Nucleic Acids Research, 2005, Vol. 33, No. 19 ***e171***
doi:10.1093/nar/gni170

An analysis of the feasibility of short read sequencing

Nava Whiteford, Niall Haslam, Gerald Weber, Adam Prügel-Bennett[1], Jonathan W. Essex, Peter L. Roach, Mark Bradley[2] and Cameron Neylon*

School of Chemistry and [1]School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK and [2]School of Chemistry, University of Edinburgh, Edinburgh EH9 3JJ, UK

Received June 2, 2005; Revised August 11, 2005; Accepted October 14, 2005

ABSTRACT

Several methods for ultra high-throughput DNA sequencing are currently under investigation. Many of these methods yield very short blocks of sequence information (reads). Here we report on an analysis showing the level of genome sequencing possible as a function of read length. It is shown that re-sequencing and *de novo* sequencing of the majority of a bacterial genome is possible with read lengths of 20–30 nt, and that reads of 50 nt can provide reconstructed contigs (a contiguous fragment of sequence data) of 1000 nt and greater that cover 80% of human chromosome 1.

INTRODUCTION

Several methods for ultra high-throughput DNA sequencing are currently under investigation (1–5). Many of these methods yield very short blocks of sequence information (reads), with some proposed methods giving reads as small as 16 nt. To be useful, these reads must be long enough to provide sufficient information to unambiguously place them on a known template sequence, or to generate unambiguous overlaps to reconstruct the sequence if no template is available. In both these processes repeats cause significant problems. Here we report on an analysis showing the level of genome sequencing possible as a function of read length. We show that re-sequencing and *de novo* sequencing of the majority of a bacterial genome is possible with read lengths of 20–30 nt, and that reads of 50 nt can provide reconstructed contigs of 1000 nt and greater that cover 80% of human chromosome 1.

The leading methods for ultra high-throughput DNA sequencing fall into two main categories; sequencing by hybridization (6) and sequencing by synthesis (5,7–11). Sequencing by hybridization is an extension of well established DNA microarray techniques that essentially aim to identify all subsequences of a specific length within a genome via their hybridization to presynthesized probes. Sequencing by synthesis utilizes a process by which nucleotides are added in a controlled fashion to isolate DNA templates. Each nucleotide is read in turn; each base being added and then read in a cyclic process. Both of these approaches produce sequence information in very short segments compared to conventional Sanger sequencing. For sequencing by hybridization the length of each fragment of sequence information (the 'read length') is limited by the length of the probe. Probes cannot be extended much beyond 30 nt as the selectivity for perfect matches over single mismatches drops to unacceptable levels (12). In the case of sequencing by synthesis obtainable read length is proportional to cycle efficiency. However the technical challenges involved in this approach mean that the length of useful sequence that can be obtained is limited. Solexa Ltd, have recently claimed reads of 25 nt with sufficient throughput to re-sequence the viral genome of ϕ X174 (http://www.solexa.com/news/2005/100305.htm). Another promising technique is the high throughput pyrosequencing approach of 454 Life Sciences who are currently reporting read lengths of $\sim$100 nt (13).

A key problem for these sequencing methodologies is that as the length of each individual read decreases, the probability that a read will occur more than once in the sequence increases. The problems that repetitions can cause for sequencing projects based on whole genome shotgun approaches, where the read length is as short as 500 nt, have recently been analysed in detail (14). There has been much debate concerning the minimum length of read required to generate useful sequence information (3). However, despite the importance of this analysis for the utility of many proposed ultra high-throughput sequencing methods, little work has been reported on the analysis or reassembly of sequence information from very short reads. Perhaps the most significant contribution to this area has been that of Chaisson *et al*. (15) who discusses the limitations of short read sequencing with read lengths starting at 70 nt, the largest genome analysed (*Neisseria meningitidis*) is $\sim$2 Mb. In contrast to this, here we describe an analysis for read lengths between 18 and 200 nt, and extend our analysis to the whole human genome. Also rather than showing the result

*To whom correspondence should be addressed. Tel: +44 23 8059 4164; Fax: +44 23 8059 6805; Email: D.C.Neylon@soton.ac.uk

of a particular reassembly tool, our analysis describes the absolute limits of sequence data that can be reassembled.

