meet this need by generating massive amounts of sequence that can be aligned to the reference sequence. The challenge is to maintain the high standards of accuracy and completeness that are hallmarks of the previous genome projects. One or more new sequencing technologies are expected to become the mainstay of future research, and to make DNA sequencing centre stage as a routine tool in genetic research in the coming years.

**Addresses**
Solexa Ltd, Chesterford Research Park, Little Chesterford, Near Saffron Walden, Essex, CB10 1XL, UK

Corresponding author: Bentley, David R (david.bentley@solexa.co.uk)

**Current Opinion in Genetics & Development** 2006, **16**:545–552

This review comes from a themed issue on
Genomes and evolution
Edited by Chris Tyler-Smith and Molly Preworski

Available online 18th October 2006

0959-437X/$ – see front matter
© 2006 Elsevier Ltd. All rights reserved.

**DOI** 10.1016/j.gde.2006.10.009

## Introduction

Much of our information about genome structure, function and evolution depends on sequence data. The availability of an increasing range of high-quality reference genome sequences for different species [1–4] provides a new opportunity to study genetic variation on an unprecedented scale. Whole genomes, regions, or genes can be re-sequenced in multiple individuals. The sequence data from each individual are aligned to the appropriate reference, and the genetic variants between the different samples can be detected as high-confidence sequence differences [5]. Similarly, sequence data from RNA samples can be used to discover new RNA species, to measure levels of gene expression and, thus, to study the transcriptional state of different cells or tissues [6]. This review explores the prospects for whole-genome re-sequencing using new technologies, with a focus on human genetic and medical applications.

Cost and throughput are the key limitations of sequencing. The human genome reference sequence cost about $1 billion to produce; it is 99.995% accurate and near-complete, containing >99% of the euchromatic region [4]. The remaining 1% consists of ~300 small gaps, many of them adjacent to recently duplicated sequence, and these may contain DNA that cannot be propagated in bacteria before sequencing. In addition to the euchromatin, there are heterochromatic regions totalling an estimated 200 Mb. These are highly repetitive and have not been sequenced. They comprise the centromeres and some additional blocks on specific chromosomes — the main ones being the short arms of chromosomes 13, 14, 15, 21 and 22, and part of the long arms of 1, 9 and Y.

An individual human genome sequenced today using the Sanger method and capillary electrophoresis would take approximately 10 000 instrument days (e.g. 30 instruments for 1 year) to complete and would cost approximately $10 million. Further cost reductions of between 100- and 10 000-fold are required, with increases in throughput of similar magnitudes, without compromising the accuracy or completeness of the data. For $100 000 each, for example, we could sequence 24–48 human genomes and obtain a baseline measure of human genetic variation at a defined population allele frequency. The resulting dataset would be an unbiased resource for human genetic studies and could be deepened when appropriate. At the same time, a bacterial genome could be sequenced for $400, similar to the cost of some routine clinical pathology tests. Targetted sequencing of a 1 Mb region of the human genome in multiple individuals (e.g. to characterize regions associated with disease) might be done for a similar cost. Much larger datasets will be generated if costs continue to fall towards $10 000 per human genome. Examples would include sequencing large collections of well-defined tumour genomes in comprehensive searches for somatic variation associated with tumorigenesis. Similarly, extensive re-sequencing of case collections in common diseases such as diabetes, obesity or cardiovascular disease would yield catalogues of germline variation to aid searches for novel risk factors. Sequencing samples from different population subgroups of humans (and other species) would reveal patterns of natural selection and help to identify regions of the genome that have evolved during adaptation to environmental changes such as diet, climate or exposure to infectious pathogens. For $1000 per human genome, the concept of personal genome sequencing would become a technical reality, and bacterial sequencing for $4 would be one of the cheapest laboratory tests available.


Petitioner Illumina, Inc., EX1017, p. 545

## New sequencing technologies

Many new sequencing methods are being explored at present (Figure 1). Some use new approaches, and the data generated from them will be in a new and rather different format from the long reads produced by Sanger sequencing (Figure 1a) [7]. Each method is at a different stage of development, and the range varies from early pilot results to full commercialisation. The following brief summary is not exhaustive, but it illustrates the diversity and some key properties of the methods that are relevant to sequencing performance. For more information see, for example, the review by Shendure *et al.* [8] and references therein.

