

# Improved data analysis for the MinION nanopore sequencer

Miten Jain<sup>1,2</sup>, Ian T Fiddes<sup>1,2</sup>, Karen H Miga<sup>1,2</sup>, Hugh E Olsen<sup>1,2</sup>, Benedict Paten<sup>1,2</sup> & Mark Akeson<sup>1,2</sup>

**Speed, single-base sensitivity and long read lengths make nanopores a promising technology for high-throughput sequencing. We evaluated and optimized the performance of the MinION nanopore sequencer using M13 genomic DNA and used expectation maximization to obtain robust maximum-likelihood estimates for insertion, deletion and substitution error rates (4.9%, 7.8% and 5.1%, respectively). Over 99% of high-quality 2D MinION reads mapped to the reference at a mean identity of 85%. We present a single-nucleotide-variant detection tool that uses maximum-likelihood parameter estimates and marginalization over many possible read alignments to achieve precision and recall of up to 99%. By pairing our high-confidence alignment strategy with long MinION reads, we resolved the copy number for a cancer-testis gene family (CT47) within an unresolved region of human chromosome Xq24.**

In 2014, Oxford Nanopore Technologies (ONT) enlisted several hundred laboratories to beta-test its 100-gram MinION sequencing device. The MinION sequences individual DNA molecules, providing very long read lengths to help overcome some of the drawbacks of short-read sequencing. As part of the MinION Access Program (MAP), we set out to characterize the performance and characteristics of the sequencing platform and to develop it to call single-nucleotide variants (SNVs) and resolve the repeat structure of highly repetitive regions. Our open-source analysis tools are available online (**Supplementary Software 1 and 2**; <https://github.com/mitenjain/nanopore> and <https://github.com/benedictpaten/margin-Align> for the nanopore and marginAlign pipelines, respectively).

The MinION reads the sequences of individual DNA strands as they are driven through biological nanopores by an applied electric field. The rate at which each DNA strand moves through a nanopore is controlled by a processive enzyme bound to the DNA at the pore orifice. Up to 512 DNA molecules can be read simultaneously using amplifiers that independently address each nanopore. Changes in ionic current, each associated with a unique five-nucleotide DNA *k*-mer, are detected as DNA molecules translocate through the nanopores at single-nucleotide precision. DNA bases are called using cloud-based software (Metrichor) provided by ONT that employs hidden Markov models (HMMs) to infer sequences from these current changes.

We determined MinION sequence-read quality and errors by analyzing the genome of M13mp18, a phage from *Escherichia coli* host strain ER2738 with a 42% average GC content and a 7.2-kb genome (Online Methods). Using expectation maximization, we inferred maximum-likelihood estimates (MLEs) for the rates of insertions, deletions and substitutions in MinION reads. We then realigned the reads to generate high-confidence alignments and used the MLE models to demonstrate that MinION reads can be used for accurate SNV calling. By coupling this alignment strategy with long MinION reads, we resolved the tandem-repeat organization of a CT47 cancer-testis gene family on an unfinished segment of human chromosome Xq24. Our results document the substantial improvements in the MinION's performance achieved during MAP.

## RESULTS

### The MinION reads both strands of duplex DNA

We prepared libraries as recommended by ONT, with modifications to ensure the integrity of high-molecular weight DNA (Online Methods). A DNA construct analyzed on the MinION (**Fig. 1**) is composed of a lead adaptor that loads the processive enzyme and facilitates DNA capture in the applied electric field; the DNA insert of interest; a hairpin adaptor that permits consecutive reading of the template and complement strands by the nanopore; and a tethering adaptor that concentrates DNA at the membrane surface.

Translocation of a single M13 genomic double-stranded DNA (dsDNA) copy through a MinION pore involves a series of steps, each associated with an identifiable ionic current pattern (**Fig. 1**). These include (i) the open pore; (ii,iii) capture and translocation of the lead adaptor; (iv) translocation of the template strand; (v) translocation of the hairpin adaptor; (vi) translocation of the complement strand (giving two-directional or 2D sequence data); (vii) translocation of the tethering adaptor; and (viii) release of the DNA strand into the *trans* compartment and return to the open-channel ionic current. At this point another DNA molecule can be captured and analyzed by the pore.

Over the first 6-month period of MAP, three MinION chemistry versions and numerous base-calling algorithm updates resulted in successive improvements in device performance (**Supplementary Fig. 1**). The average observed identity (the proportion of bases in a read that align to a matching base in a reference sequence) for

<sup>1</sup>UC Santa Cruz Genomics Institute, Santa Cruz, California, USA. <sup>2</sup>Department of Biomolecular Engineering, University of California, Santa Cruz, California, USA. Correspondence should be addressed to B.P. ([benedict@soe.ucsc.edu](mailto:benedict@soe.ucsc.edu)) or M.A. ([makeson@soe.ucsc.edu](mailto:makeson@soe.ucsc.edu)).

**Figure 1** | Molecular events and ionic-current trace for a 2D read of an M13 phage dsDNA molecule. **(a)** Steps in DNA translocation through the nanopore: (i) open channel; (ii) dsDNA with lead adaptor (blue), bound molecular motor (orange) and hairpin adaptor (red) is captured by the nanopore; capture is followed by translocation of the (iii) lead adaptor, (iv) template strand (gold), (v) hairpin adaptor, (vi) complement strand (dark blue) and (vii) trailing adaptor (brown); and (viii) status returns to open channel. **(b)** Raw current trace for the passage of the M13 dsDNA construct through the nanopore. Regions of the trace corresponding to steps i–viii are labeled. **(c)** Expanded time and current scale for raw current traces corresponding to steps i–viii. Each adaptor generates a unique current signal used to aid base calling.

2D reads was 66% in June 2014 (R6.0 chemistry release), 70% in July 2014 (R7.0 chemistry release), 78% in October 2014 (R7.3 chemistry release) and 85% in November 2014 (Metrichor R.7X 2D version 1.9 update). The present study was based on MinION R7.3 chemistry and R7.X version 1.9 base-calling algorithms.

