Table of Contents

1 MinION technology progression

Over the six-month period of MAP to date, there have been three MinION chemistry versions and numerous base-calling algorithm updates that have resulted in improvements in device performance (Supplemental Fig. 1). For example, at UCSC the average % identity (proportion of bases in a read aligned to a matching base in the reference) observed was at 67% in June 2014 (R6.0 release), 70% in July 2014 (R7.0 release), 78% in October 2014 (R7.3 release) and 85% in November 2014 (R7.3, high quality reads software filter).

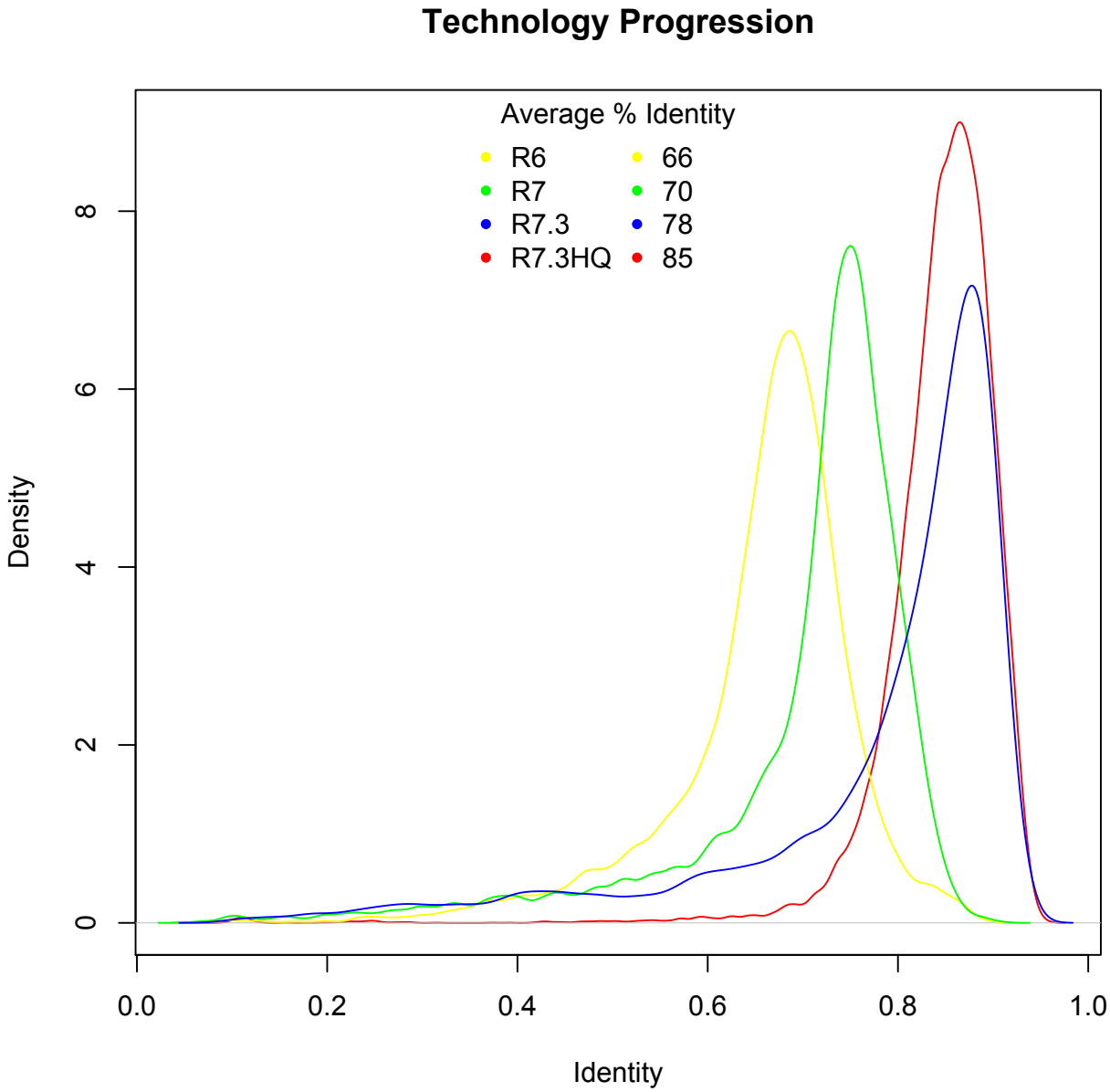


Fig. 1. Progression of read identities with MinION versions since June 2014.

2 M13 MinION Experiments

We generated three replicate experiments of M13mp18 bacteriophage DNA to establish the performance characteristics of the MinION. The throughput statistics are shown in Supplementary Note Table 1. The MinION read files were base called using Metrichor workflow R7.X 2D rev1.9. The basecaller (Metrichor) classifies reads as pass and fail. For simplicity, and to avoid doubling the exposition, all the analysis reported below, unless otherwise stated, was done using the pass reads from R7.3 chemistry.

Table 1. Number of functional channels and total amount of bases (in millions) generated as throughput from three M13 replicate experiments using R7.3 chemistry. Total throughput was obtained by adding the number of bases in the template and complement reads (from both *pass* and *fail* categories), and measures how many independent bases were read directly from the device during a run.

Experiment	Channels	*pass*			*fail*			Total
		Template	Complement	2D	Template	Complement	2D	
1	473	60	64	65	253	74	43	450
2	470	38	42	42	241	101	55	422
3	337	20	20	20	112	32	17	184

3 Establishing a read mapping strategy for MinION reads

To establish a methodology for mapping MinION reads we designed two pipelines that map the three read classes (template, complement and 2D) from the sequencing experiments described above. The nanopore pipeline (open source at `https://github.com/mitenjain/nanopore`) performs alignments, detailed analyses, and variant calling on the sequence data. Its can be used to recapitulate all the analysis in this manuscript. The marginAlign pipeline (open source at `https://github.com/benedictpaten/marginAlign`) is a lightweight, easy to install tool that performs alignment and variant calling. In the present study, FASTQ sequences were extracted from ONT base called files using custom scripts.

We experimented with four different initial read mapping programs: BLASR [1] (PacBio's long-read mapper designed for mapping PacBio reads, commit abf9c38c55c2fb5f 40316885dce39f5308c9ff25 from `https://github.com/PacificBiosciences/blasr`), BWA-MEM Release 0.7.11 [2,3] (Heng Li's popular adaptation of the BWA mapper altered for handling long-reads), LAST Version 490 [4,5] (A fast, sensitive, adaptable and popular pairwise alignment tool) and LASTZ Release 1.02.00 [6] (a more traditional BLAST type seed-and-extend program). Each mapping program was run with its default parameters, and, in addition, tuned

parameters that were determined either by experimentation, or by external expert advice, to perform well with MinION reads (see Supplementary Note 3.1 below).