### Microelectrophoresis

A 384-capillary instrument can be miniaturised using microfabrication techniques and occupies the area of a compact disk (approximately 80 $cm^2$) [9]. Miniaturised sequencing reactions are integrated with separation technology on the same solid-state device. The chief advantage of this device over current capillary systems would be the major reduction in reagent volumes (i.e. more than 1000-fold) and hence consumable costs. Another potential benefit of this system is the use of Sanger sequencing to generate long reads. However, this system is very conservative with respect to parallelisation.

### Sequencing by hybridisation

An ordered array of oligonucleotides is fabricated to query the identity of a target sequence by hybridisation (Figure 1b) [10]. The array requires a minimum of four oligonucleotides to test each base in the target sequence — one to test for each of the four alternatives, A, C, G or T, at a particular position. Each oligonucleotide (25-mer) might occupy an area of 1 $\mu m^2$, although the current arrays have larger features than this. The accuracy of base-calling relies on the ability of the method to discriminate between exact matches to those with single base differences in the 25 bp hybridisation. Some oligonucleotide sets will work better than others because the hybridisation characteristics of each set will vary depending on base composition, although there are experimental conditions that have been used to reduce this effect. Most repetitive sequences are excluded from this type of analysis, but the method has been successfully applied to re-sequence regions of bacterial genomes and human chromosomes [11] that have unique (i.e. single-copy) sequence. This method is most widely used for genotyping, in which each SNP assay can be tested for performance. A suitable subset of assays can then be selected for use in data production on the basis of the performance of each assay and the specific requirements of the study [12].

