De novo sequencing and variant calling with nanopores using PoreSeq

Tamas Szalay1 & Jene A Golovchenko1,2

The accuracy of sequencing single DNA molecules with nanopores is continually improving, but de novo genome sequencing and assembly using only nanopore data remain challenging. Here we describe PoreSeq, an algorithm that identifies and corrects errors in nanopore sequencing data and improves the accuracy of de novo genome assembly with increasing coverage depth. The approach relies on modeling the possible sources of uncertainty that occur as DNA transits through the nanopore and finds the sequence that best explains multiple reads of the same region. PoreSeq increases nanopore sequencing read accuracy of M13 bacteriophage DNA from 85% to 99% at 100x coverage. We also use the algorithm to assemble Escherichia coli with 30x coverage and the λ genome at a range of coverages from 3x to 50x. Additionally, we classify sequence variants at an order of magnitude lower of coverages from 3× to 50×. Additionally, we classify sequence variants at an order of magnitude lower

Initially proposed two decades ago, nanopore sequencing is now becoming competitive with other DNA sequencing methods. Because it provides long reads from a single DNA molecule, nanopore sequencing can resolve large-scale genomic repeats that were previously intractable due to inherent limitations in assembling short reads1. The approach uses a small transmembrane pore whose narrowest constriction is just wide enough to allow single-stranded DNA to pass through (Fig. 1a). Voltage is applied across the membrane to set up an ionic current, which electrophoretically draws the DNA into the pore. Current levels reflect the changes in conductance caused by the presence of different DNA bases in the narrowest region of the pore. An enzymatic motor, such as a polymerase2 or helicase, is used to ratchet the DNA strand through the pore one base at a time, and the resulting changes in ionic current are used to infer the sequence. Nanopores have recently been used to obtain long-read data with quantifiable accuracy3,4. Further improvements were enabled by Oxford Nanopore Technologies’ 2.048-nanopore, USB-powered MinION sequencer5-7, which can collect tens to hundreds of megabases of data in a single run, producing long sequencing reads at high coverage and a moderate accuracy of around 85% (Fig. 1b). Long-read data from the device have been used as a scaffold to aid in the assembly of shorter, more accurate reads6,8 and to study large-scale structure1-5. To correct errors in these low-accuracy reads directly, however, a more specific model of the sequencing technique is needed9.

The dominant source of error in nanopore sequencing is the simultaneous influence of multiple adjacent nucleotides on the ionic current signal. Up to five bases have been shown to influence the instantaneous current10,11. As a result, the number of distinct current levels can increase from 4 to as many as 1,024, thereby decreasing the signal-to-noise ratio for base determination (Fig. 1c). Extracting accurate sequence information is further complicated by the stochastic behavior of the DNA molecule, the enzyme and the nanopore complex, which can lead to both missing and additional current levels (Fig. 1d). Enzymes can randomly ratchet the DNA through the pore too quickly to be electronically detected, and, as a result, the discretized form of the data (Fig. 1a, red line) will have that particular level omitted. Fluctuations or conformational changes can also lead to sudden jumps in conductance that can easily be mistaken for the signal produced by DNA advancing through the nanopore, even though the enzyme stays clamped on the same base. Furthermore, certain enzymes even exhibit random backwards motion12. These confounding factors mean that a large number of possible DNA sequences could produce the observed current levels, making the true sequence difficult to obtain.

Here we show that by coupling latent information in the ionic current data from multiple reads with a statistical model of the underlying physical system, we can increase the accuracy of DNA sequence determination. We developed an algorithm, PoreSeq, that models the probability of observing a current level given a 5-mer sequence (Fig. 1c), as well as the probability of skipping or staying on a given current level (Fig. 1d and Supplementary Fig. 1). Then, given a sequence, a set of observed current data, and an alignment that maps the discretized levels to the positions in the sequence, these probabilities can be multiplied together to compute a total observation likelihood for that sequence. The best sequence is defined as the sequence with the highest likelihood, which can be obtained from one or multiple reads.