For each mapping experiment reads were mapped both to the M13 reference sequence (see Methods) and the ONT lambda control DNA. The control DNA was a 3.8 kb segment of lambda phage DNA supplied by Oxford Nanopore 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 using the default parameters (data not shown). Use of tuned parameters substantially improved the number of reads mapping to the reference sequences. Supplementary Note Fig. 3 shows the overlaps in the number of reads mapping to either reference for the different mapping programs using tuned parameters; we found that tuned LAST mapped the vast majority of reads. In addition, very few reads (e.g. 2 2D reads) mapped by the other programs using tuned parameters are not also mapped by tuned LAST.

To establish if the mappers produced substantial numbers of false positive mappings the reference sequences were reversed but not complemented and the reads mapped to these reversed sequences. The rationale for this experiment being that in the resulting reversed sequences the base composition in terms of GC content and reversible Markov chain like properties are preserved, but the sequences are highly unlikely to be similar to the reads. Supplementary Note Fig. 4 shows the results, with tuned LASTZ producing a number of mappings (454) to the reversed reference, while LAST produced 106 and the other mappers produced no or negligible numbers of such (mis)mappings.

Having determined that tuned LAST mapped almost all the reads mapped by the other mapping programs, and produced very few false positives by our reversal assessment, in subsequent figures we present the results for tuned LAST, unless noted. During development we also ran the other mapping algorithms with both tuned and untuned parameters for all the other assessments and saw similar results to those presented.

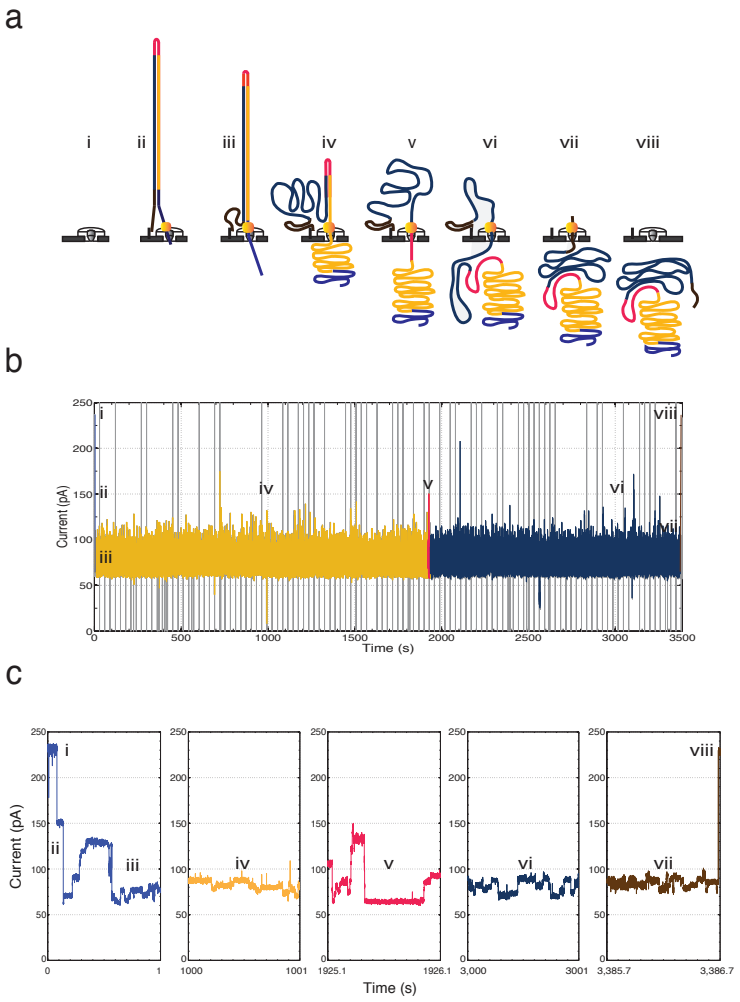


Fig. 2. Full Length (48kb) Lambda DNA Nanopore Data. (a) Molecular events for translocation of a single 48kb Lambda dsDNA molecule through the nanopore sequencer. DNA length and conformation are simplified for purposes of illustration. i - Open channel. ii - dsDNA with ligated loading (blue and brown) and hairpin adaptors (red) captured by the nanopore with the aid of a membrane anchor and an applied voltage across the membrane. iii -Translocation of the 5′ end of the loading adaptor through the nanopore under control of a molecular motor and driven by the applied potential across the membrane. DNA translocation through the nanopore starts. iv - Translocation of the template strand of DNA (gold). v - Translocation of the hairpin adaptor (red). vi - Translocation of the complement strand (blue). vii-Translocation of the 3′ portion of the loading adaptor. viii - Return to open channel nanopore. (b) Raw current trace for the entire passage of DNA construct through the nanopore (approximately 2789 seconds). Regions of the ionic current trace corresponding to steps i-viii are labeled. (c) Expanded 1 second time scale of raw current traces for DNA capture and translocation of 5′ loading adaptors (i-iii); template strand (iv); hairpin adaptor (v); complement strand (vi); 3′ loading adaptor, and return to open channel (vii-viii). Each adaptor generates a unique signal used for position reference in base determination.