### Sequencing by synthesis on arrays

In these systems, single DNA molecules are clonally amplified in spatially separate locations in a highly parallel array and used as templates for sequencing by synthesis. The use of clonal single-molecule templates is very useful for studying heterogeneous samples; for example, either for detecting heterozygous SNP (single nucleotide polymorphism) alleles in diploid DNA or, alternatively, for identifying somatic mutations in a tumour. Because each sequencing template is clonal in origin, high-quality base-calls are generated separately for each variant. These systems also avoid bacterial cloning steps and can potentially generate sequence data from DNA that cannot be propagated in bacteria. The three systems outlined below have the potential for parallelisation by several orders of magnitude. Scalability may vary between systems, depending on the nature and sophistication of the array surface. The three systems also differ in the type of sequencing chemistry they use.

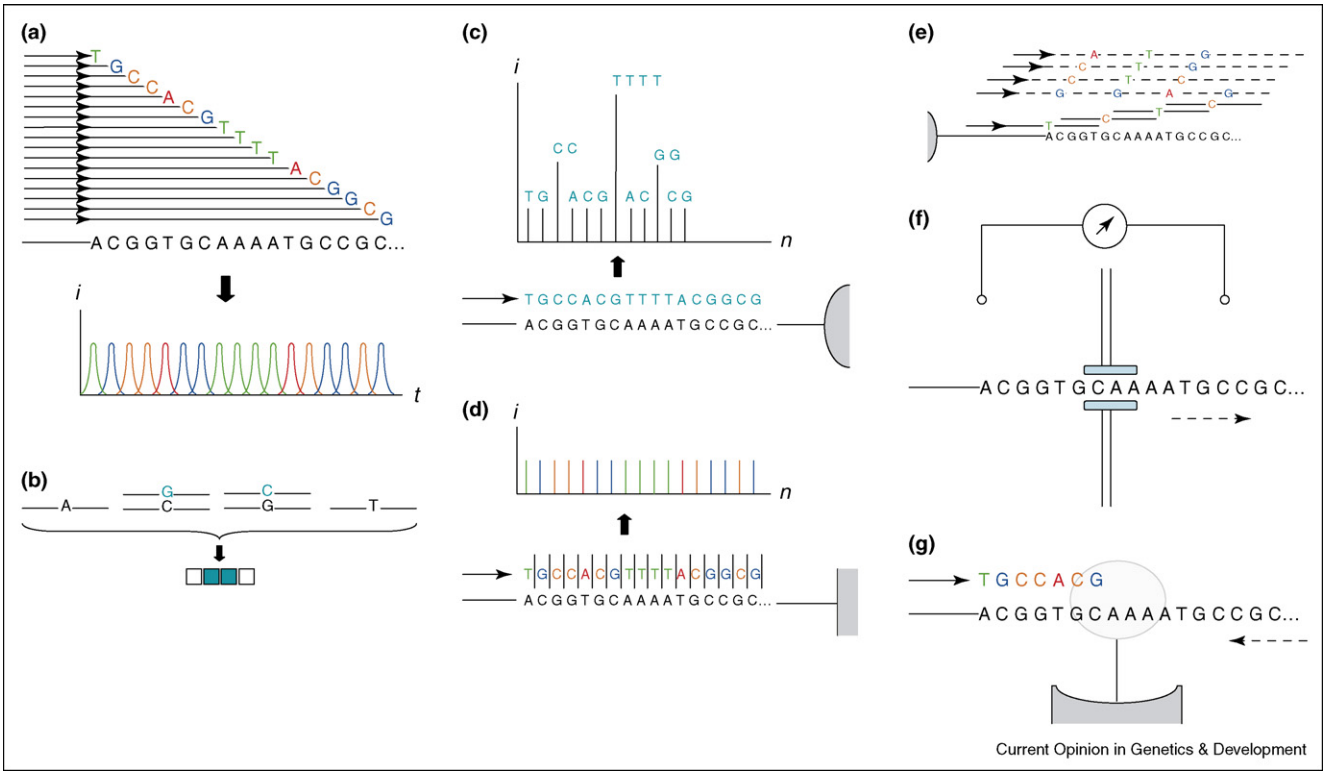
The 454 system (www.454.com) [13] uses templates that are immobilised and amplified on beads in aqueous-oil emulsion. Beads with DNA are purified and placed in individual microfabricated picolitre wells (50 $\mu m$ spacing; not all wells are occupied) for pyrosequencing [14]. One deoxynucleoside triphosphate (dNTP) plus DNA polymerase is added per cycle, and the 3′ end of the nascent chain is extended from a primer until a position is reached where a different base is required. A chemiluminescent signal is released in solution in the well and is quantitated to determine the number of bases incorporated. The process is then reiterated with each dNTP (A, C, G or T) in turn. Typically, 80 cycles of incorporation yield reads of approximately 100 bases, making it possible to either align the reads to a reference for re-sequencing or use them in *de novo* assemblies. Up to 40 Mb of raw data are produced per experiment — and recent announcements indicate that we may expect this figure to increase two- or threefold. Quantitation of the signal to estimate the length of a monomeric tract does not necessarily compromise sequencing accuracy in the first few positions of a tract but becomes progressively more of a problem in longer tracts. Quality values can be derived individually wherever the detected base is flanked by two different bases in the sequence, because the measured signal of the detected base is derived only from the raw data derived from sequencing that base. However, for runs of more than one base, the base-call and quality values for the second and subsequent bases are all derived from one signal (Figure 1c).