Although techniques exist for finding the maximum-likelihood sequence using the dynamic programming13-based Viterbi algorithm14,15, they are limited to analyzing only one or two reads at a time owing to the computational complexity of simultaneously aligning the
current levels and finding the true sequence (Supplementary Note 1 and Supplementary Fig. 1). Other methods such as dynamic time warping\(^3,16\) can align many sets of levels to one another without prior knowledge about the sequence, but in doing so, they sacrifice valuable information about the sequence-specific behavior (e.g., current level distributions) contained in the statistical model. The alignment of multiple DNA sequences is a related problem in which various iterative optimization approaches have proven useful\(^17,18\); we show here that the same general class of algorithm can reliably surpass the accuracy of single-molecule reads (Fig. 1b) when the implementation is tailored appropriately to the task of nanopore sequencing.

### RESULTS

#### PoreSeq algorithm

The base-calling algorithm presented here is designed to accurately determine the \textit{de novo} sequence using the discretized ionic current data from an arbitrary number of independent nanopore reads of the same region of DNA, including partial or reverse complement reads. It does this by iteratively finding the sequence that maximizes the total observation likelihood for all of the reads according to the statistical model. Optimizing the sequence in this iterative way requires searching through the vast space of all possible sequences of a given length. We limit this search to sequences that could plausibly fit the data by repeatedly testing and making only those local changes that are expected to improve the observation likelihood.

The components of the PoreSeq algorithm are illustrated in Figure 2a. The process starts with a candidate sequence considered to be an initial best guess for the optimization routine and a number of observed nanopore reads that cover some of the region over which we are optimizing. A natural choice for the initial guess is one of the read’s single-molecule sequences computed by existing Viterbi methods (e.g., as provided directly by Oxford Nanopore’s cloud-based service for the MiniON). This guess is then gradually improved by introducing artificial mutations into the candidate sequence. These mutations are drawn from alternate versions of the candidate sequence, generated by a modified single-molecule Viterbi algorithm that deliberately introduces some randomness (Supplementary Note 2). Although these alternate sequences globally fit the data slightly worse than the best candidate, they may contain a short region (anywhere from one to dozens of bases) that has a higher sequence-aligned likelihood in that region compared to the current best candidate (Fig. 2a, iii, shaded and Supplementary Fig. 2). This stretch of sequence is then locally substituted into the best candidate’s sequence, and the full observation likelihood recalculated. The mutated sequence is kept if this likelihood exceeds the current best. Once all such mutations have been exhaustively tested, new alternates are generated, and the procedure repeats until no more changes are found. This technique makes sequence space optimization feasible by screening for likely mutations and thus avoiding a prohibitively dense search over all possible sequences.

The details of the likelihood calculation are illustrated in Figure 2b. The matrix depicts the alignment of a single read with a candidate sequence and yields the maximum likelihood (over all possible alignments) of observing that read given the sequence. The calculation of each cell in the matrix proceeds in a similar manner to canonical alignment algorithms (dynamic time warping\(^16\) or Smith-Waterman\(^19\)): the row and column coordinates of a single read represent the indices of the current levels and the bases, respectively, and the value contained in that cell is the maximum likelihood over all possible paths that can be taken to reach that cell. The cells are filled in using the likelihood values from the surrounding three cells directly above and to the left, as they correspond to the different possible cases of enzymatic motion according to the statistical model (e.g., a step from the left is a skip, shown in red, Fig. 2b, as the sequence advances without a corresponding current level). The resulting likelihood value of the new matrix cell is the maximum likelihood out of all of these starting cells multiplied by the probability of the new sequence/observation pair and the skip/stay probabilities, if applicable (Supplementary Note 3 and Supplementary Fig. 3). At the end of the procedure, the highest likelihood in the entire matrix is then the maximum likelihood over all alignments, and the best alignment itself can be obtained by tracing the matrix backwards (Fig. 2b, shaded blue, arrows) from the maximum. A separate matrix is calculated for each read and the resulting likelihoods are then multiplied together to get the total observation.
likelihood; the computation time hence scales linearly with the coverage depth.

De novo accuracy

We used PoreSeq to obtain de novo sequence information using MinION nanopore data (Fig. 3a). The data are taken from the M13mp18 phage vector digested by EcoRI, yielding identical double-stranded fragments 7,249 bases long, as well as from λ phage DNA sheared to ~8-kb fragments (Supplementary Fig. 4) and from a published E. coli data set 5. After preparing the molecules as recommended, we ran a 24-h sequencing protocol. The initial sequence analyses were done using Oxford Nanopore Technologies’ cloud-based Metrichor service, which computed the sequence corresponding to each detected molecule separately. The models (internally trained on the E. coli genome) used by Oxford to map 5-base sequences to observed current levels were also stored, and these models were used in this work without modification along with the offsets and scaling provided. The additional skip and stay parameters for the model were trained using the results of a sequencing run of identical 3.6-kb fragments of λ DNA that are included with the MinION device for calibration.