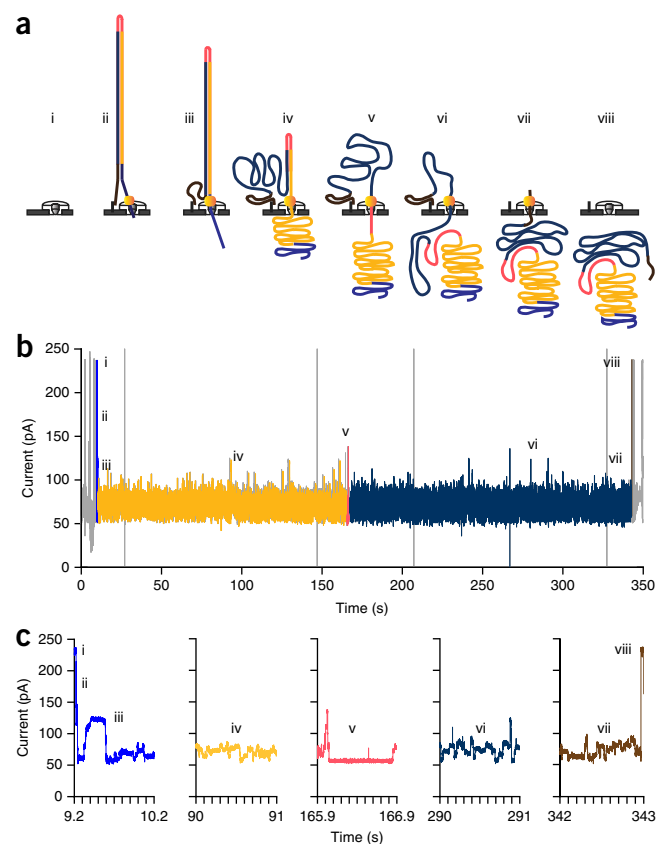
### MinION throughput

We sequenced intact replicative-form M13 phage dsDNA using three MinION flow cells that contained 337–473 functional channels (Online Methods). Reads were characterized as ‘template’, ‘complement’ or ‘2D’, with ‘2D’ representing reads obtained by computationally merging template and complement data from the same hairpin-linked molecule. Each 48-h replicate run generated between 184 million and 450 million bases from 63% template, 24% complement and 13% 2D reads (Supplementary Table 1). Results presented in this paper are based on reads classified by Metrichor as high quality, which totaled between 60 million and 189 million bases per M13 sequencing run.

### Establishing a mapping pipeline for MinION reads

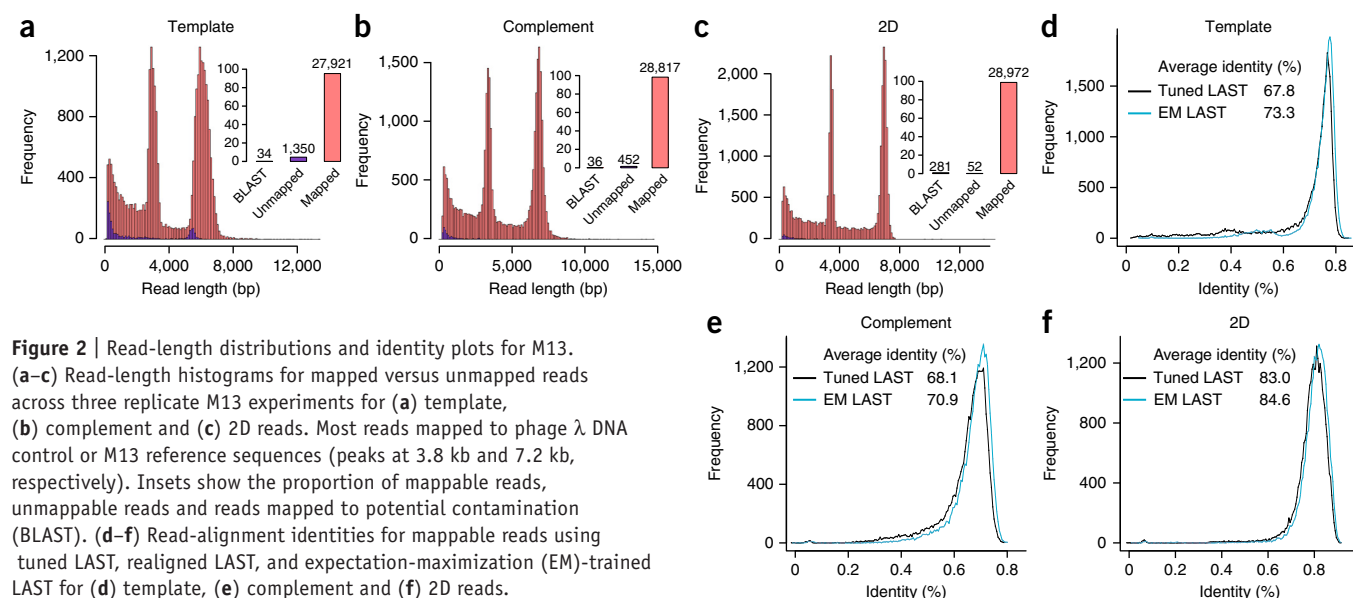
To evaluate the quality of these reads, we experimented with four different alignment programs<sup>1–4</sup> (Online Methods). Each was run with its default parameters and with tuned parameters that were selected on the basis of experimentation or expert advice from other participants in MAP (Supplementary Table 2).