In a sequencing project based on short reads, repeated sections of the genome will cause several types of problems. If re-sequencing uses a known template sequence, repetitions will prevent an unambiguous assignment of reads to a single position on the template (16). Therefore the uniqueness of reads within the genome will be a key factor in determining how successful such re-sequencing can be. However, if the aim is *de novo* sequencing, a different problem is encountered. In this case repetitions will cause significant problems for assembly of contigs and will severely limit the amount of the sequence that can be effectively reconstructed.

MATERIALS AND METHODS

The analysis shown in this paper was generated using the RepAnalyse (17) sequence analysis tool which we have developed for this purpose. RepAnalyse uses linear time suffix (18) and LCP (19) array construction algorithms. The uniqueness analysis and simulated reassembly are performed by analysing this suffix array. A technical report (17) for RepAnalyse is available on our website (www.4g.soton.ac.uk) or from the corresponding author. Details of the analysis methods and other applications will be reported elsewhere. Using the algorithms developed for RepAnalyse to process the suffix array we are able to produce the analysis shown far faster than with traditional techniques. Processing the whole human genome uniqueness analysis, required the construction of a full human genome suffix array for both forward and reverse strands, a total of 6.1×10^9 symbols. The analysis took two days and 70 Gb of RAM on a 1.6 GHz Itanium2 Processor.

RESULTS

Uniqueness

We first focused on the simple issue of what percentage of all sequences are unique within a series of model genomes. This uniqueness measure is directly linked to the effect of read length on re-sequencing, where sequence information is assembled on a known template. In its simplest form, re-sequencing requires that a probe is uniquely identifiable with a single location on the template sequence.

To evaluate the feasibility of sequencing using a template we calculated the percentage of unique reads within a given genome using suffix array (20) derived analysis (17). We consider all possible reads of length l in the genome. Each read is compared with every other read of the same length and all reads that occur only once are counted. The uniqueness for read length l is then given by $U_l = n_{u,l}/n_{p,l}$, where $n_{u,l}$ and $n_{p,l}$ are the number of unique and possible reads, respectively. We have analysed the uniqueness of reads of all lengths in several model genomes. In Figure 1 we show two representative examples, a viral genome; λ-phage (GenBank accession no. NC_001416 size: 48 kb) and a bacterial genome, *Escherichia coli* (K12 MG1655, accession no. NC_000913 size: 4.6 Mb). In both cases uniqueness has an approximately sigmoidal dependence on read length. This closely follows the behaviour of randomly generated sequences of the same size. The position of the rise in the *E.coli* genome is at higher read length, largely due to the increased sequence length and therefore higher probability that any given sub-sequence will repeat. The most pronounced difference between the data for *E.coli* and λ-phage is that uniqueness in the bacterial genome reaches a plateau at around 97% from where the value then rises only very slowly. For instance while 97% of 18 nt reads are unique, a read length of 475 nt is required for 99% of reads to be unique. This plateau is a result of repeated segments such as the seven copies of the genes that code for ribosomal proteins and RNA and frequently repeated intergenic sequences such as Intergenic Repeat Unit (IRUs) and Enterobacterial Repetitive Intergenic Consensus (ERICs) (21). Our analysis of the whole human genome shows a similar pattern, with a plateau being reached at a higher read length (Figure 2a).

Reassembly

Re-sequencing applications require that there is a known template sequence of sufficient quality to provide a map on