We calculated the de novo sequence accuracies of M13mp18 as a function of single-molecule coverage (Fig. 3a), with each molecule consisting of both a template and complement read. We measured the minimum and maximum accuracy obtained from trials using different random subsets of molecules at the specified coverage. We emphasize that no information about the true sequence was used, nor any information (e.g., statistics or fit parameters) from molecules outside of those included in a single trial. Because of the higher levels of coverage needed to reach a practical 99% accuracy, small DNA samples might require the use of PCR, in which case the mutation scoring would be modified to discard the lowest-scoring strands so as to account for the errors introduced during replication.

We also used PoreSeq in a standard genome assembly pipeline by assembling λ DNA from nanopore reads at a range of coverages, and by correcting errors and assembling E. coli from a previously published data set (Fig. 3a). Our technique is similar to the Hierarchical Genome Assembly Pipeline (HGAP) 9 developed for Pacific Biosciences sequence data, consisting of three stages: fragment error correction, assembly and assembled genome refinement. When the existing pipeline was run on MinION sequence data directly, the accuracy was limited to around 96% (Fig. 3a, green line); however, replacing both the existing fragment error correction and refinement stages with our technique increased the accuracy to 99% (Fig. 3a, red line). Notably, because long sequences are split into smaller segments for parallel processing, the technique is inherently scalable to larger genomes. We show this by running the same pipeline on a published E. coli data set from the MinION sequencer 2. We used our algorithm to correct errors in sequencing reads and assembled fragments into a single contiguous genome. When the algorithm was used to correct errors in the assembled sequence a second time, the final accuracy was 98.5% (Fig. 3a and Supplementary Note 4), which we would expect to improve further if more recent MinION data were used.

Variant calling

Another benefit of this approach is that it easily enables sequence variant comparison (Fig. 3b,c). It is often necessary to distinguish between known single-nucleotide variants (SNV) occurring at a low density ( ~1 SNV per 2 kb in humans 20). We took the actual M13mp18 sequence and computationally mutated it in one position (replacing, inserting or deleting a single base) to generate a SNV of the original sequence. The observation likelihood feature was then used to compute the likelihood score of both the original and mutated sequences in order to call the correct variant as the higher of the two likelihoods. Even at low coverages the correct unmodified sequence was accurately identified as having a higher likelihood score (Fig. 3b). This variant calling feature outperformed de novo sequencing at lower coverages because the vast majority of single-nucleotide-mutated sequences would yield a considerably different current signature, making them easy to distinguish.

We also compared our approach to previous sequence-based variant calling analyses using nanopore data 1 (Fig. 3c). Substitution errors were introduced into the M13mp18 sequence at a rate of 1%, and we then attempted to correctly identify them and recover the original sequence. By setting a threshold on the difference in observation likelihood when calling each base variant, we calculated the probability of false positives and false negatives (referred to as precision and recall, respectively) at a range of coverages, to find the maximum classification accuracy. We found a peak F-score accuracy of 99.1% at a coverage of only 16×, demonstrating similar performance to previous results 1 but at an order-of-magnitude lower coverage.

Figure 2 PoreSeq algorithm. (a) A high-level overview of the algorithm. The process starts with multiple nanopore reads and a single candidate sequence shown in blue (i). Aligned alternate sequences from which mutations are drawn are shown in red with a specific one highlighted (ii). The alternates must be similar enough to the candidate that their sequences can be aligned but otherwise vary throughout their full length. The plot contains the aligned local likelihood scores across all reads (calculated as in b) for the candidate and each alternate (iii). Local regions where an alternate’s likelihood score exceeds the candidate’s are shaded and correspond to a short region where the alternate sequence is likely more accurate than the best candidate. The alternate’s local sequence from this region is substituted into the best candidate and the full maximum likelihood score is recomputed (iv), at which point the higher-scoring sequence becomes the new best (v). The vertical dotted lines denote a specific mutation taken from the highlighted red strand in parts (i) through (iv). (b) The maximum likelihood over all possible alignments of a single read and sequence is computed using a Smith-Waterman style dynamic programming approach. The values of the matrix elements are the maximum likelihood over all valid paths to that element, whereas the lighter colors represent a higher local probability of that particular sequence/current pair according to the statistical model. The maximum likelihood score is the highest value in the whole matrix and the full best alignment can be found by backtracking through the matrix as illustrated by the arrows. The local probabilities of each aligned read can be mapped back to the sequence and combined (bottom), at which point they can be compared to the scores from other sequences (2a, iii, comparison between red and blue lines).