The proportion of reads that mapped to reference sequences (M13 or ONT λ DNA control) varied by aligner (Supplementary Fig. 2). LAST<sup>3</sup> with tuned parameters was the most inclusive program, and stringency analysis indicated that few of its alignments were false

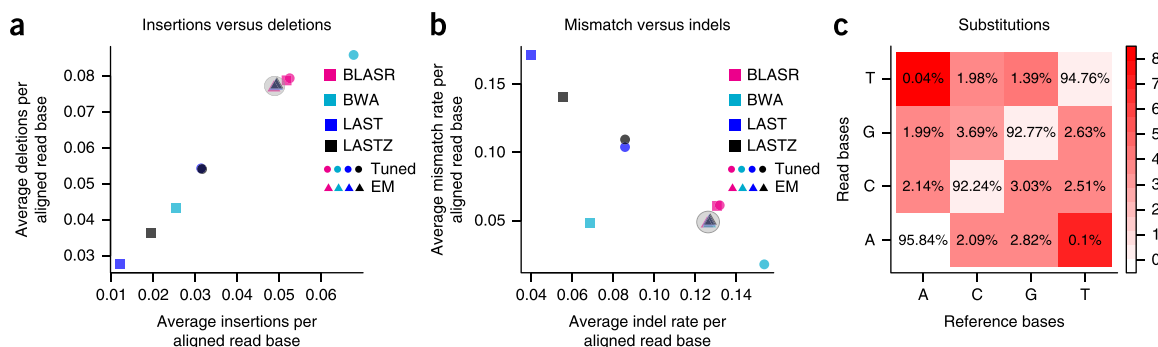


positives (Supplementary Fig. 3). For data pooled from the three M13 experiments, tuned LAST mapped 95.3% of template, 98.3% of complement and 89.9% of 2D reads. Most unmapped reads were homologous to *E. coli*, indicating minor contamination<sup>5,6</sup> (Online Methods, Fig. 2a–c and Supplementary Table 3).

We observed distinct peaks at 7.2 kb, corresponding to full-length M13 DNA, and at 3.8 kb, corresponding to the ONT λ phage DNA control (Fig. 2a–c). A large number of reads spanned the full M13 genome, whereas unmappable reads made up a small proportion (<0.2% of all 2D reads) and were generally shorter than mappable reads.



**Figure 2** | Read-length distributions and identity plots for M13. **(a–c)** Read-length histograms for mapped versus unmapped reads across three replicate M13 experiments for **(a)** template, **(b)** complement and **(c)** 2D reads. Most reads mapped to phage λ DNA control or M13 reference sequences (peaks at 3.8 kb and 7.2 kb, respectively). Insets show the proportion of mappable reads, unmappable reads and reads mapped to potential contamination (BLAST). **(d–f)** Read-alignment identities for mappable reads using tuned LAST, realigned LAST, and expectation-maximization (EM)-trained LAST for **(d)** template, **(e)** complement and **(f)** 2D reads.



**Figure 3** | Maximum-likelihood alignment parameters derived using expectation maximization (EM). The process starts with four guide alignments, each generated with a different mapper using tuned parameters. (a) Insertion versus deletion rates, expressed as events per aligned base. (b) Indel events per aligned base versus rate of mismatch per aligned base (Online Methods). Rates varied strongly between different guide alignments; however, EM training and realignment resulted in very similar rates (gray shading in circles), regardless of the initial guide alignment. (c) The matrix for substitution emissions determined using EM reveals very low rates of A-to-T and T-to-A substitutions. The color scheme is fitted on a log scale, and the substitution values are on an absolute scale.

### Expectation maximization generates high-confidence read alignments

We found substantial disagreement among rates of substitution, insertion and deletion for alignments generated by different mapping programs (Fig. 3a,b). A more principled way to estimate true error rates is to propose a reasonable model of the error process and calculate MLEs of the parameters (Online Methods)<sup>7</sup>. Using expectation maximization to train an HMM (Supplementary Fig. 4) and alignment-banding heuristics for efficiency<sup>8</sup>, we obtained robust convergence of parameter MLEs across all replicate experiments, guide alignments and random starting parameterizations (Fig. 3a,b and Supplementary Fig. 5). This showed that insertions were less frequent than deletions by about twofold in 2D reads and about threefold in template and complement reads. The combined insertion-deletion (indel) rate was between 0.13 (2D reads) and 0.2 (template and complement reads) events per aligned base. For all read types, indels were predominantly single bases (Supplementary Fig. 6). Substitutions varied from 0.21 (for template reads) to 0.05 (for 2D reads) events per aligned base (Fig. 3c and Supplementary Figs. 7 and 8). Substitution errors were not uniform; in particular, A-to-T and T-to-A errors were estimated to be very low, at 0.04% and 0.1%, respectively (Supplementary Note 1).

Realigning reads using the MLE parameters and the AMAP objective function<sup>9</sup> yielded substantial improvements over the initial alignments for every tuned program (Online Methods, Fig. 2d–f and Supplementary Fig. 9). For high-confidence alignments, there were no clear correlations between read length and errors

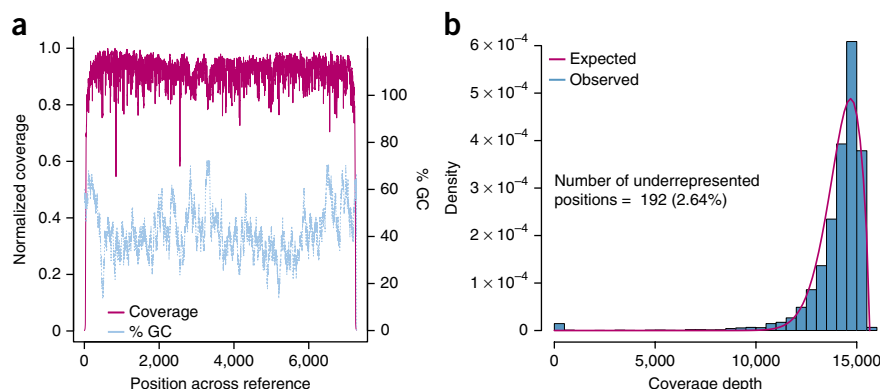
(Supplementary Fig. 10). However, there were positive correlations among the rates of insertions, deletions and substitutions in 2D reads (Supplementary Fig. 11 and Supplementary Note 2).

We also analyzed our data using a newly available Burrows-Wheeler Aligner (BWA) mode (ont2d) optimized for nanopore reads. The average percent identity obtained with ont2d was slightly less than the value obtained through expectation maximization (Supplementary Table 4); however, error rates were substantially closer to the MLE parameters estimated by expectation maximization, which suggests that ont2d is an improvement over the pacbio mode (for Pacific Biosciences) that we used originally.