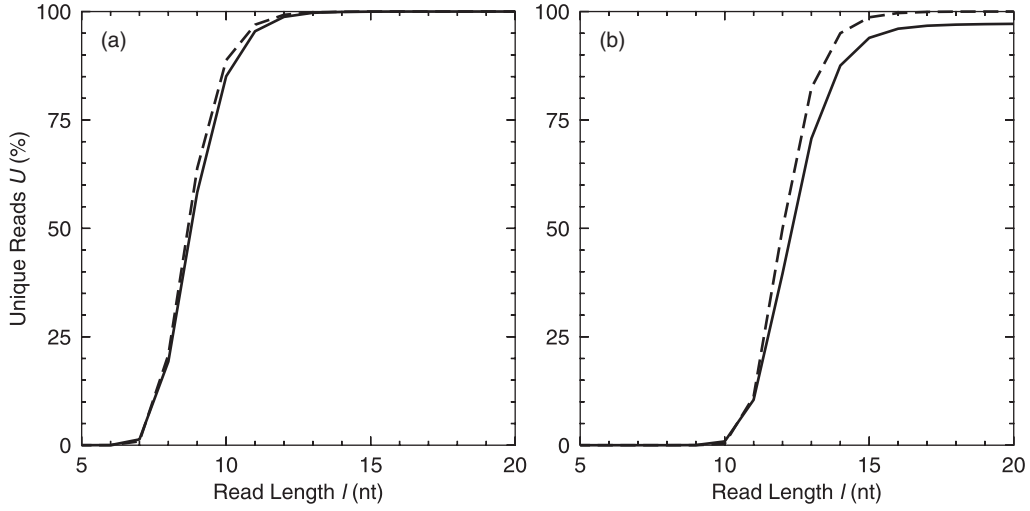

Figure 1. Percentage of unique reads as a function of read length for (**a**) λ-phage and (**b**) *E.coli* K12. The dashed curves show results for randomly generated sequences of the same size, which are content-biased to yield the same relative proportion of nucleotides given in ref. (27).

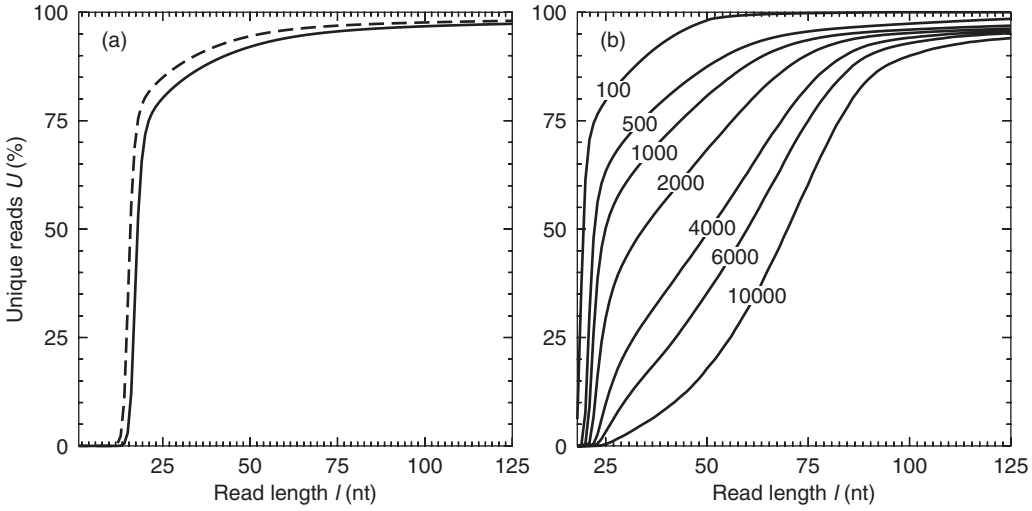

Figure 2. (**a**) Percentage of unique sub-sequences (*U*) for varying read length (*l*), the solid line shows uniqueness in the whole human genome, the dashed line shows uniqueness in human chromosome 1. (**b**) Percentage of human chromosome 1 covered by contigs greater than a threshold length as a function of read length. The horizontal axis starts at 18 nt, due to the limitations of reassembly below this length.

which reads can be placed. In the case of *de novo* sequencing, such as the sequencing of a new organism, this map is not available. In the extreme case where no map information is available it is necessary to assemble these reads into contigs without the aid of any external information.

As more complete genome sequences become available the quality of map information will increase. However it is not currently clear how reliable much of this information is, nor how best to apply it to the problem of reassembly from very short reads. We have therefore restricted our analysis to reassemble with no additional mapping information. In this analysis we use a suffix array to examine all possible fragments of a given size and use this information to predict which contigs could be assembled.