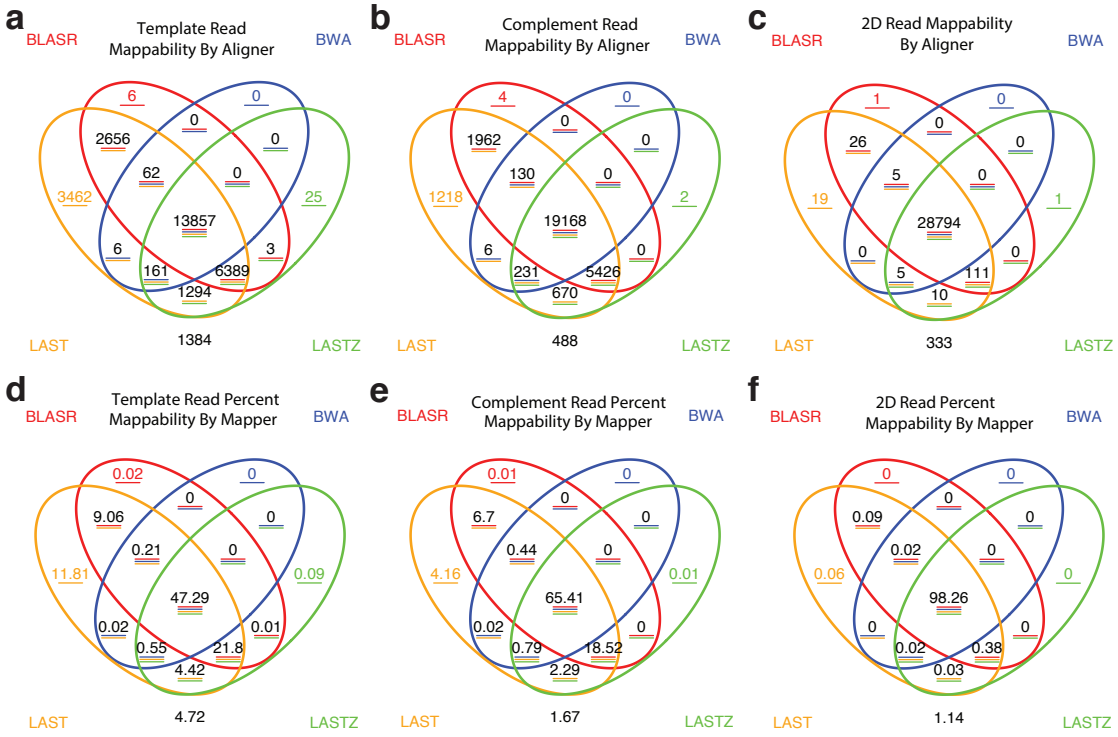


Fig. 3. Venn diagram representing read mappability for MinION reads across three replicate M13 experiments using R7.3 chemistry. Mappability represents the proportion of reads that can be aligned to either the M13 or phage lambda DNA using the tuned parameters for each mapper. In our analysis, 2D reads have the highest mappability, with 99% of reads being mappable, followed by complement and template reads at 98% and 95% of their respective read proportions being mappable. Among the four aligners used, LAST and LASTZ performed the best for M13, with LAST capturing the most proportion of mappable reads on its own.

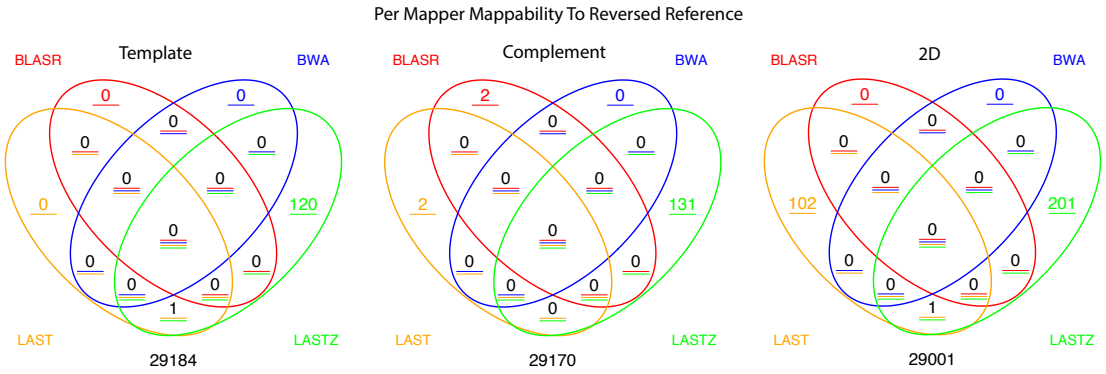


Fig. 4. Venn diagram representing read mappability to a reversed reference for MinION reads across three replicate M13 experiments using R7.3 chemistry. Results are using the tuned parameters. Since the reference was reversed, effectively no reads should map.

3.1 Mapping program parameters

In the figures and tables presented each mapper was run both with its default parameters and with the parameters described below, which are denoted 'tuned'. The parameters we used for tuning any of the mappers came mostly from recommendations from either Oxford Nanopore, fellow participants from MinION Access Program (MAP), and parameters tuned in house. The parameters we used for the tuned versions of each mapper are shown in Supplementary Note Table 2.

Table 2. Parameters used for different mappers and their sources.

Program	Parameters	Source/Recommendation
BLASR	-sdpTupleSize 8 -bestn 1 -m 0	MAP participants, tweaking at UCSC
BWA	-x pacbio	Heng Li for long reads
BWA	-x ont2d	Heng Li for MinION^{TM} long reads
LAST	-s 2 -T 0 -Q 0 -r 1 -a 1 -b 1 -q 1	Quick *et al*[7], MAP participants
LASTZ	–hspthresh=1800 –gap=100,100	Oxford Nanopore

4 *E. coli* contamination explains most unmapped 2D 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 NT database. The NT database contains entries from all of the traditional

divisions of GenBank, EMBL and DDBJ[8,9]. The majority of unmapped 2D reads had BLAST hits (See Fig. 2 in the main text and Supplementary Note Tables 3, 4 and 5), most representing a low level of Escherichia coli contamination. For the unmapped template and complement reads there were very few BLAST hits, but those that did map also mostly mapped to Escherichia coli family members.

Table 3: Table of BLAST hits for 2D reads unmapped by any mapper.

2D Unmapped Reads BLAST Hits	
Sequence Name	Counts
Escherichia coli KLY, complete genome	173
Escherichia coli B7A, complete genome	25
Escherichia coli O157:H7 str. EDL933, complete genome	11
Escherichia coli strain ST540, complete genome	7
Escherichia coli C321.deltaA, complete sequence	5
Escherichia coli UMNK88, complete genome	4
Escherichia coli str. K-12 substr. MC4100 complete genome	4
Escherichia coli str. K-12 substr. MG1655, complete genome	2
Escherichia coli LY180, complete genome	2
Escherichia coli plasmid pIS04_68, strain ISO4, complete sequence	2
Escherichia coli HS, complete genome	2
Escherichia coli P12b, complete genome	2
Escherichia coli E24377A, complete genome	2
Escherichia coli BL21(DE3), complete genome	2
Adenovirus type 2, complete genome	2
Human adenovirus C strain human/USA/Pitts_00109/1992/2[P2H2F2], complete genome	2
E. coli; the region from 81.5 to 84.5 minutes	2
Escherichia coli plasmid pH1038-142, complete sequence	1
Uncultured bacterium clone nbw890d10c1 16S ribosomal RNA gene, partial sequence	1
Homo sapiens chromosome 15, clone RP11-97H17, complete sequence	1
Escherichia coli SE15 DNA, complete genome	1
Homo sapiens 3 BAC RP11-208P4 (Roswell Park Cancer Institute Human BAC Library) complete sequence	1
Escherichia coli plasmid pH2291-144, complete sequence	1
Human alphoid repetitive DNA, subclone pHS53	1
Escherichia coli O145:H28 str. RM12581, complete genome	1
Escherichia coli DH1 (ME8569) DNA, complete genome	1

Homo sapiens 12 BAC RP11-478B9 (Roswell Park Cancer Institute Human BAC Library) complete sequence	1
Insertion sequence IS3 (from E.coli) inversion termini	1
Homo sapiens chromosome 18, clone RP11-210K20, complete sequence	1
Escherichia coli ABU 83972, complete genome	1
Homo sapiens 3-hydroxyisobutyryl-CoA hydrolase (HIBCH), RefSeqGene on chromosome 2	1
Escherichia coli O104:H4 str. 2009EL-2071 plasmid pAA-09EL71, complete sequence	1
Escherichia coli 042 complete genome	1
Escherichia coli strain ST2747, complete genome	1
Homo sapiens BAC clone CH17-417G10 from chromosome 1, complete sequence	1
Escherichia coli ATCC 8739, complete genome	1
Escherichia coli ETEC H10407, complete genome	1
Lactobacillus helveticus H9, complete genome	1
Salmonella enterica subsp. enterica serovar Typhimurium plasmid R64 DNA, complete sequence	1
Uncultured bacterium clone nck212c03c1 16S ribosomal RNA gene, partial sequence	1
Escherichia coli O157:H7 str. SS17, complete genome	1
Vibrio sp. 04Ya090 plasmid pAQU2 DNA, complete sequence	1
Shigella sonnei 53G main chromosome, complete genome	1
Achromobacter xylosoxidans A8, complete genome	1
Shigella boydii CDC 3083-94 plasmid pBS512_211, complete sequence	1
Homo sapiens 12 BAC RP11-693J15 (Roswell Park Cancer Institute Human BAC Library) complete sequence	1
Escherichia coli B7A plasmid pEB4, complete sequence	1
Shigella boydii CDC 3083-94, complete genome	1
Homo sapiens chromosome 15, clone RP11-483O19, complete sequence	1

Table 4: Table of BLAST hits for Complement reads unmapped by any mapper.