To see whether our analysis pipeline produced similar results with larger, more complex genomes, we analyzed the *E. coli* data set released by Quick *et al.*<sup>10</sup>, which used R7.3 chemistry and Metrichor R7.3 2D version 1.5. The most recent Metrichor update was not available when Quick *et al.*<sup>10</sup> released their data set. We observed an improvement in average identity from 80.1% with tuned LAST to 81.8% after realignment using the AMAP objective function with MLE parameters. In addition, the MLEs for the rates of insertions (0.0598 events per aligned base), deletions (0.0910) and substitutions (0.0531) were very similar to those found for the M13 data.

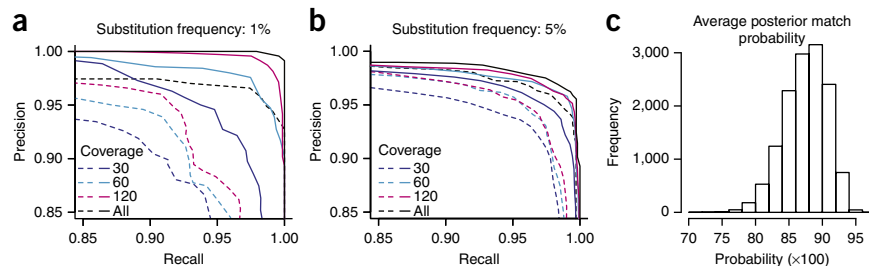
### M13 sequencing depth and *k*-mer analysis

Sequencing depth was generally consistent across the 7.2-kb M13 genome (Fig. 4 and Supplementary Fig. 12); however, 192 positions (2.6%) were underrepresented (Supplementary Note 3). Approximately 50% of these positions appeared at the beginning and end of the reference, and were likely the result of adaptor trimming by Metrichor. A majority of the remaining underrepresented positions were associated



**Figure 4** | M13 sequencing depth. (a) The magenta line denotes coverage by position in the genome (binned over a sliding 5-bp window), and the blue line depicts the local percentage of GC for that position (binned over a 50-bp sliding window). (b) Coverage-depth distribution fitted with a generalized extreme-value distribution.

**Figure 5** | Exploring SNV calling with MinION reads. (a,b) Variant calling with substitution frequencies of (a) 1% and (b) 5%. Dashed lines in both a and b represent results from variant calling using a transducer model conditioned on a fixed, tuned LAST alignment. Different sampled read coverages are shown. Each curve was produced by varying the posterior base-calling threshold to trade precision for recall. Solid lines in both a and b represent results from variant calling using the same transducer model as used for the tuned LAST alignments but incorporating marginalization over the read to reference alignments using a trained alignment model. Results shown are averaged over three replicate M13 experiments and, for each coverage level, three samplings of the reads. The “All” curve reflects all the available data for each experiment. (c) The distribution of posterior match probabilities shows that there was substantial uncertainty in most matches and demonstrates that marginalizing over the read alignments is a powerful approach.



with 5-mers rich in polymeric nucleotide runs (Supplementary Table 5). To determine whether the MinION has an inherent bias toward certain *k*-mers, we compared counts of 5-mers for all three read types (template, complement and 2D) with the M13 reference sequence. The most underrepresented 5-mers were homopolymers of poly(dA) or poly(dT), whereas the most overrepresented 5-mers were GC-rich and absent homopolymer repeats (Supplementary Note 3 and Supplementary Table 6). These findings are consistent with observations from Ashton *et al.*<sup>11</sup>.

### MinION reads can call SNVs with high recall and precision

SNV detection is important for metagenomics and microbial-strain detection<sup>12–14</sup>. To determine whether MinION reads could be used for SNV discovery in monoploid genomes, we computationally introduced random substitutions into the M13 reference sequence at 1%–20% frequency. Using this altered sequence as an alignment reference, we attempted to recover these substitutions using a Bayesian transducer framework<sup>15</sup> (Online Methods and Supplementary Note 4) and assessed performance in terms of precision, recall and *F*-score. These experiments also addressed the accuracy of our alignments and error models while avoiding

issues of reference-allele bias, to which simple metrics, such as alignment identity, are prone.

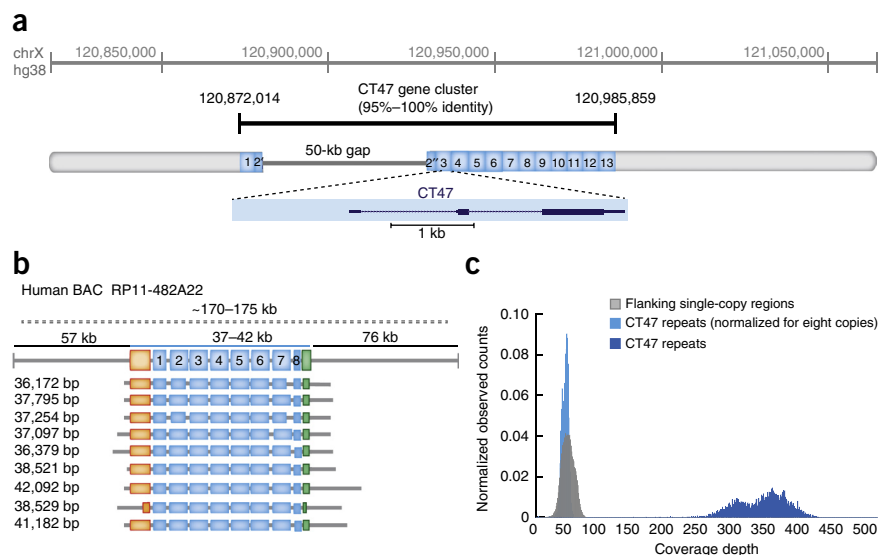
Using all the 2D read data and a posterior base-calling threshold that gave the optimal *F*-score, we achieved a recall of 99% and precision of 99% at 1% substitution frequency (Fig. 5a). When we reduced the sequencing depth down to a more reasonable 60× by sampling, we achieved recall and precision of 97%. Increasing the mutation frequency decreases the *F*-score progressively, presumably because alignment between the reads and the mutated reference becomes more difficult (Fig. 5b).