For two reads to be reconstructed into a single contig, an unambiguous overlap between them is needed, i.e. there must not be two or more potential reads that would extend a contig differently. It is obviously possible to reassemble sequences where all potential overlaps are unique. It is also possible to reconstruct contigs from sections of DNA that have two or more identical repeats elsewhere in the genome. However it is not possible to extend the contig beyond the point where the repeats diverge. For example if there are two identical copies of a gene in the genome which differ only in the downstream non-coding sequence then it is possible to build a contig that contains the gene but this will be broken at the point where the downstream sequences differ. The gene would be reconstructed, but its location would be ambiguous.

Figure 3 shows how the percentage of the *E.coli* genome covered by contigs greater than a given length is affected by read length and Figure 2b shows the analysis for human chromosome 1.

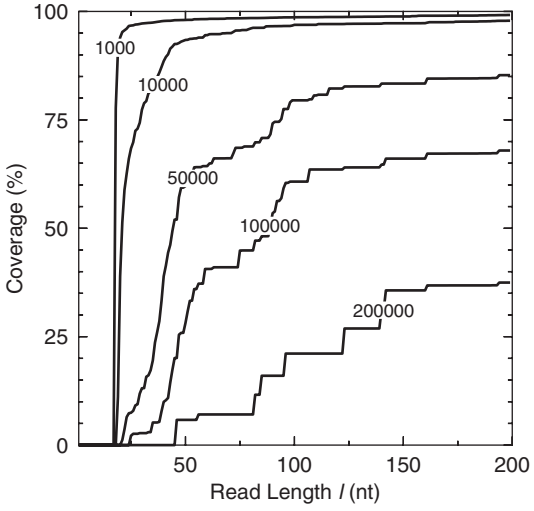

Figure 3. Percentage of the *E.coli* genome covered by contigs greater than a threshold length as a function of read length.

DISCUSSION

Viral sequencing

For the very short genome of λ-phage reads of 12 nt are 98% unique. Therefore re-sequencing of this genome should be straightforward given a complete and error free read set. Re-sequencing by hybridization using probes of around 25 nt is now a reasonably well established technology and has been reported for the SARS Coronavirus (22). Other viral genomes we have analysed show similar trends (data not shown). As the genome is very short and contains very few repeats it is also possible to reassemble the genome of λ-phage phage from short reads. Our analysis shows that it is possible to completely reassemble the genome from 18 nt reads and that 17 nt reads can cover 99.6% of the genome with contigs larger than 10 000 nt. Below this read length, coverage falls dramatically. This confirms and extends the result of Chaisson *et al*. (15) where it was found that a read length of 70 nt would allow a viral genome (*Adenovirus* 35 kb accession no. AF394196) to be completely reassembled. And indicates that *de novo* sequencing of genomes of viral size should be well within the limitations of short read methods. The results on *de novo*

sequencing show that, in principle at least, it is possible to sequence these genomes without any prior knowledge. However, highly repetitive elements can be found in specific viral genomes which will increase the read length required in these cases. It should also be noted that when sequencing a viral population it will not be possible to determine individual genome variability without first isolating that individual sequence.

Bacterial and small eukaryote sequencing

The uniqueness and reassembly analysis of the bacterial genome of *E.coli* K12 MG1655 show one main difference to that of both the viral genome and randomly generated sequences. While 97% of the genome can be unambiguously probed by reads of 17 nt (Figure 1b) the remaining 3% is ambiguous for much longer read lengths. This difference between the bacterial and viral genome sequences is a result of much more repetition of both large (kilobases) and small (20–200 nt) sequences in the *E.coli* sequence. The plateau in *E.coli* uniqueness suggests that, for bacterial genomes, once a threshold read length has been reached ($\approx$ 17 nt for *E.coli*) further increases in read length will not provide a significant increase in the quality of information. This indicates that for bacterial genomes, methodologies that can provide reads of 18 nt or more are adequate for re-sequencing. However, it is important to note that specific bacterial genomes may contain highly repetitive elements which will increase the required read length. Sequencing by hybridization and other high-throughput methods can provide this information so the expansion of existing methods for re-sequencing viral genomes will provide sufficient data for re-sequencing most of a given bacterial genome.