Variant calling Another benefit of this approach is that it easily enables sequence variant comparison (Fig. 3b,c). It is often necessary to distinguish between known single-nucleotide variants (SNV) occurring at a low density (~1 SNV per 2 kb in humans 20). We took the actual M13mp18 sequence and computationally mutated it in one position (replacing, inserting or deleting a single base) to generate a SNV of the original sequence. The observation likelihood feature was then used to compute the likelihood score of both the original and mutated sequences in order to call the correct variant as the higher of the two likelihoods. Even at low coverages the correct unmodified sequence was accurately identified as having a higher likelihood score (Fig. 3b). This variant calling feature outperformed de novo sequencing at lower coverages because the vast majority of single-nucleotide-mutated sequences would yield a considerably different current signature, making them easy to distinguish.

We also compared our approach to previous sequence-based variant calling analyses using nanopore data 1 (Fig. 3c). Substitution errors were introduced into the M13mp18 sequence at a rate of 1%, and we then attempted to correctly identify them and recover the original sequence. By setting a threshold on the difference in observation likelihood when calling each base variant, we calculated the probability of false positives and false negatives (referred to as precision and recall, respectively) at a range of coverages, to find the maximum classification accuracy. We found a peak F-score accuracy of 99.1% at a coverage of only 16×, demonstrating similar performance to previous results 1 but at an order-of-magnitude lower coverage.
DISCUSSION

Our results show that multiple nanopore reads can be combined to reach accuracies over 99%, compared to the ~85% seen with single molecules, and many further improvements are expected both in the nanopore biochemistry and the physical models capturing their behavior. In particular, the inclusion of current level durations will be necessary in dealing with homopolymer DNA regions, which we have found are responsible for over half of the errors at 99% accuracy (Supplementary Note 5 and Supplementary Fig. 5). The model can also be extended to detect base methylation, the identification of which requires an estimated 5–19 repeated reads21,22. Other improvements are possible through better control of enzymatic ratcheting23 or the inclusion of a wider variety of pore mutants24–26 to obtain pore-specific current data on the same sequence. The code was designed to be flexible in handling such modifications with far fewer constraints than Viterbi-based approaches. All code is publicly available at http://github.com/tszalay/poreseq and a brief description is included in Supplementary Note 6.

During the review of this manuscript, a related approach was submitted and published27. Both methods employ a similar error model but differ in that Loman et al.27 use only sequence information for the initial pre-assembly error correction, whereas our approach also uses nanopore current data for both the initial correction and the final refinement pass (made possible by the mutation finding process in Fig. 2). This allows us to find the consensus sequence and analyze the error across a wider range of coverages, from as low as 4× and 8× (Fig. 3). The coverage of *E. coli* in both our paper and Loman et al.’s27 is roughly 30× and the difference in accuracy (98.5% vs. 99.5%, respectively) is likely a consequence of the great improvement in MinION chemistry and models since the *E. coli* data used in our work were generated2. Our demonstration of variant calling additionally provides insight into the capabilities of these methods beyond error correction.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. All of the M13 and lambda MinION sequencing data used for this manuscript are available in the public Dryad archive (10.5061/dryad.84d4j), and the *E. coli* data can be found at the European Nucleotide Archive under accession number ERP007108. All code is publicly available at http://github.com/tszalay/poreseq.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank E. Brandin for molecule preparation, D. Branton for helpful algorithmic discussions and Figure 1a, and A. Kuan and M. Burns for feedback on this manuscript. The computations in this paper were run on the Odyssey cluster supported by the Faculty of Arts and Sciences Division of Science, Research Computing Group at Harvard University, and the work was supported by the National Institutes of Health Award no. R01HG008703 to J.A. Golačenko and D. Branton.