The Solexa system (www.solexa.com) provides several contrasts to the above system. Single molecules are covalently attached to a planar surface and amplified *in situ*. Sequencing by synthesis is carried out by adding a mixture of four fluorescently labelled reversible chain terminators and DNA polymerase to the template. This results in addition of a single reversible terminator to each template (Figure 1d). The fluorescent signal is detected for each template, and the flourophore and the reversible block are removed. The terminator–enzyme mix is then



Petitioner Illumina, Inc., EX1017, p. 546

**Figure 1**




Sequencing methodologies. **(a)** Synthetic chain-terminator chemistry. A universal primer (arrow) directs synthesis of DNA products, using a template (black letters) incorporating dNTP, and each chain terminates when a chain terminator (dideoxynucleoside triphosphate) with a base-specific fluorescent dye attached to it is incorporated (coloured letters). The mixture of fluorescently labelled fragments is resolved electrophoretically (one sample per lane or channel) on the basis of length, and the sequence is decoded from the order of tags passing the detector. The detected signal can be viewed as a trace that displays intensity *i* with respect to time of detection *t*. **(b)** Sequencing by hybridisation. Four oligonucleotides, each identical except for the central base, are immobilised on arrays, each in its own location. A DNA sample is labelled with a fluorescent dye and hybridised to the array. In the example shown, a heterozygous position (G/C) provides fragments that hybridize to two features, each representing one allele. Intensity is detected and quantified from each feature (depicted as squares). **(c)** Pyrosequencing. Template (black letters) is attached to a bead surface (shaded), and a primer directs DNA synthesis, which involves addition of one deoxynucleoside triphosphate per sequencing cycle. One signal is generated pre cycle by assaying the amount of pyrophosphate (PPi) released using a two-step reaction: PPi is used to form ATP catalysed by ATP sulfurylase, and the ATP is used for production of light from luciferin oxidation catalysed by luciferase. The products of the reaction are released in solution and are contained within small wells to enable mapping of the signal to each template. The number of bases present in a homopolymeric run is decoded by quantifying the amount of PPi released in a reaction (in the figure, for example, the single peak representing a run of four Ts is four times the height of a peak representing a single base.) Intensities (*i*) are plotted with respect to the order of successful signal detections (*n*) to create a 'flowgram'. **(e)** Base-by-base sequencing by synthesis. Template (black letters) is attached to a flat surface (shaded), and polymerase and a mix of four base-specific fluorescently labelled reversible chain terminators is added. A primer directs incorporation of one base, which is decoded following signal detection. Extension is blocked (depicted by the vertical line after each base) until the block is removed. The process is iterated, enabling addition and detection of one base on each template per cycle. Signal intensity is plotted with respect to cycle number, generating a trace. Not shown: signals for all four fluorescent channels are collected and plotted at each position, enabling quality per scores to be derived using four-colour information if desired. **(f)** Sequencing by ligation. An anchor primer (arrow) is annealed to a template attached to a bead. A complex mixture of oligonucleotides (representing all possible 9-mers) is added, each labelled with one of four fluorophores that is specific for the identity of the central base. Incorporation of one oligonucleotide is directed by the template, and the central base is decoded (in this case T) following signal detection. In one iteration of this method, part of the incorporated oligonucleotide is removed before each subsequent cycle of ligation. Three more cycles are illustrated (resulting in decoding of C, T and C, respectively, at every fifth position). The entire process is repeated with four additional anchor primers, each shifted one base from the previous primer, to enable decoding of the intervening bases in four additional rounds. Sequencing by ligation can be carried out in either direction from the primer (i.e. an oligonucleotide can be ligated to either the 5′ or the 3′ end of the anchor primer). **(g)** Nanopore technology. A DNA molecule is passed through a 1.5 nm wide pore (e.g. haemolysin in a membrane) in an electric field (the dotted arrow denotes direction of motion). Changes in conductance through the pore are influenced by the identity of the base at the pore. **(h)** Single-molecule sequencing by synthesis in real time. A polymerase is immobilised on a solid surface. Template and annealed primer are added, and incorporation of dNTPs takes place at the rate catalysed by the anchored polymerase in an uninterrupted reaction. Various assays for the identity of each incorporated nucleotide are under investigation. Two examples are the use of a mixture of four-colour base-specific labelled dNTPs, and a fluorescence resonance energy transfer (FRET) system that uses a donor attached to the polymerase and acceptors attached to the gamma-phosphate of each dNTP (see text and references for details).



Petitioner Illumina, Inc., EX1017, p. 547

added to start the next cycle, and the process is reiterated until the end of the run. Given that all four are present in the reaction, the risk of misincorporation is minimised, increasing sequencing accuracy. Accuracy is also independent of sequence context, and a discrete signal is generated for every base. Read lengths are 30–50 bases, which are of sufficient length for re-sequencing applications (see below).

Shendure *et al.* [15] attach template DNA molecules to 1 μm beads that are embedded in polyacrylamide for amplification, forming 'polonies' (polymerase colonies). Sequencing is performed using multiple cycles of ligation of labelled 9-mers from a start point determined by an anchor primer. A mixture of 9-mers is added to the reaction, tagged using a four-colour fluorescent labelling system that distinguishes the central base (nnnnAnnnn, nnnnCnnnn, nnnnGnnnn and nnnnTnnnn). The signal provided by the successfully ligated 9-mer — this should be an exact match to the template — identifies the central base. Multiple cycles of ligation can be carried out to decode every fifth base, and the process can be repeated with different anchor primers to sequence the intervening bases (Figure 1e). In the original report, data from1.6 million templates (filtered from 14 million) each provided 26 bases of information and were used to re-sequence the genome of a strain of *Escherichia coli*. Although there are variants to this format, as adopted by Applied Biosystems (www.appliedbiosystems.com), this example illustrates the possibility of sequencing by decoding discontinuous bases. The method also illustrates the use of a stringent filter to retain high quality data at the expense of yield.

### Single-molecule sequencing

Several groups have investigated the possibility of sequencing directly from single DNA molecules. Potential benefits include using very small amounts of sample, avoiding template amplification, generating a homogeneous signal, and the possibility of obtaining a very high data-density. These are offset by the requirement for a detection system of very high sensitivity. There are also risks of losing data because each sequencing read is susceptible to transient or permanent loss of signal by a single event such as DNA damage or a signal detection failure.