Analysis of the reassembly of the *E.coli* genome show that with a read length of 20 nt, over 98% of the genome can be re-assembled into contigs larger than 100 nt, and 91% into contigs larger than 1000 nt. However, only 10% is covered by contigs larger than 10 000 nt. At this read length over 75% of genes are entirely covered by a single contig. A read length of 30 nt improves the coverage for 10 000 nt contigs significantly to 75%, with over 96% of genes entirely covered by a single contig, this rises to 90 and 98%, respectively at a 50 nt read length. *De novo* sequencing of 90–97% of the *E.coli* genome into gene-sized contigs is therefore possible with reads of around 50 nt and a significant portion can be reconstructed with much shorter read lengths. This compares well with the result of Chaisson *et al*. (15) where it was found that a 70 nt read length would allow 93.4 and 99.5% of bacterial genomes (*Campylobacter jejuni*, 1.6 Mb and *N meningitidis*, 2.1 Mb accession nos: CJ11168X1 and NMA1Z2491) to be reconstructed into contigs of 1000 nt and greater.

Our analysis of the *Caennorhabitis elegans* genome (97 Mb, Figure 4b) showed that in a predicted reassembly from 50 nt reads 88% of the sequence is covered by contigs larger than 1000 nt and 51% is covered by contigs larger than 10 000 nt. By a read length of 100 nt a significant percentage (8.4%) of the genome is covered by contigs larger than 100 000 nt. This indicates that the whole genome shotgun sequencing of small eukaryotes is within the limitations of short read sequencing.

Human genome sequencing

The ultimate goal for all high-throughput sequencing technologies is to enable affordable human genome sequencing. The human genome is much larger and much more repetitive than that of *E.coli* and therefore the challenges for short read sequencing approaches are expected to be much greater.

This is immediately evident when analysing uniqueness as a function of read length for the whole human genome (Build 35.1 size: 3×10^9 nt). Our uniqueness analysis for the whole human genome is shown in Figure 2a. This result compares well with the estimation of Shendure *et al*. (1) where they state that reads of $\approx$ 60 nt will be required for 95% uniqueness.

While the overall shape of the curve is similar to that for *E.coli* the rate of increase after the plateau is reached is much slower. Read lengths of 25 nt would in principle be able to probe 80% of the genome, however read lengths of at least 43 nt are required to cover 90% of the genome.

Such a length is beyond the reach of sequencing by hybridization as selectivity for single mismatches will not be possible (12). Single molecule (23) and sequencing by synthesis approaches currently claim read lengths of 25 to 115 nt