AUTHOR CONTRIBUTIONS

T.S.: algorithm development, data analysis and interpretation, writing of manuscript; J.A.G.: data analysis and interpretation, writing of manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Jain, M. et al. Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* **12**, 351–356 (2015).
2. Lieberman, K.R. et al. Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *J. Am. Chem. Soc.* **132**, 17961−17972 (2010).
3. Laszlo, A.H. et al. Decoding long nanopore sequencing reads of natural DNA. *Nat. Biotechnol.* **32**, 829–833 (2014).
4. Loman, N.J. & Quinlan, A.R. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* **30**, 3399−3401 (2014).
5. Quick, J., Quinlan, A.R. & Loman, N.J. A reference bacterial genome dataset generated on the MinION(TM) portable single-molecule nanopore sequencer. *Gigascience* **3**, 22 (2014).
6. Ashton, P.M. et al. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.* **33**, 296−300 (2014).
7. Bayley, H. Nanopore sequencing: from imagination to reality. *Clin. Chem.* **61**, 25−31 (2014).
8. Koren, S. et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693−700 (2012).
9. Chin, C.-S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563−569 (2013).
10. Nanrao, E.A. et al. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat. Biotechnol.* **30**, 349−353 (2012).
11. Nanrao, E.A., Derrington, I.M., Pavlenok, M., Niederweis, M. & Gundlach, J.H. Nucleotide discrimination with DNA immobilized in the MspA nanopore. *PLoS ONE* **6**, e25723 (2011).
12. Cheff, G.M. et al. Automated forward and reverse ratchetting of DNA in a nanopore at 5-A precision. *Nat. Biotechnol.* **30**, 344−348 (2012).
13. Belman, R. Dynamic Programming: A Bibliography of Theory and Application (Dover Publications, Reprint Edition (2003), 1957).
14. Viterbi, A.J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory* **13**, 260−269 (1967).
15. Ting, W., Corner, J. & Aksimentiev, A. DNA base-calling from a nanopore using a Viterbi algorithm. *Biophys. J.* **102**, L37–L39 (2012).
16. Vintsyuk, T.K. Speech discrimination by dynamic programming. Cybernetics 4, 52–57 (1972).
17. Gotoh, O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J. Mol. Biol. 264, 823–838 (1996).
18. Brudno, M. & Morgenstern, B. Fast and sensitive alignment of large genomic sequences. Proc. IEEE Comput. Soc. Bioinform. Conf. 1, 138–147 (2002).
19. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197 (1981).
20. Sachidanandam, R. et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409, 928–933 (2001).
21. Schreiber, J. et al. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. Proc. Natl. Acad. Sci. USA 110, 18910–18915 (2013).
22. Wessoe, Z.L., Schreiber, J. & Akeson, M. Nanopores discriminate among five C5-cytosine variants in DNA. J. Am. Chem. Soc. 136, 16582–16587 (2014).
23. Gollnick, B. et al. Probing DNA helicase kinetics with temperature-controlled magnetic tweezers. Small 11, 1273–1284 (2015).
24. Howorka, S., Cheley, S. & Bayley, H. Sequence-specific detection of individual DNA strands using engineered nanopores. Nat. Biotechnol. 19, 636–639 (2001).
25. Butler, T.Z., Pavlenok, M., Derrington, I.M., Niederweis, M. & Gundlach, J.H. Single-molecule DNA detection with an engineered MspA protein nanopore. Proc. Natl. Acad. Sci. USA 105, 20647–20652 (2008).
26. Maglia, G., Restrepo, M.R., Mikhailova, E. & Bayley, H. Enhanced translocation of single DNA molecules through alpha-hemolysin nanopores by manipulation of internal charge. Proc. Natl. Acad. Sci. USA 105, 19720–19725 (2008).
27. Loman, N.J., Quick, J. & Simpson, J.T. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Methods 12, 733–735 (2015).
Online Methods

M13 restriction digest. Four micrograms of M13mp18 RFI (New England BioLabs, cat. no. N4018S) DNA were digested with EcoRI restriction enzyme in a 100 µl reaction volume for 2 h at 37 °C, and then heated for 30 min at 65 °C to inactivate the enzyme. The digested DNA mix was applied to Micro Bio-Spin 30 chromatography columns (Bio-Rad Laboratories Inc., cat. no. 732-6223) equilibrated with molecular grade water, centrifuged according to the manufacturer’s instructions, and a column flow-through volume containing 1 µg of desalted DNA was adjusted to 85 µl with water for subsequent library preparation.