Solexa initially used four colour-coded reversible terminators, added simultaneously in the reaction, to sequence single-molecule templates by synthesis [16]. Quake *et al.* [17] reported adding single, fluorescently labelled nucleotides sequentially in a proof-of-principle study. Helicos (www.helicosbio.com) have adopted the latter approach. If this method succeeds in a production environment, we can envisage that the type of data produced will resemble the output of the multimolecular sequencing by synthesis on arrays described above (Figure 1c or 1d, respectively, depending on sequential or simultaneous addition of the four dNTPs). There would be potential for massively parallel production of short reads. Possible sources of error include those associated with single- or four-colour nucleotide systems described above. In addition, misincorporation by the polymerase will yield a high-confidence wrong base-call when a single molecule template is used, unlike the equivalent system using multimolecular templates. As long as these errors are randomly distributed they will not affect consensus base-calls.

Alternative single-molecule formats under investigation include the use of nanopores, whereby a DNA molecule is driven by an electric field through a 1.5 nm pore at a rate of at least ~1000 bases per second (Figure 1f) [18]. Conductance through the pore can be measured and provides a signal that is potentially characteristic of the identity of the base obstructing the pore. Other proposed arrangements use polymerase-based sequencing in real-time, detecting base-specific incorporation as the polymerase moves along a single DNA molecule at a rate of ~100 bases per second (www.pacificbiosciences.com; www.visigenbio.com) [19]. If they are reduced to practice, these methods could potentially produce long reads very fast. Reagent costs would also be very low, because the entire experiment would be carried out in a single, small reaction volume. Extensive parallelisation might be a major challenge but is much less crucial given the theoretical rate of detection. Given the early state of these technologies, it is not possible to predict the scope of these methods.

## Cost, throughput, accuracy and completeness

Cost, throughput, accuracy and completeness are interrelated, and efforts to decrease project cost or to increase throughput have sometimes been accompanied by a reduction in accuracy or completeness. The challenge for new technologies is to achieve massive improvement in one or more of these components without compromising the others.

### Cost

To achieve dramatic improvements in cost and throughput compared with those of current macrocapillary systems, a radical change in sequencing methodology is needed. The single most effective parameter that changes cost is massive parallelisation. In contrast to capillary systems (96 or 384 channel), sequencing by synthesis on arrays has already achieved a parallelisation of $10^5$–$10^7$ reactions — the exact degree of improvement depends on the system — and these methods have the potential to support in excess of $10^8$ reactions per experiment (albeit with shorter read lengths and hence fewer bases of sequence per reaction). Miniaturisation to achieve high data-density is an important contributor to cost reduction — assuming comparable rates of throughput between systems. Comparing estimates of data density (measured in bases/10 $\mu m^2$) of the different



Petitioner Illumina, Inc., EX1017, p. 548

platforms, microelectrophoresis has a value of 0.0034; sequencing by hybridisation (assuming 1 μm features) has an estimated value of 2.5. The current three systems for sequencing on arrays achieve densities of between 10 and 500 (i.e. up to five orders of magnitude improvement over microelectrophoresis). The potential increase in data density has a positive impact on reagent cost. Reagent volumes are significantly reduced, whereas it may be assumed that concentrations of reagents in polymerase reactions are generally comparable, with less than tenfold variance between platforms. Instrument costs also remain relatively comparable.

### Throughput

The intrinsic throughput of a sequencing system depends on speed of detection and degree of parallelisation. A capillary sequence reads 0.17 bases per second per channel. Allowing for 96 channels, the total throughput in continuous operation is 17 bases per second, limited only by the rate of electrophoresis. Sequencing by synthesis on arrays has a much slower cycle time because reaction chemistry is carried out in-between read-outs at each cycle. This time penalty is more than compensated for by the degree of parallelisation, and the total throughput in continuous operation of current systems is between 1400 and 4000 bases per second. The theoretical potential of the nanopore technologies is evident in this comparison: if an instrument can sequence single molecules during polymerase synthesis in real time, a single-channel instrument would have a total throughput of 100–1000 bases per second (see above). Modest parallelisation could make this system very competitive with regard to throughput. For example, 10 channels each delivering 1000 bases per second would have a throughput comparable with array-based sequencing by synthesis, and such a system would be competitive if other factors — in particular accuracy — were equal.

It is important to note that sequencing instruments rarely work in continuous operation. To consider throughput in sequence production, it is necessary to take into account time between runs, time for servicing, staff availability and failure rates. Continuous operation is less compatible with normal working hours if run times are short (a few hours or less), because the instruments might not be manned overnight or at weekends. A feature of the latest 96-capillary instrument, the AB3730 (Applied Biosystems; www.appliedbiosystems.com), was to overcome this problem by providing the necessary automation for the instrument to run unattended for multiple runs.