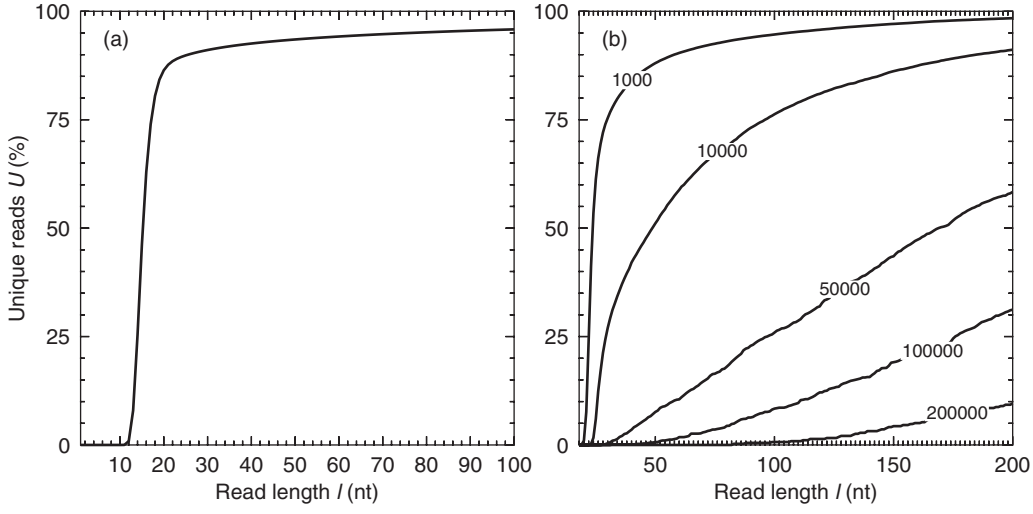

Figure 4. (**a**) Percentage of unique sub-sequences (*U*) for varying read length (*l*), in the *C.elegans* genome. (**b**) Percentage of the *C.elegans* genome covered by contigs greater than a threshold length as a function of read length. The horizontal axis starts at 18 nt, due to the limitations of reassembly below this length.

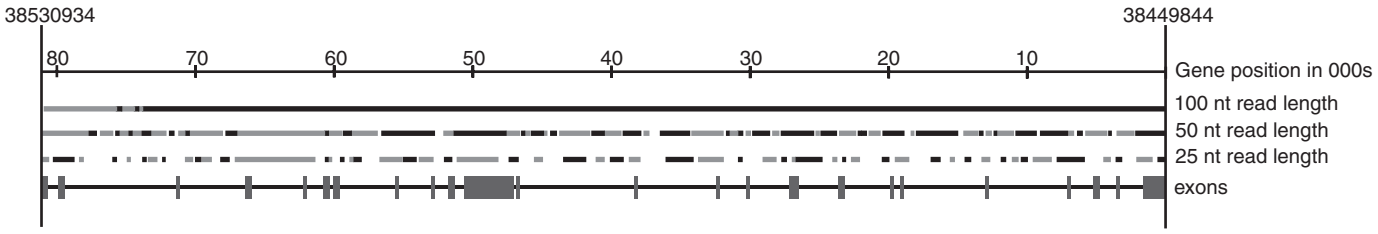

Figure 5. Reassembled contigs longer than 200 nt in the 81 090 nt of the BRCA1 gene. Reassembly was simulated from 25, 50 and 100 nt reads covering the whole of chromosome 17. Reassembled contigs are shown in alternating black and grey. Contigs maybe next to each other, or overlapping slightly without an unambiguous overlap existing between the contigs.

and therefore the application of the techniques to whole human genome re-sequencing shows promise.

The largest re-sequencing project reported for the human genome focussed on the non-repetitive sequences within chromosome 21, one of the smallest of the human autosomes. This allows the problem to be reduced to a series of smaller tasks both in terms of isolating a smaller section of the genome and the computational problems that are encountered. By reducing the size of the problem and focussing on regions that are not repeated it was possible to gain high quality sequence information for ~65% of the 21.7 Mb under study using just over three billion 25 nt oligonucleotide probes(24).

Our whole human genome uniqueness analysis shows that relatively large (25 to 50 nt) read lengths will be required for a significant percentage of sub-sequences to be unique making re-sequencing difficult. Uniqueness also places an absolute limit on the read length required for *de novo* reassembly. As these results suggest that *de novo* reassembly of the whole genome is likely to be difficult and due to the amount of computation time required to process a whole human genome reassembly analysis we also elected to reduce the problem to single chromosomes.

For the 246 Mb of chromosome 1, the largest human chromosome, the uniqueness analysis (Figure 2a) shows a rise reaching 50% uniqueness with 16 nt reads, 2 nt fewer than for the full genome. Uniqueness also plateaus earlier and continues to rise faster with reads of 20 nt able to probe 80% of the chromosome and 35 nt required for 90% coverage. Other human chromosomes we have analysed show a similar pattern (data not shown) where the rate of increase after the plateau has been reached being dependent on the size of the chromosome.

Figure 2b shows the analysis of a simulated reassembly of a complete set of reads obtained from chromosome 1. The analysis shows that with reads of 50 nt it is possible to reassemble 80% of the chromosome into contigs larger than 1000 nt; 17% is covered by contigs larger than 10 000 nt. Even with relatively modest read lengths of 25–30 nt significant proportions of chromosome 1 can be reassembled into contigs of 1000 to 10 000 nt. With a 500 nt read length 98.4% of the chromosome maybe reassembled into contigs larger than 10 000 nt.