λ DNA shearing. λ genomic DNA for sequencing was prepared by shearing the DNA using the Covaris g-TUBE (Covaris, Inc., cat. no. 520079) to generate targeted DNA fragment sizes of ~8 kbp. One microgram of Lambda DNA (New England BioLabs, cat. no. N30135) in a total volume of 80 µl water was applied to a Covaris g-TUBE and centrifuged, according to manufacturer’s instructions, for 1 min at 7,200 r.p.m. in a Heraeus Biofuge fresco (Kendro Laboratory Products) at 30 °C. Five microliters of DNA CR Oxford Nanopore Technologies (ONT) was added to the sheared DNA and the sample prepared for sequencing following the Oxford Nanopore MAP Genomic DNA Sequencing Kit protocol (ONT prod. Code SQK-MAP003).

Sequence library preparation. End repair of the digested DNA samples was done using NEBNext End Repair module (New England BioLabs, cat. no. E6050S) by adding 10 µl of reaction buffer and 5 µl of enzyme mix and incubating at room temperature for no longer than 30 min. To purify the end-repaired DNA, 100 µl (1x volume) of Agencourt AMPure XP beads (Beckman Coulter Inc., cat. no. A63880) were gently mixed in, incubated at room temperature for 15 min, and the beads separated from the reaction supernatant using a magnetic rack. The DNA was eluted from the beads in 25 µl of 10 mM Tris-HCl pH 8.5 and dA-tailing performed at room temperature for 60 min by adding 3 µl of reaction buffer and 2 µl of enzyme supplied in the NEBNext dA-Tailing module (New England BioLabs, cat. no. E6053S).

The Oxford Nanopore MAP Genomic DNA Sequencing Kit (ONT prod. Code SQK-MAP004 for M13, SQK-MAP003 for λ) was used to further prepare the DNA samples for sequencing. Kit reagents were thawed and stored on ice (HP adaptor, adaptor mix and fuel mix), or at room temperature (2× wash buffer, elution buffer, and EP buffer) before use. Following the manufacturer’s protocol, a ligation step was performed by combining (in order) 30 µl of dA-tailed DNA, 8 µl of water, 10 µl of Adaptor mix, 2 µl of HP adaptor, 50 µl of Blunt/TA Ligase Master Mix (New England BioLabs, cat. no. M0367S), and incubating the mixture at room temperature for 10 min. Ten microliters of His-Tag Dynabeads (Life Technologies, cat. no. 10103D) were washed twice with 250 microliters of 1× wash buffer and suspended with 100 microliters of 2× wash buffer. This volume of washed His-Tag beads was added directly to the ligation mixture, incubated for 5 min at room temperature, and the beads collected using a magnetic rack. Once separated from the reaction supernatant, the pelleted beads were carefully rinsed twice (without suspension) using 250 µl of 1× wash buffer. All wash buffer was subsequently removed and the DNA library eluted from the beads with 25 µl of elution buffer. Prior to sequencing on the MinION, 140 µl of EP buffer was added to a 6 µl aliquot of the DNA library, followed by 4 µl of fuel mix and thorough mixing.

Data acquisition. Data were acquired with an Oxford MinION device connected to a sufficiently capable computer (Windows 7; MinKNOW software; USB3; SSD; i7 processor). A new version R7.3 flowcell was used for each run; the flowcell was initially primed with two injections of 150 µl of EP Buffer, spaced 10 min apart, after which the MinKNOW software was used to measure pore quality and activity. The 150 µl of previously prepared sequencing mix was then loaded into the MinION flowcell as specified by the manufacturer. A 24-h sequencing protocol was selected in the software and the device was allowed to run until all pores and sample were exhausted.

Data analysis. Initial data analysis was performed by Oxford Nanopore Technologies’ cloud-based Metrichor service, which runs proprietary versions of the Viterbi algorithm similar to those described in Supplementary Note 1. The service uploads the basic HDF5 (‘fast5’) files generated by the sequencing software and returns the files with additional processed data appended. A number of these analyzed features were used in the present work: the parsing of each molecule to separate the current levels corresponding to the template and complement strands in each double-stranded molecule (and the removal of hairpin/adaptor levels); the Oxford-trained statistical models used to map 5-mers to current levels; the offset and scaling between the molecule and model current levels; estimates of the skip and stay probabilities per molecule; and the two one-dimensional Viterbi and one two-dimensional (2D) Viterbi-computed sequence for each molecule. The details of the extraction and use of each of these features is well-documented in Supplementary Note 6 and Supplementary Data 1. Only the models and scaling are required; the Viterbi sequences were merely used to help speed up convergence of the algorithm.