### Accuracy

Capillary sequencing using Sanger chemistry generates high-quality raw sequence data over most of the length of each read. Accuracy is measured using Phred [20], an algorithm that provides a numerical score (i.e. quality value) that is an estimate of the probability of error for each base-call in a raw read [21]. The algorithm takes into account various parameters extracted from the raw data. The quality values can be submitted to the public databases along with the base-calls, thus providing a convenient measure of data quality. For many projects, the traces are also available in the trace repository (http://trace.ensembl.org; www.ncbi.nlm.nih.gov/Traces). Making the quality values available enables users to choose a threshold for filtering the data depending on the application. For example, multiple reads from the same sample can be aligned and used to derive a consensus base-call at each position. In this case the depth of the consensus provides high confidence base-calling, and the quality of the raw data may be considered sufficient at Q15 [4]. Aligned reads, and consensus and individual base-quality values can also be used for calling SNP alleles. By contrast, if single reads are used to call a SNP with high confidence, a much higher threshold is applied in selecting the raw data that is used to support the base-call [5].

New technologies work on a range of different principles that will require different ways to define good quality metrics. Success in this area is important to make reliable comparisons of the accuracy of different systems. At present, however, there are insufficient data available from most technologies. Base-by-base sequencing, as used in the Solexa system, provides information on the raw data collected at each base, enabling derivation of quality values that are analogous to Phred scores. In a similar manner to capillary data, the quality value is relatively even along most of the run. Consensus scoring and quality values can be used to call SNPs in the same way as that developed for capillary sequence data (illustrated schematically in Figure 2). Pyrosequencing data can also be used to derive quality metrics, but dividing up a single signal between two or more bases is not directly analogous to a strict 'per base' quality metric. For all methods, it is necessary to calibrate quality values empirically, by generating large amounts of sequence data from a known template, measuring the frequency of incorrect base-calls and assessing the validity of various quantitative parameters in the raw data in correlation with observed error rates. More work is required in this area to assess the accuracy of each sequencing method, to look for loss of accuracy in particular sequence contexts for each method, and to enable assembly of sequence data obtained using more than one technology.

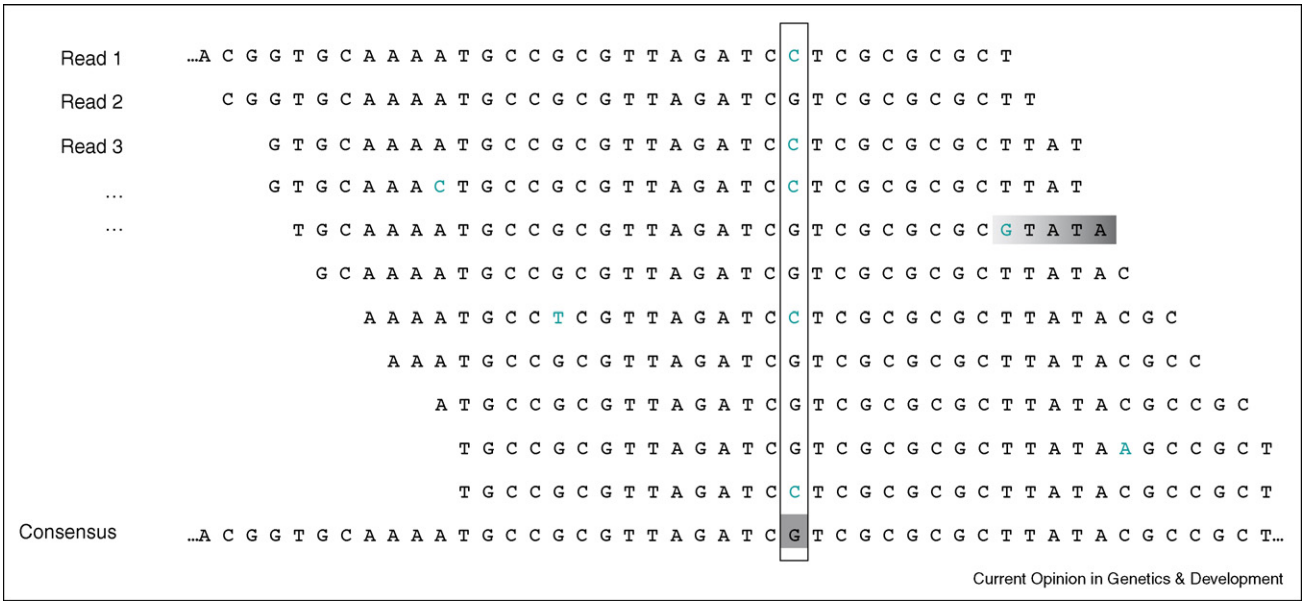
### Completeness

Long reads of high accuracy provide a very high level of completeness in most sequencing projects. Regions that may confound long read-sequencing include extremes of base context such as long polyA stretches or hairpin structures in templates, wherein progress of the DNA polymerase along the template is compromised. Assembly or alignment of long reads is impossible when there