Complement Unmapped Reads BLAST Hits	
Sequence Name	Counts
Escherichia coli KLY, complete genome	15
Escherichia coli O157:H7 str. EDL933, complete genome	6
Escherichia coli C321.deltaA, complete sequence	2
Escherichia coli strain ST2747, complete genome	2
Escherichia coli B7A, complete genome	2
Escherichia coli 042 complete genome	1
Escherichia coli Trp repressor binding protein (wrbA) gene, complete cds	1
Escherichia coli W, complete genome	1
Escherichia coli 1540 plasmid pIP1206 complete genome	1
Escherichia coli O157:H7 str. EDL933 plasmid, complete sequence	1
Human adenovirus C strain DD28, complete genome	1
Escherichia coli strain D183 beta-lactamase TEM-1-like gene, partial sequence	1
Shigella dysenteriae strain 225-75 RNA polymerase subunit sigma-38-like (rpoS) gene, partial sequence	1
Enterobacter asburiae L1, complete genome	1

Table 5: Table of BLAST hits for Template reads unmapped by any mapper.

Template Unmapped Reads BLAST Hits	
Sequence Name	Counts
Escherichia coli KLY, complete genome	14
Escherichia coli B7A, complete genome	5
Escherichia coli O157:H7 str. EDL933, complete genome	2
Escherichia coli gene for hypothetical protein, partial cds, clone: pYU38	1
Shigella flexneri 2a str. 301, complete genome	1
Escherichia coli APEC O78, complete genome	1
Escherichia coli C321.deltaA, complete sequence	1
Escherichia coli W, complete genome	1
Enterobacteriaceae bacterium strain FGI 57, complete genome	1
Acidilobus saccharovorans 345-15, complete genome	1
Burkholderia cenocepacia MC0-3 chromosome 1, complete sequence	1
Uncultured bacterium clone PL06G10 16S ribosomal RNA gene, partial sequence	1
Uncultured soil bacterium clone GO0VNXF07H12HG 16S ribosomal RNA gene, partial sequence	1
Rattus norvegicus 8 BAC CH230-416D7 (Children's Hospital Oakland Research Institute Rat (BN/SsNHsd/MCW) BAC library) complete sequence	1
Shigella flexneri 5 str. 8401, complete genome	1
Shigella dysenteriae Sd197, complete genome	1

5 Analyzing Read Length

Read length distributions for mapped vs. unmapped reads across three replicate M13 experiments using R7.3 chemistry for template, complement, and 2D reads are shown in Fig. 2(a-c).

5.0.1 Most mapped 2D reads span the full length of the M13 genome. We observed two distinct peaks for 2D reads, one at about 7.2 kb, corresponding to full-length M13, and one at 3.8 kb, corresponding to ONT lambda phage DNA control. The very small proportion of unmappable 2D reads (<0.2%) were generally shorter than the mappable reads.

6 Learning the MinION error model

Counting the number of substitutions, insertions and deletions in alignments we found substantial disagreement in the rates of these errors between different mapping programs and parameter variations (Fig. 3 A-B). A more principled way to estimate the true rates of these errors is to propose a model of the error process and calculate maximum likelihood estimates of the parameters of the model.

The model we propose is a five state pair-HMM [11] which has two sets of insertion/deletion states (Supplementary Note Fig. 5), one set for modeling short insertions/deletions and one for modeling long insertions/deletions.

The latter were included to account for large gaps at the beginning and ends of the alignments, i.e. to convert a local alignment model into a global alignment, as described in Durbin et al. [11]. 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 [12] for each read, the guide alignments being provided by a mapping program, and the band being constructed as described in Paten et al.. [12], using C code adapted from the Cactus alignment program [13]. In contrast to learning an alignment model from sequences related by evolution, no assumption of reversibility (and therefore symmetry) was made, and parameters for each transition and emission were learnt independently.

For each possible combination of guide mapping program (tuned versions of BLASR, BWA-MEM, LAST and LASTZ, see Supplementary Note 3.1), MinION run (of three replicates) and read type set (template, complement and 2D) we trained the alignment model. For each training experiment we performed three independent runs, in each case starting from a randomly parameterized model and running for 100 iterations. Supplementary Note Fig. 6 shows the results for one training experiment, showing convergence of log-likelihood for all three runs to essentially the same value. Supplementary Note Fig. 5 shows the resulting transition parameters for each read type; for each read type we observe 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 are robust.

Fig. 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 starting guide alignments having very different estimates of these error rates the realigned alignments give consistently close error rates for these parameters. Interestingly, these relatively closely agree with the starting tuned BLASR alignments, indicating it was most closely parameterized to our estimates of the maximum likelihood rates.