One particularly powerful strategy that we employed was marginalization over many possible alignments for each read, which helped factor out the considerable alignment uncertainty (Fig. 5c). In contrast, using fixed LAST alignments but otherwise keeping the method the same resulted in substantially higher rates of false positives for a given recall value (Fig. 5a,b).

### Resolving the organization of a cancer-testis gene family

A strength of the MinION device is its ability to produce long, single-molecule reads. In addition to routinely observing full-length 2D reads of M13 genomic DNA (Fig. 2), we found substantially

**Figure 6** | Resolution of CT47 repeat copy-number estimate on human chromosome Xq24. (a) BAC end sequence alignments (RP11-482A22: AQ630638 and AZ517599) span a 247-kb region, including 13 annotated CT47 genes<sup>16</sup> (each within a 4.8-kb tandem repeat), and a 50-kb scaffold gap in the GRCh38/hg38 reference assembly. (b) Nine MinION reads from high-molecular weight BAC DNA span the length of the CT47-repeat region, providing evidence for eight tandem copies of the repeat. Insert size estimated from pulse-field gel electrophoresis (dashed line) with flanking regions (black lines) and repeat region (blue line) are shown. Single-copy regions before and after the repeats are shown in orange (6.6 kb) and green (2.6 kb), respectively, along with repeat copies (blue) and read alignment in flanking regions (gray). The size of each read is shown to its left. (c) Shearing BAC DNA to increase sequence coverage provided copy-number estimates by read depth. All bases not included in the CT47 repeat unit are labeled as flanking regions (gray distribution; mean: 46.2-base coverage). Base coverage across the CT47 repeats was summarized over one copy of the repeat to provide an estimate of the combined number (dark blue distribution; mean: 329.3-base coverage) and was similar to single-copy estimates when normalized for eight copies (light blue distribution; mean: 41.15-base coverage).





longer reads, but at a lower frequency, when very large intact DNA fragments were delivered to the sequencer (for example, a full-length 48-kb 2D read of phage  $\lambda$  DNA mapped back to the reference with 87% identity (**Supplementary Fig. 13**)). We reasoned that long MinION reads, coupled with our high-confidence alignment strategy, could be used to resolve complex and often unfinished regions of genomes.

To test this, we examined the organization of a human-specific tandem-repeat cluster spanning a putative 50-kb assembly gap on human Xq24 (hg38 chrX:120,814,747–121,061,920) (**Fig. 6a**). Each 4,861-bp tandem repeat in this region contains a single annotated cancer-testis gene from the CT47 gene family with observed expression in testes, lung and esophageal cancer cells<sup>16</sup>. The high level of homology between adjacent copies (95%–100% sequence identity) is likely to result in recombination or replication errors, leading to alleles with different numbers of repeats that are often difficult to represent accurately by standard short-read assembly<sup>17</sup>. Furthermore, copy-number expansion and contraction involving genes contribute to variability in gene expression, epigenetic regulation and association with human disease<sup>18,19</sup>.

We used the MinION to acquire very long reads from a human BAC (RP11-482A22) that contained the CT47 repeats within the unresolved Xq24 segment. Nine 2D reads from 36 kb to 42 kb spanned all the repeats and together indicated eight tandem copies within the gap (Online Methods, **Fig. 6b** and **Supplementary Data**). This copy-number prediction was supported by pulse-field gel electrophoresis, which revealed a repeat array of 37–42 kb, or 7.5–8.6 copies of the 4.8-kb repeat (**Supplementary Fig. 14**). As an additional test, we obtained 40 $\times$ –60 $\times$  sequence coverage of the unresolved Xq24 segment using shorter ( $\sim$ 10 kb) MinION reads from sheared BAC DNA. A copy-number estimate based on these reads also indicated eight CT47 repeats within the unresolved region (**Fig. 6c**).

## DISCUSSION

We began this study by documenting MinION performance using M13 phage dsDNA. We found that consecutive reads of adaptor-linked template and complement DNA strands ( $\sim$ 14.4 kb total) were routinely achieved. Approximately 99% of 2D reads mapped to a reference (M13 or phage  $\lambda$  DNA control) and yielded 85% average identity. Using expectation-maximization training of an HMM, we were able to robustly parse the error sources into mismatches, insertions and deletions. This information was used to generate high-confidence alignments that allowed us to call SNVs accurately and characterize an unresolved region of human Xq24 rich in repetitive DNA. A dual-MinION sequencing strategy that employed both long-read scaffolds and higher-coverage shorter reads was essential for copy-number estimates in that region.

Comparisons with prior results<sup>11,20</sup> demonstrated improved read quality during MAP. We anticipate that the number of correct base calls will continue to increase beyond the average 85% identity observed in the current study. We also expect that the MinION will be used to report features of genomic DNA that are observable because the nanopore sensor directly touches each base on native DNA strands. These features include epigenetic modifications<sup>21–23</sup>, abasic residues<sup>24,25</sup>, DNA adducts<sup>26</sup>, thymine-thymine dimers and strand breaks.

In summary, we have shown that the MinION has sufficient accuracy to resolve important biological questions by sequencing long, native DNA strands. This accuracy is improving rapidly.

## METHODS

Methods and any associated references are available in the [online version of the paper](#).

**Accession codes.** ENA: [PRJEB8230](#), [ERP009289](#).

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

## ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Human Genome Research Institute of the US National Institutes of Health under award numbers HG006321 (M.A.), HG007827 (M.A.) and U54HG007990 (B.P.). The authors thank Oxford Nanopore Technologies for their gift to the UCSF Nanopore Group. The authors also thank D. Deamer for support, reading of the manuscript and helpful discussion. The authors gratefully acknowledge D. Haussler and J. Kent for their support.

## AUTHOR CONTRIBUTIONS

M.A. conceived experiments and directed research. B.P. conceived and directed bioinformatics analysis. B.P., M.J., I.T.F. and K.H.M. were responsible for bioinformatics analysis and software development. M.J. and H.E.O. were responsible for the completion of sequencing experiments and data processing. M.J. and H.E.O. were responsible for preparing DNA sequencing standards. H.E.O. was responsible for Sanger sequencing of M13 dsDNA. B.P. and I.T.F. were responsible for *k*-mer and BLAST analysis. B.P. and M.J. were responsible for SNV analysis. B.P. developed and implemented expectation-maximization and realignment strategies. K.H.M. conceived and directed BAC experiments and data analysis. All authors contributed to the writing, editing and completion of the manuscript.

## COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests. Details are available in the [online version of the paper](#).

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

1. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
2. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <http://arxiv.org/pdf/1303.3997.pdf> (2013).
3. Frith, M.C., Wan, R. & Horton, P. Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic Acids Res.* **38**, e100 (2010).
4. Harris, R.S. *Improved Pairwise Alignment of Genomic DNA*. PhD thesis, Pennsylvania State Univ. (2007).
5. Benson, D.A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
6. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
7. Do, C.B. & Batzoglou, S. What is the expectation maximization algorithm? *Nat. Biotechnol.* **26**, 897–899 (2008).
8. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18**, 1814–1828 (2008).
9. Schwartz, A.S. & Pachter, L. Multiple alignment by sequence annealing. *Bioinformatics* **23**, e24–e29 (2007).
10. Quick, J., Quinlan, A. & Loman, N. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *Gigascience* **3**, 22 (2014).
11. Ashton, P.M. *et al.* MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.* doi:10.1038/nbt.3103 (8 December 2014).
12. Davey, J.W. *et al.* Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510 (2011).

13. Bourlat, S.J. *et al.* Genomics in marine monitoring: new opportunities for assessing marine health status. *Mar. Pollut. Bull.* **74**, 19–31 (2013).
14. Stucki, D. & Gagneux, S. Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database. *Tuberculosis (Edinb.)* **93**, 30–39 (2013).
15. Holmes, I. & Bruno, W.J. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics* **17**, 803–820 (2001).
16. Chen, Y.T., Iseli, C. & Venditti, C. Identification of a new cancer/testis gene family, CT47, among expressed multicopy genes on the human X chromosome. *Genes Chromosomes Cancer* **45**, 392–400 (2006).
17. Treangen, T.J. & Salzberg, S.L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2012).
18. Tremblay, D.C., Alexander, G., Moseley, S. & Chadwick, B.P. Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. *BMC Genomics* **11**, 632 (2010).
19. Brahmachary, M. *et al.* Digital genotyping of macrosatellites and multicopy genes reveals novel biological functions associated with copy number variation of large tandem repeats. *PLoS Genet.* **10**, e1004418 (2014).
20. Mikheyev, A.S. & Tin, M.M. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* **14**, 1097–1102 (2014).
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. Laszlo, A.H. *et al.* Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proc. Natl. Acad. Sci. USA* **110**, 18904–18909 (2013).
23. Wescoe, Z.L., Schreiber, J. & Akeson, M. Nanopores discriminate among five C5-cytosine variants in DNA. *J. Am. Chem. Soc.* **136**, 16582–16587 (2014).
24. Cherf, G.M. *et al.* Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nat. Biotechnol.* **30**, 344–348 (2012).
25. Lieberman, K.R., Dahl, J.M., Mai, A.H., Akeson, M. & Wang, H. Dynamics of the translocation step measured in individual DNA polymerase complexes. *J. Am. Chem. Soc.* **134**, 18816–18823 (2012).
26. Schibel, A.E. *et al.* Nanopore detection of 8-oxo-7,8-dihydro-2'-deoxyguanosine in immobilized single-stranded DNA via adduct formation to the DNA damage site. *J. Am. Chem. Soc.* **132**, 17992–17995 (2010).

## ONLINE METHODS

**M13 MinION experiments.** We generated three replicate experiments with M13mp18 phage dsDNA to establish the reproducibility and performance characteristics of the MinION. Below we describe the M13 sequencing-standard preparation and MinION sequencing protocols.

**M13mp18 DNA sequencing standard.** M13mp18 dsDNA was obtained from New England Biolabs (NEB) (catalog no. N4018S). The host for this phage is *E. coli* strain ER2738, and the genome is 7.2 kb in size with a 42% average GC content. Thirty micrograms of M13mp18 was linearized by means of overnight double digestion with High-Fidelity HindIII (NEB, catalog no. R3104S) and High-Fidelity BamHI (NEB, catalog no. R3136S). Digests were performed according to NEB recommendations using Cut Smart Buffer supplied with restriction enzymes. Two hundred nanograms of M13mp18 double digest was run on a 1% Tris borate EDTA (TBE) agarose gel to confirm complete linearization of the circular replicative-form genome. The restriction digest was then extracted once with an equal volume of TE buffer (10 mM Tris, 1 mM EDTA, pH 8)-equilibrated phenol:chloroform (OmniPur, catalog no. 6805) and twice with TE buffer-equilibrated chloroform, pH 8, and then ethanol precipitated by the addition of 1/10 volume of 3 M sodium acetate, pH 5.2 (Teknova, catalog no. S0296), and 2 volumes of ice-cold 100% ethanol. Samples were centrifuged to pellet DNA, and the M13mp18 pellet was washed twice with 70% ethanol, allowed to dry to remove ethanol, resuspended in MilliQ water and quantitated using a Nanodrop. The M13 sequence was confirmed using Sanger sequencing (UC Berkeley DNA Sequencing Facility, with an ABI Model 3730 XL DNA Sequencer (Applied Biosystems, Life Technologies, Thermo Fisher Scientific)). Sequencing primers TAAGGTAATTCACAATGATTAAGTTG, CTGTGGAATGCTACAGGC, CACCTTTAATGAATAATTTCCGTC, CATGCTCGTAAATTAGGATGG, GTTTTACGTGCTAATAATTTTGATATG, CAAGGCCGATAGTTTGAGT, CACTGGCCGTCGTTTTA, GAGGCTTTATTGCTTAATTTTGC, AGGTCTTTACCCTGACTATTATAG, AGGCTTTGAGGACTAAAGAC, AATGGATCTTCATTAAAGCCAG, CAGCCTTTACAGAGAGAATAAC, TCCGCTTAGGTTGGG, GTGAGGCGGTCAGTATTAAC, GAGATAGGGTTGAGTGTGT and TTCTCCGTGGGAACAAAC were obtained from Integrated DNA Technologies (<http://www.idt.com/>).