Petitioner Illumina, Inc., EX1017, p. 549

**Figure 2**




Viewing sequence alignments. A schematic diagram of the viewing alignments of 35-base reads derived from a diploid DNA sample, and the resulting consensus, in the style of the gap4 program for sequence alignment and viewing [23] (http://staden.sourceforge.net). The gap4 viewer contains extensive editing and viewing features, including the ability to view quality scores, traces and sequence differences. The schematic shows errors that are distributed randomly (one each of C, T, G and A in blue) and excluded from being true base-calls because they are isolated examples of a base call, different from all the other calls at that position in other reads that contribute to the consensus. In one position (boxed), there are multiple incidences of two base-calls. The consensus quality value is low (grey shading) because of high representation of two base-calls, and this indicates the presence of a heterozygous position, with both base-calls being true. In addition to the use of majority voting and the consensus quality value to make the call, and depending on the sequencing chemistry used, there may be a quality value for each base of each read (see text).

are identical long repeats (i.e. when the length of the repeat is longer than the read-length). A few sequences are refractory to cloning in bacteria, and templates cannot be prepared for them. The completeness of the reference human genome [5] (see above) demonstrates the utility of long reads generated by Sanger sequencing. Long reads generated by other methods would be expected to have the same utility if the accuracy was equivalent and if there was no systematic bias towards failure in certain sequence contexts. However, it is not clear how particular sequence motifs might affect migration of DNA through a nanopore, or how evenly a polymerase migrates along a single template in real-time and whether this would affect the accuracy of sequence detection.

The current array-based sequencing technologies are accompanied by a substantial reduction in read-length to 25–100 bases. However, this is an acceptable trade-off for many applications, particularly re-sequencing. The availability of a reference genome sequence renders short reads very powerful as a means to obtain re-sequencing data. The only requirement of a short read is that it must be long enough, and sufficiently accurate, to align uniquely to the correct position in the reference. From simulations [22], reads of length 25–30 bases can be aligned uniquely so as to cover 80% of the human genome sequence. For 1 Mb human regions or 4 Mb bacterial genomes, reads can be aligned uniquely to 95–99% of all positions in the sequence.

Reads of 100 bases or more are currently used for *de novo* assembly of genomes. 25–30 base reads can also potentially be used for *de novo* assembly. In particular, the ability to generate paired 25–30 base reads separated by a known distance — or, preferably, multiple, paired-read datasets with a range of defined separations (e.g. 1, 2, 5, 10 and 40 kb) — would also aid *de novo* assembly of genomes and characterisation of copy number variations that are different from the reference sequence. More experience is required to explore this area, but it is likely that the use of short reads will not greatly compromise the degree of completeness that can be obtained. Instead, multiple new strategies will be developed for recovering specific fractions of sequence (e.g. those generated by long PCR or hybrid selection) that might be missed in a first round of sequencing. The concept of a random shotgun phase — low cost, high throughput and highly automated — followed by a directed finishing phase — costly, directed, using custom reagents and less automated — proved to be very successful in producing 'finished' sequence of much



Petitioner Illumina, Inc., EX1017, p. 550

better quality and utility than 'unfinished' of 'draft' sequence [4].

## Conclusions and prospects

Fifteen years ago, biology entered the genome sequencing era. Since then, a series of high quality, near-complete reference sequences have been produced for a wide range of organisms. Important lessons have been learned: making the data publicly available, stimulating rapid technology development, forming international consortia and establishing universal high standards have all contributed to the provision of long-term resources that will underpin genetics and medicine for years to come. These foundations are stimulating the study of genetic variation as a means to understand complex disease, pathogenicity, evolution and individuality. Whole-genome re-sequencing is an essential research tool to characterize genetic variation in all these contexts but has been limited until now by cost and throughput of the current technologies. A growing recognition of the need for re-sequencing is driving the invention of new sequencing methods. Those that succeed will bring improvements in productivity by several orders of magnitude, enabling us to decode the genetic make-up of a bacterium, a fungus or a human quickly and accurately, making knowledge of multiple individual genome sequences a key component to aid research, clinical decisions and lifestyles.