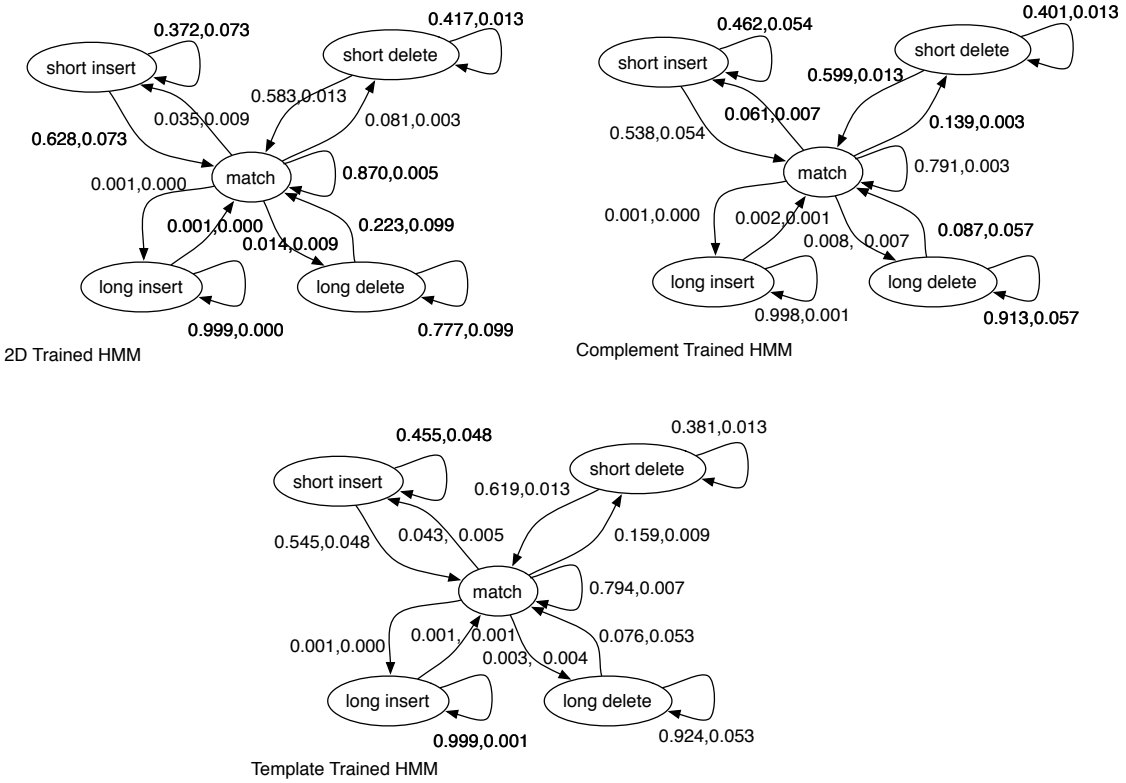


Fig. 5. Structure for the Hidden Markov Model (HMM) used for EM, along with the estimated parameters for transition probabilities for template, complement, 2D reads. For each transition in order the mean estimate and standard error across all experiments for that read type are shown.

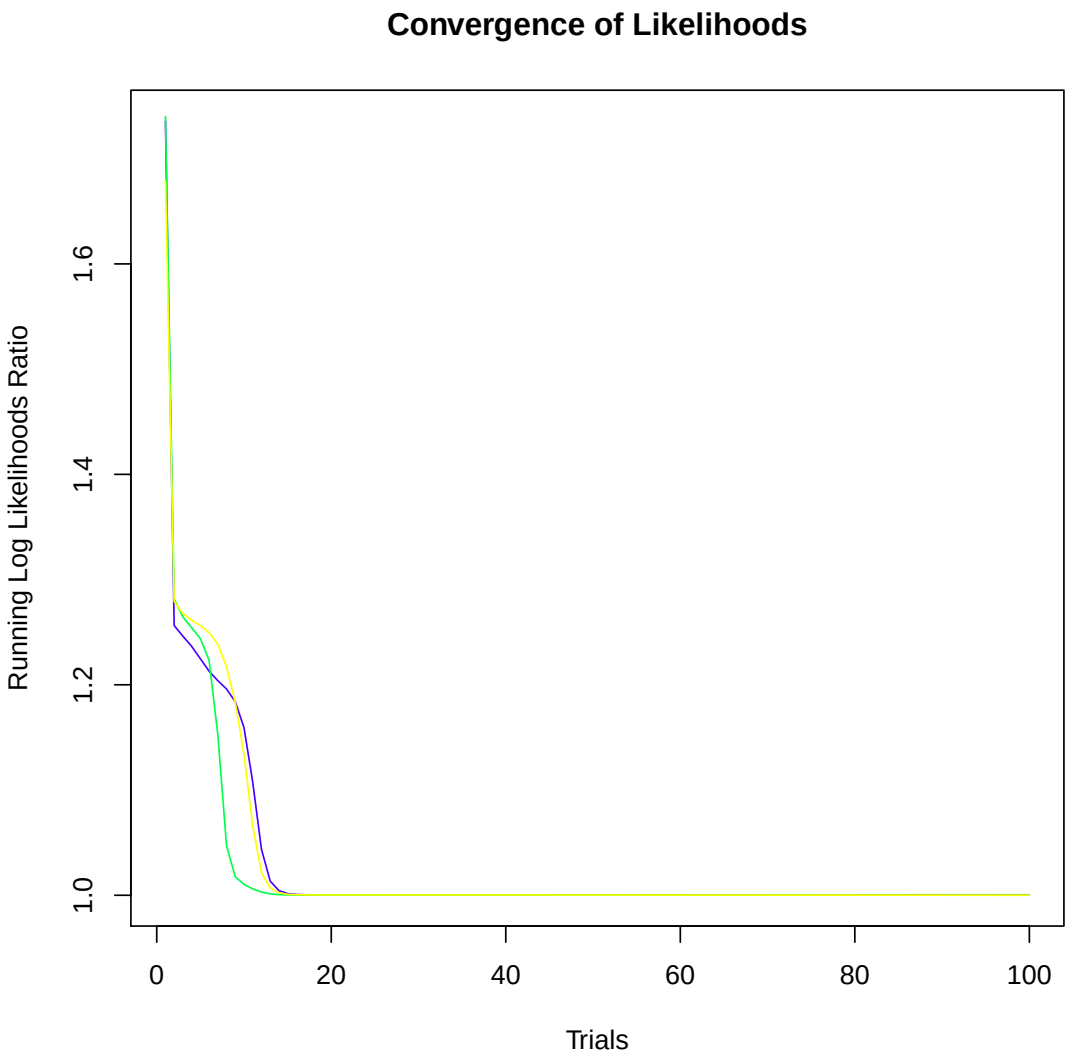


Fig. 6. Convergence of log-likelihood for three independent runs of expectation-maximization, each from a randomly parameterised model, each run for 100 iterations of training. The y-axis gives likelihood normalized by the highest log-likelihood found. The training used 2D reads from one MinION run of the M13 data using release R7.3 chemistry and a guide alignment generated by tuned BLASR.