**M13 MinION sequencing.** The libraries for MinION runs were prepared as recommended by ONT. Unsheared DNA was used for preparation of the M13 sequencing library. For BAC DNA, sequencing libraries were prepared using unsheared DNA as well as DNA sheared to an average length of 10 kb using g-TUBE (Covaris, catalog no. 520079). Briefly, the DNA sample was spiked with ONT  $\lambda$  DNA control, end-repaired using NEBNext End Repair Module (NEB, catalog no. E6050S) and cleaned up using Agencourt AMPure XP beads (Beckman Coulter, catalog no. A63880). The purified end-repaired DNA then underwent dA tailing with the NEB dA-Tailing Module (NEB, catalog no. E6053S). This was followed by ligation of ONT sequencing adaptors (adaptor Mix and HP adaptor) using Blunt/TA Ligase Master Mix (NEB, catalog no. M0367S). Using Dynabeads His-Tag Isolation and Pulldown (Life Technologies, catalog no. 10103D), we enriched the library for DNA molecules ligated to the ONT

HP adaptor. The adapted and enriched DNA was eluted in ONT-supplied elution buffer. This prepared library was then mixed with proprietary ONT EP Buffer and ONT Fuel Mix before being added to the MinION flow cell. Three 48-h sequencing runs were performed, each using a new flow cell.

The MinION data were base called using ONT Metrichor software (workflow R7.X 2D rev1.9). The base caller used classifies reads as pass or fail. Unless otherwise noted, all the analyses reported in this paper were performed using the 'pass' reads from R7.3 chemistry.

**Establishing a mapping strategy for MinION reads.** We experimented with four different initial read-mapping programs: BLASR<sup>1</sup> (PacBio's long-read mapper designed for mapping PacBio reads; commit abf9c38c55c2fb5f40316885dce39f5308c9ff25 from <https://github.com/PacificBiosciences/blasr>), BWA-MEM Release 0.7.11 (refs. 2,27) (H. Li's popular adaptation of the BWA mapper altered for handling long reads), LAST Version 490 (refs. 3,28) (a fast, sensitive, adaptable and popular pairwise-alignment tool) and LASTZ Release 1.02.00 (ref. 4) (a more traditional BLAST-type seed-and-extend program).

For each mapping experiment, reads were mapped both to the M13 reference sequence and to control DNA, a 3.8-kb segment of  $\lambda$  phage DNA supplied by ONT to be used in each experiment to measure baseline performance. For each mapping program, a sizable fraction of reads could not be aligned to either reference when the default parameters were used (data not shown). The use of tuned parameters substantially improved the number of reads mapped to the reference sequences.

To establish whether the mappers produced substantial numbers of false positive mappings, the reference sequences were reversed but not complemented, and the reads were mapped to these reversed sequences. The rationale for this experiment was that in the resulting reversed sequences, the base composition in terms of GC content and reversible Markov chain-like properties would be preserved, but it was highly unlikely that the sequences would be similar to the reads (**Supplementary Fig. 3**).

**BLAST analysis for unmapped reads.** In order to characterize the small minority of unmapped reads, we used BLAST 2.2.29 to align the unmapped reads to the NCBI Nucleotide database. The Nucleotide database contains entries from all of the traditional divisions of GenBank, the European Molecular Biology Laboratory and the DNA Data Bank of Japan<sup>5,6</sup>. The majority of unmapped 2D reads had BLAST hits (**Fig. 2** and **Supplementary Table 3**), most representing a low level of *E. coli* contamination.

**Learning the MinION error model.** The MinION error model we propose is a five-state pair HMM<sup>29</sup> that has two sets of insertion-deletion states (**Supplementary Fig. 4**), one set for modeling short insertions and deletions and one for modeling long insertions and deletions. The latter was included to account for large gaps at the beginnings and ends of the alignments—that is, to convert a local alignment model into a global alignment, as described by Durbin *et al.*<sup>29</sup>. To train the model we used a hybrid form of the Baum-Welch algorithm (a form of expectation maximization) that, for speed, works within an alignment band around a fixed guide alignment<sup>8</sup> for each read, with the guide alignments provided by a mapping program and the band constructed as

described by Paten *et al.*<sup>8</sup>, using C code adapted from the Cactus alignment program<sup>30</sup>. In contrast to alignment models learned from sequences related by evolution, no assumption of reversibility (and therefore symmetry) was made, and parameters for each transition and emission were learned independently.

We trained the alignment model for each possible combination of guide mapping program (tuned versions of the four mapping programs tested), MinION run (of three replicates) and read-type set (template, complement and 2D). For each training experiment we performed three independent runs, in each case starting from a randomly parameterized model and running for 100 iterations. **Supplementary Figure 5** shows the results of one training experiment, in which there is convergence of log-likelihood for all three runs to essentially the same value. **Supplementary Figure 5** also shows the resulting transition parameters for each read type; we observed excellent agreement in parameter estimates both between runs for the same training experiment and between training experiments with different MinION runs and different guide alignments, indicating that our parameter estimates were robust.

**Figure 3a,b** shows, as a cross-check, the calculation of insertion, deletion and substitution rates for 2D reads from realignments computed (see below) from each guide alignment using the alignment and the trained model. In each case, despite the fact that the starting guide alignments had very different estimates of these error rates, the realigned alignments gave consistently close error rates for these parameters. Interestingly, these values agreed relatively closely with the starting tuned-BLASR alignments, indicating it was most closely parameterized to our estimates of the maximum-likelihood rates.

**Realignment with a trained model.** For each possible combination of guide mapping program (tuned versions of BLASR, BWA-MEM, LAST and LASTZ; see **Supplementary Table 2**), MinION run (of three replicates) and read-type set (template, complement and 2D), we trained the alignment model and then realigned the reads using the resulting model. We call such alignments trained realignments. To realign the reads we used the aforementioned banding strategy around the guide alignment and picked a single alignment using the AMAP objective function<sup>9</sup>, which calculates an alignment that accounts for the posterior expectation of each match and indel. As a control experiment to account for the effects of realigning the reads, we also realigned the reads using the same guide-alignment strategy and objective function, but with an untrained model, the default HMM used by Cactus<sup>30</sup>, which was parameterized for vertebrate sequences related by natural selection. The control experiment showed that such alignments had substantially lower identity, indicating that the training, and not the process of realignment, was responsible for the improvement in identity (**Supplementary Fig. 9**).