Figure 5 gives an example of how contigs obtained by reassembling reads from a single chromosome cover a region of interest in the human genome. The reassembly of contigs from a complete set of 25, 50 and 100 nt reads of chromosome 17 was simulated. Figure 5 shows how these contigs cover the BRCA1 (25) gene, implicated in susceptibility to breast cancer. As can be seen the majority of exons are covered at a 50 nt read length, while 25 nt produces significantly less coverage. Using a 100 nt read length, two contigs are constructed which cover all of the exons in this gene. While contigs assembled from 50 nt reads cover the majority of the gene they cannot be positioned without additional information. However the fact that 87.5% of exons are covered by contigs of 200 nt or larger suggests that by using mapping information it should be possible to reconstruct the majority of the coding sequences in the genome. The efficient use of mapping information [such as matepair sequence data used to build 'scaffolds' by the Celera assembler (26)] is therefore an important area for future investigation.

CONCLUSION

Our analysis links the frequency analysis of repeated regions to the problem of genome sequence reassembly from short read information. We have defined the limits on the amount of unambiguous sequence that can be obtained with a given read length for a range of model genome sequences. Re-sequencing and *de novo* sequencing of viral genomes should be straightforward with read lengths of 18–25 nt depending on the length of the genome, and the number and length of repetitive regions. For the re-sequencing of bacterial genomes, once a threshold read length is reached, increasing the read length yields only small gains. *De novo* sequencing of a large proportion of the *E.coli* genome is possible with read lengths of 20–50 nt. In the case of the much larger and more repetitive human genome sequence whole genome re-sequencing will be limited with current proposed technology. However partitioning the problem, by focussing on single chromosomes or by neglecting the more difficult and repetitive parts of the genome, makes the problem more tractable.

As the analysis reported assumes a perfect data set these results represent the upper bounds of what information can be obtained by short read sequencing. The effect of random and systematic errors will be to reduce the amount of sequence information that can be obtained from reads of a given length. Specifically, erroronious reads may result in incorrect sequence data, breaks within contigs and contig mis-assembly. The challenge for the future is to identify in detail the effect of different potential experimental errors and to work to eliminate those that have the most detrimental effect on data quality so as to approach the upper bounds identified here. Our key finding is that with high quality data and currently available read lengths significant amounts of useful data can be obtained both by using a template sequence and for *de novo* sequencing. Thus, while there are distinct limits on the completeness of the

sequence information it is possible to obtain, short read sequencing has the potential to provide useful sequence information on large proportions of the genome for a range of model organisms cheaply and rapidly.

ACKNOWLEDGEMENTS

Requests for materials and software should be addressed to C.N. We thank the Information System Services department of the University of Southampton for access to the equipment on which this analysis was run. This work was supported by Research Councils UK through the Basic Technology Programme.

Conflict of interest statement. None declared.