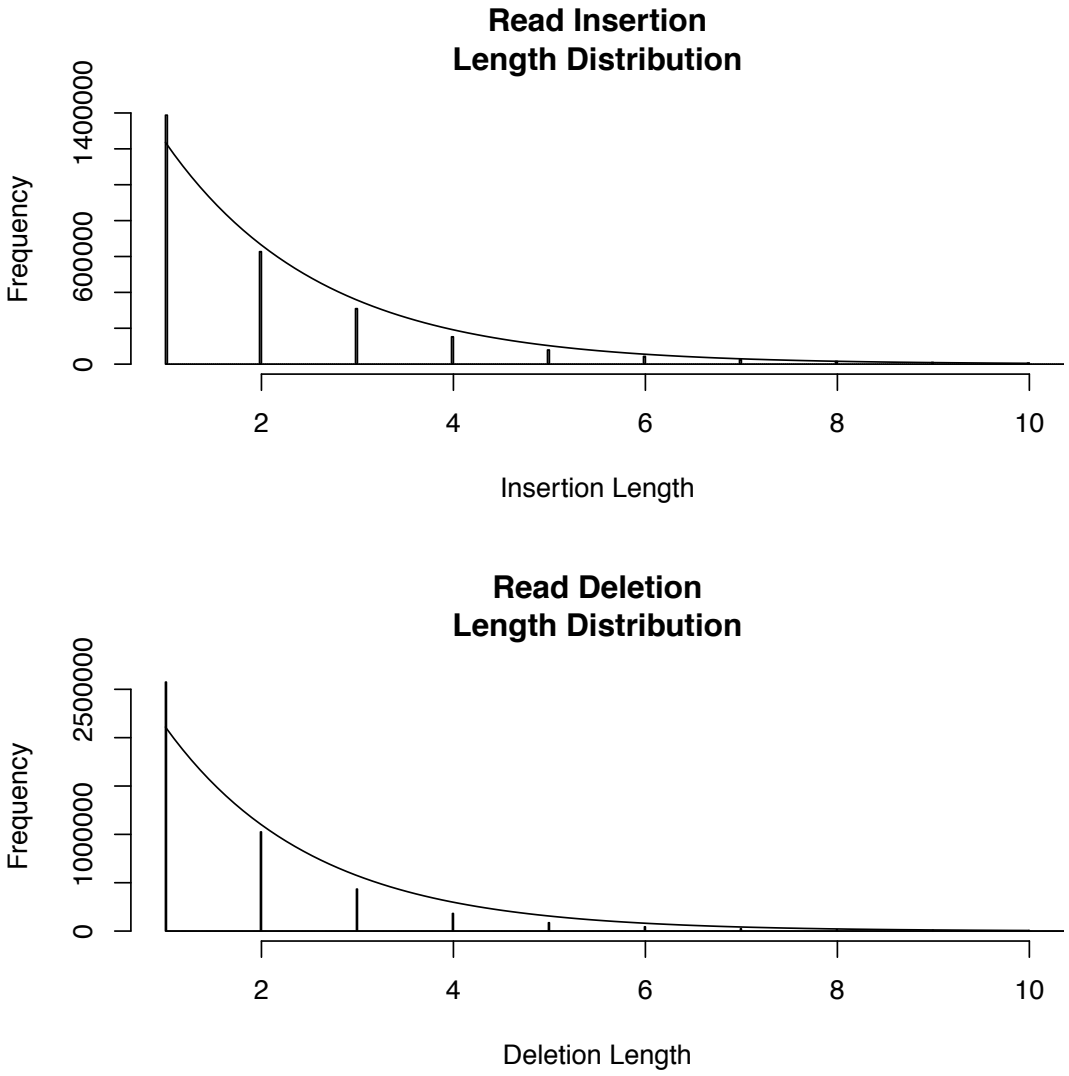


Fig. 7. Representative insertion and deletion plot for one M13 experiment using R7.3 chemistry, and aligned using LAST post-EM.

Recently, we performed alignments using the new BWA release (version 0.7.12) that includes the ont2d mode for nanopore reads (commit 8211fbcb625bef6480d04fa196e7514cbb31eb84 from `https://github.com/lh3/bwa/`). The rate of insertions, deletions, and substitutions for BWA (pacbio and ont2d modes) and EM-based LAST are shown in Supplementary Note Table 6. The average % identity decreased from 85% for BWA pacbio mode to 83% for BWA ont2d mode. However, the error rates for BWA ont2d are now closer (though still a distance from) our MLE estimates. Using EM and our realignment strategy we observed convergence

between different starting alignments. We expect this also to be true if starting from a BWA alignment using the new ont2d mode (Figure 3a-b).

Table 6. Error rates obtained using tuned BWA (*pacbio* and *ont2d* modes), and EM-based LAST.

Program	Parameters	*Rate (%)*			Average % Identity
		Insertions	Deletions	Substitutions	
BWA	-x pacbio	6.8	8.6	1.8	85
BWA	-x ont2d	3.1	5.4	10.4	83
LAST	EM	4.9	7.8	5.1	85

6.1 Adenosine to thymine and thymine to adenosine substitution errors are rare in MinION reads

Fig. 3c and Supplementary Note Fig. 8 shows the trained estimates of the substitution parameters of the model, for each of the read types. Surprisingly the proportion of adenosine to thymine errors was estimated to be very low, and similarly, but slightly less strongly, the proportion of thymine to adenosine errors was also estimated to be low. To check that these rather striking results were not training artifacts we calculated estimates of the substitutions directly from alignments produced by the different mapping programs (Supplementary Note Fig. 9), in each case seeing the same trend. To ascertain if the very low substitution error rates were influencing the transition parameters during training (e.g. certain substitutions being traded for higher rates of insertions/deletions, Supplementary Note Fig. 7), we tied the emission parameters during training so that substitutions occurred at the same rate regardless of the bases involved, and so that indel emissions were flat (the same for each base regardless of type). The resulting trained HMMs had virtually the same transition parameters as the untied models (data not shown), suggesting that the trained transition parameters were not biased by the asymmetries of the trained emission parameters. Though more data on a diversity of different sequencing samples was needed to confirm these results, we note that mapping results could probably be improved by taking into account these bias in substitution errors when considering seed alignments (e.g. discounting seed matches with numerous adenosine to thymine matches).

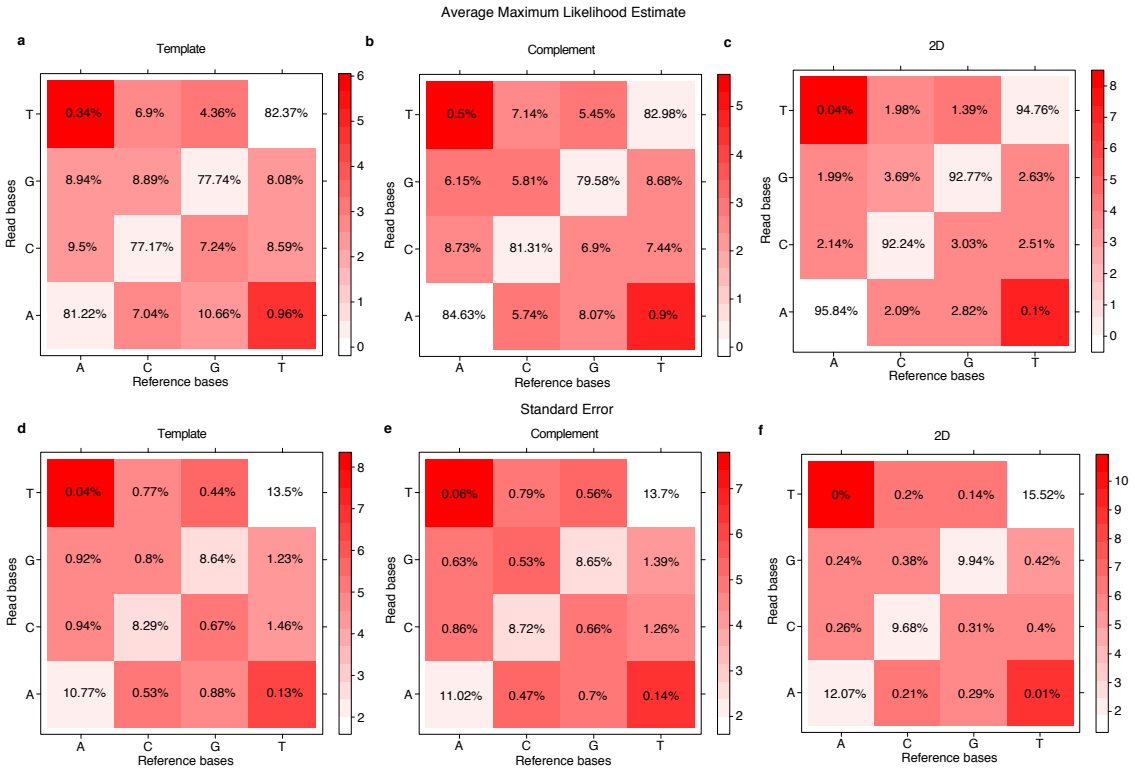


Fig. 8. Maximum-likelihood estimates and standard error parameters for substitution matrices show trends across template, complement, and 2D reads across three M13 experiments using R7.3 chemistry. The top panel illustrates the average maximum likelihood estimate for these substitutions, with the standard error represented in the lower panel.