Model training. The skip, stay and insertion parameters of the model were trained using a sample of ONT provided “DNA CS” (calibration sample) prepared in the same manner as the M13mp18 post-digest using the material supplied with the SQK-MAP004 kit. Of the 907 molecules sequenced, a subset of 29 was selected at random and de novo sequencing performed. The parameters were arbitrarily seeded to 5% for skips/stays and 2% for insertions (Supplementary Note 3), then randomly perturbed and the change kept if the resulting DNA CS sequence accuracy increased. After a few hundred iterations, the parameters had converged and no improvements were found; these were the parameters used in the final analysis used to generate Figure 3. The data used to generate Figure 3 are presented in Supplementary Data 1. The authors acknowledge that more precise means of training the parameters are possible, but it was found that the de novo accuracies were fairly insensitive to small (~20%) changes in the parameters, and as a result the present method was found to be sufficient.

Accuracy calculation. When the term “accuracy” was used to refer to a DNA sequence relative to the M13mp18 reference (Figs. 1b and 3a), this accuracy was calculated as follows: first, the optimal alignment between the candidate sequence and a reference was found using MatLab’s swalign function (Smith-Waterman alignment) using default parameters, which uses a match score of +5, a mismatch penalty of −4, and a gap/extend penalty of −8. The alignment is assumed to be good enough to cover all of the candidate sequence unless otherwise noted. Next, the accuracy was computed as the number of matching bases divided by the total number of bases in the alignment, defined as matches + mismatches + insertions + deletions. For the variant calling and de novo sequencing trials of M13, the accuracy was calculated in the region covered by at least 75% of the strands, to compensate for end-trimming in the data. In the case of λ DNA, the accuracy reported is % identity as calculated by MUMmer28, which was found to not differ significantly from other definitions at the error rates shown.

M13 de novo sequencing. To reconstruct the original sequence from nanopore reads of M13 bacteriophage DNA, the code was executed on a specific number of fragments as shown in the coverage plot of Figure 3. Suitable fragments were defined as those with double-stranded information available whose 2D basecalls had between 5,000 and 8,000 bases (compared vs. the full M13 with 7,249 bases), in order to filter out partial or repeated molecules. Optimization was then run using each molecule’s own 2D Viterbi sequence generated by Metrichor as the alternate sequence inputs to the algorithm, until no mutations were found, typically taking around three iterations (with up to 1,000 mutations possible in each iteration). Once this phase was completed, alternate sequences were generated as described in Supplementary Note 2, to find likely mutations that did not appear in any of the individual 2D sequences. In each iteration, 12 alternate sequences were generated with identity/similarity to the candidate ranging from 60% to 100%, and these iterations were alternated with the testing of every possible single-base mutation (Supplementary Note 6) to ensure that no mutations were missed; this was repeated at most five times and found to sufficiently capture all mutations.

Lambda assembly. The code was used to correct errors in each individual read as follows: first, the 2D-basecalled sequences of all reads were extracted from the returned ONT files. LAST was then used to perform overlap alignment of each reads sequence to all other sequences, and these alignments were used to seed the algorithm and improve the accuracy of each fragment sequence. The latest Celera assembler (v6383 from SVN) was then used to assemble the corrected sequences into a single contig. All reads were re-aligned with the contig once again using LAST and corrected in 6,000-base fragments which were finally reassembled into a
single, high-accuracy sequence. All commands and parameters used can be found in the Supplementary Data. Each point in Figure 3a is from a single random subset of all reads, and the coverage shown is the theoretical maximum value of (total bases in reads)/48,502. The specifics of the iterations for the correction are the same as those described in the previous section.

**E. coli assembly.** We also used the algorithm to correct errors in and assemble a previously published E. coli MinION data set5 using a similar procedure to that for DNA. Briefly, pre-correction was done with our code using all of the 2D data but limited to 10× coverage for speed, followed by PBcR/Celera assembly of the corrected reads, and finished with a final refinement pass of our code at a 25× coverage limit. The assembled and refined *de novo* genome is 4.584 Mb long and covers 99.65% of the *E. coli* reference at an identity of 98.48% (see Supplementary Note 4 for more details). Due to the lower accuracy inherent in the older data and chemistries used, we were not able to assemble the sequences without using our code for pre-correction, as the self-correction pipeline in PBcR was unable to improve the 2D sequences beyond 87%.

28. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* 5, R12 (2004).