REFERENCES

1. Shendure,J., Mitra,R.D. and Church,G.M. (2004) Advanced sequencing technologies: methods and goals. *Nature Rev. Gen.*, **5**, 335–344.
2. Brenner,S., Johnson,M., Bridgham,J., Golda,G., Lloyd,D., Johnson,D., Luo,S., McCurdy,S., Foy,M., Ewan,M. *et al.* (2003) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630–634.
3. Kling,J. (2003) Ultrafast DNA sequencing. *Nat. Biotechol.*, **21**, 1425–1427.
4. Miller,R.D., Duan,S., Lovins,E.G., Kloss,E.F. and Kwok,P.-Y. (2003) Efficient high-throughput resequencing of genomic DNA. *Genome Res.*, **13**, 717–720.
5. Ronaghi,M., Uhlen,M. and Nyren,P. (1998) DNA sequencing: a sequencing method based on real-time pyrophosphate. *Science*, **281**, 363–365.
6. Drmanac,R., Drmanac,S., Chui,G., Diaz,R., Hou,A., Jin,H., Jin,P., Kwon,S., Lacy,S., Moeur,B. *et al.* (2002) Sequencing by hybridization (sbh): advantages, achievements, and opportunities. *Adv. Biochem. Eng. Biotechnol.*, **77**, 75.
7. Seo,T.S., Bai,X., Kim,D.H., Meng,Q., Shi,S., Ruparel,H., Li,Z., Turro,N.J. and Ju,J. (2005) Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc. Natl Acad. Sci. USA*, **102**, 5926–5931.
8. Braslavsky,I., Hebert,B., Kartalov,E. and Quake,S.R. (2003) Sequence information can be obtained from single DNA molecules. *Proc. Natl Acad. Sci. USA*, **100**, 3960–3964.
9. Kartalov,E.P. and Quake,S.R. (2004) Microfluidic device reads up to four consecutive base pairs in DNA sequencing-by-synthesis. *Nucleic Acids Res.*, **32**, 2873–2879.
10. Mitra,R., D., Shendure,J., Olejnik,J. and Edyta-Krzymanska-Olejnik, church,G.M. (2003) Fluorescent *in situ* sequencing on polymerase colonies. *Anal. Biochem.*, **320**, 55–65.
11. Metzker,M., Raghavachari,R., Richards,S., Jacutin,S., Civitello,A., Burgess,K. and Gibbs,R. (1994) Termination of DNA synthesis by novel 3′-modified-deoxyribonucleoside 5′-triphosphates. *Nucleic Acids Res.*, **22**, 4259–4267.
12. Relogio,A., Schwager,C., Richter,A., Ansorge,W. and Valcarcel,J. (2002) optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Res.*, **30**, e51.
13. Margulies,M., Egholm,M., Ahman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braueman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
14. She,X., Jiang,Z., Clark,R.A., Liu,G., Cheng,Z., Tuzun,E., Church,D.M., Sutton,G., Halpern,A.L. and Eichler,E.E. (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*, **431**, 927–930.
15. Chaisson,M., Pevzner,P. and Tang,H. (2004) Fragment assembly with short reads. *Bioinformatics*, **20**, 2067–2074.
16. Fedrigo,O. and Naylor,G. (2004) A gene-specific DNA sequencing chip for exploring molecular evolutionary change. *Nucleic Acids Res.*, **32**, 1208–1213.
17. Whiteford,N. (2005) RepAnalyse: A Genome Repeat Analysis tool. Chemistry Department University of Southampton, Technical Report.
18. Kärkkäinen,J. and Sanders,P. (2003) Simple linear work suffix array construction. In Goos,G., Hartmanis,J. and Van Leeuwen,J. (eds). *Lecture Notes in Computer Science*. Springer-Verlag, New York, Vol. 2719, pp. 943–955.
19. Manzini,G. (2004) Two Space Saving Tricks for Linear Time LCP Computation. Technical Report 124, Dipartimento di Informatica, Universita del Piemonte.
20. Manber,U. and Myers,G. (1993) Suffix arrays: a new method for on-line string searches. *SIAM J. Comput.*, **22**, 935–948.
21. Rudd,K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids. Res.*, **28**, 60–64.
22. Wong,C.W., Albert,T.J., Vega,V.B., Norton,J.E., Cutler,D.J., Richmond,T.A., Stanton,L.W., Liu,E.T. and Miller,L.D. (2004) Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays. *Genome Res.*, **14**, 398–405.
23. Bennett,S. (2004) Solexa Ltd. *Pharmacogenomics*, **5**, 433–438.
24. Patil,N., Berno,A.J., Hinds,D.A., Barrett,W.A., Doshi,J.M., Hacker,C.R., Kautzer,C.R., Lee,D.H., Marjoribanks,C., McDonough,D.P. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.
25. Albertsen,H., Plaetke,R., Ballard,L., Fujimoto,E., Connolly,J., Lawrence,E., Rodriguez,P., Robertson,M., Bradley,P., Milner,B. *et al.* (1994) Genetic mapping of the brca1 region on chromosome 17q21. *Am. J. Hum.*, **54**, 516–525.
26. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A. and Holt,R.A. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
27. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (2000) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.