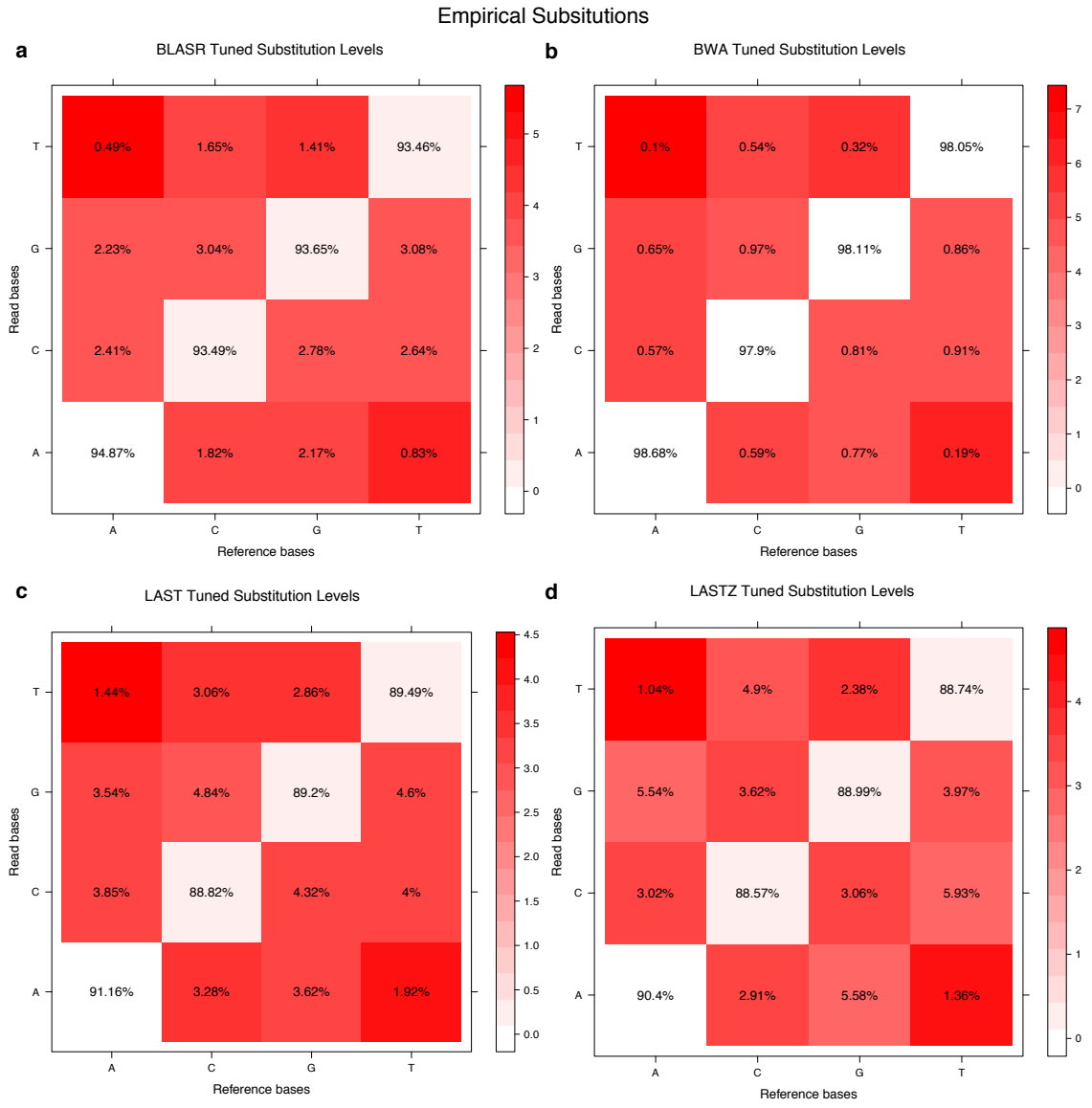


Fig. 9. Substitution matrices representing for each of the four tuned aligners across three M13 experiments using R7.3 chemistry. For all the aligners, thymine to adenosine and adenosine to thymine substitution rates are low, indicating that the device rarely miscalls one as the other.

7 Read alignment identity was increased by realigning reads with a trained model

We define the *identity* of a read alignment as the proportion of read bases aligned to reference bases without mismatches. We realigned the reads using trained models to see if this altered the identity of the

read alignments. For each possible combination of guide mapping program (tuned versions of BLASR, BWA-MEM, LAST and LASTZ, see Supplementary Note 3.1), 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 same banding strategy around the guide alignment, and picked a single alignment using the AMAP objective function[14], 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 using an untrained model, the default HMM used by Cactus, which was parametrized for vertebrate sequences related by natural selection. We call these alignments *naive realignments*.

Supplementary Note Fig. 10 and Fig. 2d-f show the resulting distribution of alignment identity, aggregated across replicates for the LAST trained realignments. The trained LAST realignments, but not the naive realignments, show a substantial boost in identity (see Fig. 2d-f) over the tuned LAST alignments. This was evident for all other guide mappers (data not shown).

8 Errors in mappable reads are not clearly correlated with read length

We compared read lengths of mappable reads across all three read types to common alignment metrics - mismatches, insertions, deletions, and identity (Supplementary Note Fig. 11 shows results for 2D reads, other read types were similar). Though the patterns are complex, partly because of the two different reference sequences (M13 and Lambda control DNA), there are no clear overall linear correlations between read length and any given mutation frequency.