100-fold improvements in cost and throughput are anticipated in the coming year. It is clear that DNA sequence data will be produced in new formats, especially as short reads in the near future. It is becoming widely recognized that long reads are not required for re-sequencing, because the reference sequences provide an essential backbone against which even short reads of 25 bases can be aligned uniquely to more than 80% of any human genome. Targetted sequencing (e.g. using long PCR) can be used to obtain most or all of the remaining 20%. The same approaches may be used to maximise the yield of current and future methods for other genomes, and it has already been demonstrated that small genomes such as those of bacteria can be sequenced in a day.

Accuracy and completeness are hallmarks of genome sequencing, and the lessons learned over the past 15 years should be applied rigorously in the evaluation of new sequencing methods. Standards must be maintained, because accurate information underpins excellent science and ultimately saves money. New technologies need to be measured and compared on the basis of the principles and metrics reviewed in this article. Given the current rate of progress, we can confidently expect many individual genome sequences of humans and other species to become commonplace in the next few years. Our next challenge will be to analyse all the data, and to use it responsibly.

## Acknowledgements

The author gratefully acknowledges the assistance of Shankar Balasubramanian, Colin Barnes and Tony Smith for discussions and critical reading of the manuscript.

### Disclosure statement

The author is a full-time employee of Solexa, a company that is developing and marketing a new DNA sequencing technology.

## References

1. The Yeast Genome Directory: **The yeast genome directory**. *Nature* 1997, **387**(6632 Suppl):5.
2. C. *elegans* Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The *C. elegans* Sequencing Consortium**. *Science* 1998, **282**:2012-2018.
3. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE III *et al.*: **Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence**. *Nature* 1998, **393**:537-544.
4. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**:931-945.
5. The International SNP Map Working Group: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms**. *Nature* 2001, **409**:928-933.
6. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression**. *Science* 1995, **270**:484-487.
7. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proc Natl Acad Sci USA* 1977, **74**:5463-5467.
8. Shendure J, Mitra RD, Varma C, Church GM: **Advanced sequencing technologies: methods and goals**. *Nat Rev Genet* 2004, **5**:335-344.
9. Paegel BM, Blazej RG, Mathies RA: **Microfluidic devices for DNA sequencing: sample preparation and electrophoretic analysis**. *Curr Opin Biotechnol* 2003, **14**:42-50.
10. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP: **Accessing genetic information with high-density DNA arrays**. *Science* 1996, **274**:610-614.
11. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP *et al.*: **Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21**. *Science* 2001, **294**:1719-1723.
12. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR: **Whole-genome patterns of common DNA variation in three human populations**. *Science* 2005, **307**:1072-1079.
13. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**:376-380.
14. Hardenbol P, Baner J, Jain M, Nilsson M, Namsaraev EA, Karlin-Neumann GA, Fakhrai-Rad H, Ronaghi M, Willis TD, Landegren U *et al.*: **Multiplexed genotyping with sequence-tagged molecular inversion probes**. *Nat Biotechnol* 2003, **21**:673-678.
15. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome**. *Science* 2005, **309**:1728-1732.
16. Balasubramanian S, Bentley DR: **Polynucleotide arrays and their use in sequencing**. *Patent WO 01/157248* 2001.



Petitioner Illumina, Inc., EX1017, p. 551

17. Braslavsky I, Hebert B, Kartalov E, Quake SR: **Sequence information can be obtained from single DNA molecules**. *Proc Natl Acad Sci USA* 2003, **100**:3960-3964.

18. Winters-Hilt S, Vercoutere W, DeGuzman VS, Deamer D, Akeson M, Haussler D: **Highly accurate classification of Watson-Crick basepairs on termini of single DNA molecules**. *Biophys J* 2003, **84**:967-976.

19. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW: **Zero-mode waveguides for single-molecule analysis at high concentrations**. *Science* 2003, **299**:682-686.

20. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using Phred. I. Accuracy assessment**. *Genome Res* 1998, **8**:175-185.

21. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Res* 1998, **8**:186-194.

22. Whiteford N, Haslam N, Weber G, Prugel-Bennett A, Essex JW, Roach PL, Bradley M, Neylon C: **An analysis of the feasibility of short read sequencing**. *Nucleic Acids Res* 2005, **33**:e171.

23. Bonfield JK, Smith K, Staden R: **A new DNA sequence assembly program**. *Nucleic Acids Res* 1995, **23**:4992-4999.




Petitioner Illumina, Inc., EX1017, p. 552