**SNV calling with the MinION.** To determine how useful MinION reads are for simple SNV discovery in monoploid genomes, we took the M13mp18 reference sequence and randomly introduced substitutions at frequencies of 1%, 5%, 10% and 20%, picking the alternate allele with equal probability for each possible alternate base. We called each altered sequence a mutated reference sequence. For each read type for each replicate of the M13mp18 experiment, we aligned the reads to each mutated reference

sequence with a given mapper and ran an algorithm to call SNVs with respect to the mutated reference sequence.

Briefly, the SNV-calling algorithm (see **Supplementary Note 4** for a full description) has two steps: computing posterior alignment match probabilities between the bases in the reads and the reference, and calculating posterior base-calling probabilities for each reference base. By varying the threshold on the posterior base-calling probability, we traded precision for recall (**Fig. 5**). The reported precision and recall values were chosen to optimize the overall *F*-score.

The posterior match probabilities were computed using the guided-realignment strategy described above. The HMM used was composed by combining the described pair HMM (trained using expectation maximization on 2D reads with tuned LAST used as the guide alignment, as described earlier) with a substitution model that accounts for the introduced mismatches. Each model was described as a branch transducer<sup>15</sup>, and the models were combined to create an overall HMM, using the evolutionary HMM formalism<sup>15</sup>. The addition of the substitution model was found to be essential for high performance; **Supplementary Note 4** describes the parameters used and algorithm variations.

**Sequence scaffolding across the CT47 repeat cluster.** High-molecular weight BAC DNA (RP11-482A22) was isolated using standard methods for purification of large constructs (QIAGEN Large-Construct Kit, catalog no. 12462). To avoid DNA shearing for high-molecular weight sequencing, we performed NotI-HF (NEB, catalog no. R3189S) restriction digestion (expected to isolate the insert from pBACe3.6 cloning vector, gi|4878025) followed by end repair using Klenow in the same mix. This mixture underwent dA tailing directly after being added with separately end-repaired ONT-supplied control DNA, and the rest of the steps then proceeded according to the standard ONT recommendations, as mentioned above. The device was operated using ONT's MinKNOW software according to the provided instructions. The flow cells used were chemistry version R6.0 and R7.0. The read files were base called using ONT's Metrichor software, version 2D base calling, v1.2 and v1.3.1.

Long reads spanning the CT47-repeat cluster were identified using three sequence models<sup>31</sup>: a single-copy sequence directly upstream of the repeat array (6.6 kb, hg38 chrX:120865735-120872351), the CT47 repeat (4.8 kb, hg38 chrX:120932375-120937233) and a single-copy sequence directly downstream from the repeat array (2.7 kb, hg38 chrX:120986928-120989651). Reads were trimmed to the only present sequences involved in the repeat-classification models, and Pecan software<sup>8</sup> was used to generate multiple alignment of reads (data available in the European Nucleotide Archive; the primary accession number is PRJEB8230, and the secondary accession number is ERP009289).

**Copy-number estimates by sheared BAC sequencing.** To increase the MinION sequence throughput, we sheared RP11-482A22 BAC DNA to an average fragment length of 10 kb using g-TUBE (Covaris, catalog no. 520079). By alignment to the hg38 reference sequence (hg38 chrX:120,814,747–121,061,920, omitting a 50-kb scaffold gap), using tuned BLASR (as described above), we identified 2,006 2D reads that mapped to the RP11-482A22 DNA. Base coverage was determined from a sorted-alignment RP11-482A22 BAM file using bedtools genomecov<sup>32</sup> with the





command `bedtools genomecov -d -ibam mapping.sorted.bam`. Coverage estimates were converted to a BED file with each row entry defining coverage at a single base and at base + 1, and then they were subdivided into bases that overlapped with the CT47 repeat region and those that did not, with the latter labeled as flanking regions (bedtools intersect -woa and -v, respectively)<sup>32</sup>. A histogram of base coverage was generated to encompass all flanking bases and was determined to have a mean coverage value of 46.2 bases. Base-coverage estimates across the CT47 repeats were merged to represent a combined coverage depth over a single 4.8-kb repeat unit (mean observed base coverage: 329.3). Normalization of the read depth for eight copies of the repeat predicted an average read depth of 41 bases. We obtained the distribution of the normalized read depth by dividing by 8 across all base positions of the repeat with combined sequence depth.

**Pulse-field gel electrophoresis validation.** The RP11-482A22 BAC insert length estimate of NotI-HF-digested (NEB, catalog no. R3189S) or AatII-digested (NEB, catalog no. R0117S) DNA (1 µg) was determined by pulse-field gel electrophoresis (PFGE) using a CHEF-DRII system (Bio-Rad). Length estimates were determined using standard PFGE markers Low-range (NEB, catalog no. N0350S) and MidRange I (NEB, catalog no. NE551S).

Samples were run for 15 h (gradient, 6.0 V/cm; angle, 120°; switch time, linear; initial ramping, 0.2 s, finishing at 26 s) in 1% Pulsed Field Certified Agarose (Bio-Rad) and 0.5× TBE buffer at 4 °C. Banding was identified using standard SYBR Gold (Life Technologies) staining.

**Code availability.** The analysis software is open source and is available online (nanopore pipeline at <https://github.com/mitenjain/nanopore> and marginAlign pipeline at <https://github.com/benedictpaten/marginAlign>).

27. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM <https://github.com/lh3/bwa/blob/master/NEWS.md/#release-079-19-may-2014> (2014).
28. Frith, M.C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. *BMC Bioinformatics* **11**, 80 (2010).
29. Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. (The Press Syndicate of The University of Cambridge, 1998).
30. Paten, B. *et al.* Cactus: algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011).
31. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
32. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).