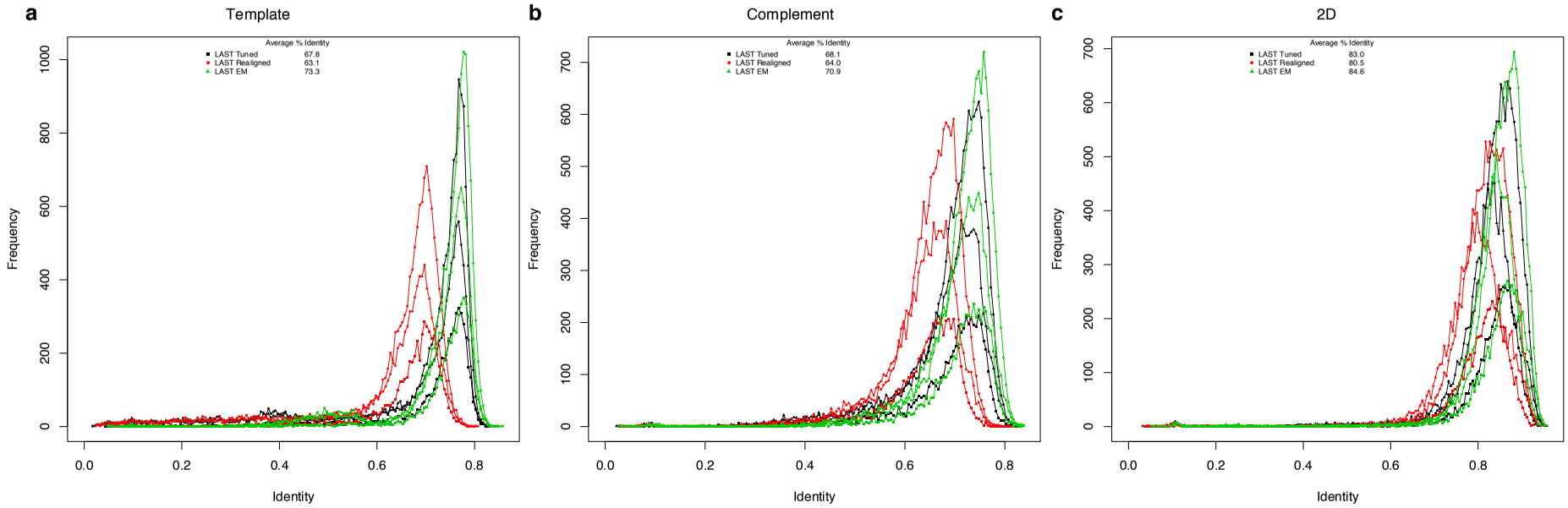


Fig. 10. Read identity for template, complement, and 2D reads for three M13 replicate experiments using R7.3 chemistry, aligned using LAST. Three versions of the LAST alignment are shown: tuned LAST, trained LAST realignments and naive LAST realignments.

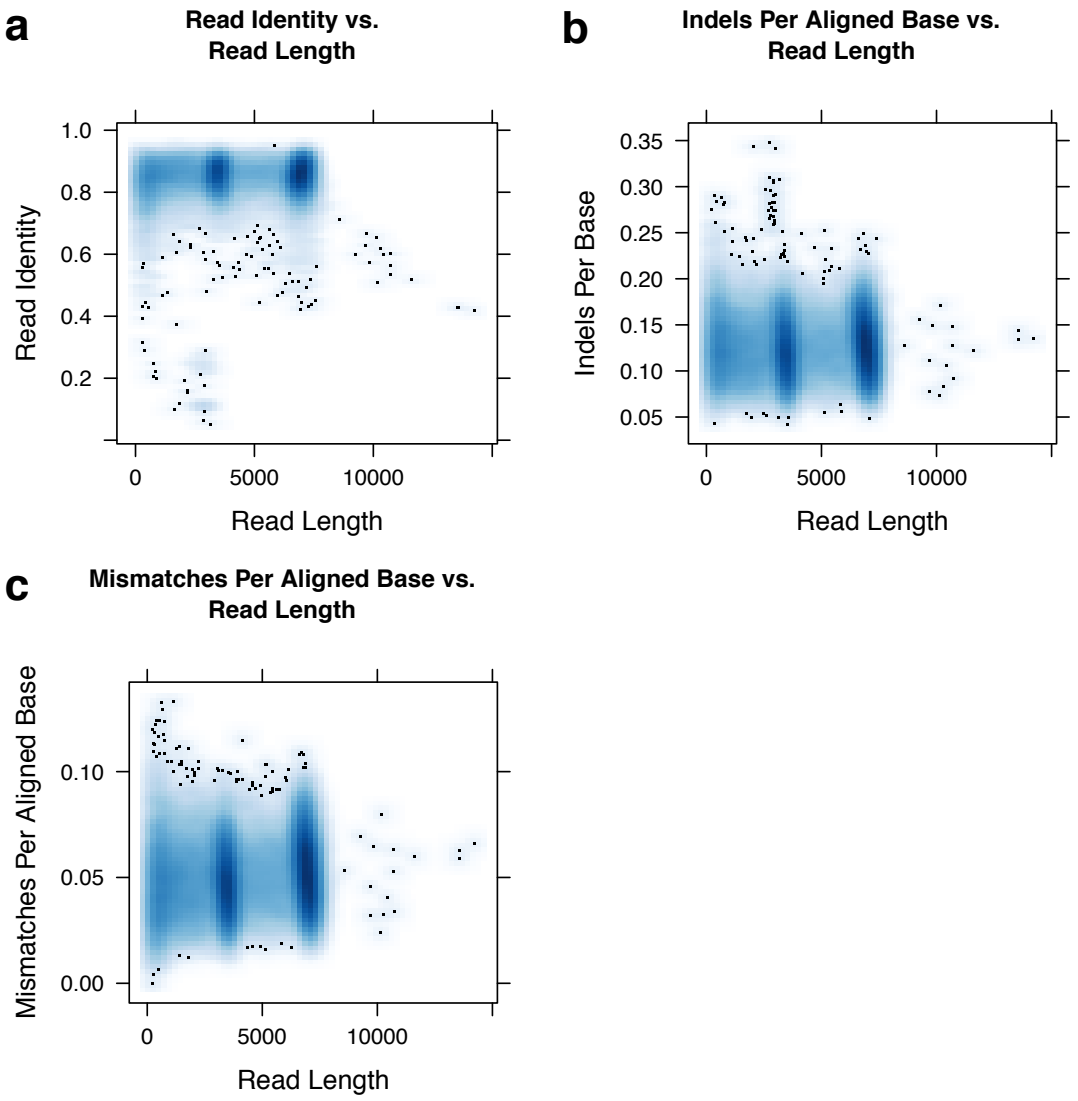


Fig. 11. Alignment quality measurements for 2D reads across three M13 replicate experiments. Alignments were obtained using trained LAST realignments.

9 Insertion, deletion and substitution errors correlate in 2D reads

We compared rates of insertion, deletion and mismatch against each other for all three replicates of M13 (Supplementary Note Fig. 12). For 2D reads, we found a correlation between the rate of mismatches and indels, $R^2 = 0.735$, and a suggestive correlation between the rates of insertions and deletions, $R^2 = 0.387$. Looking at the template and complement reads we did not find any such correlation (data not shown). One plausible hypothesis to explain the apparent correlation was that error rates for 2D reads were dictated by

the ratio of the lengths of its constituent template and complement reads. E.g. if there was a full template read but the complement read was short, much of the 2D read would be inferred only from the template read, without the benefit of having a full second observation of the read sequence. We did not find a convincing correlation between read identity for 2D reads and the number of segments in their respective template and complement reads (data not shown). Using R7.3 chemistry with older versions of Metrichor (R7.3 2D Version 1.5), Quick *et al.* observed a correlation between read identity for 2D reads and the number of segments in the template and complement reads[7].

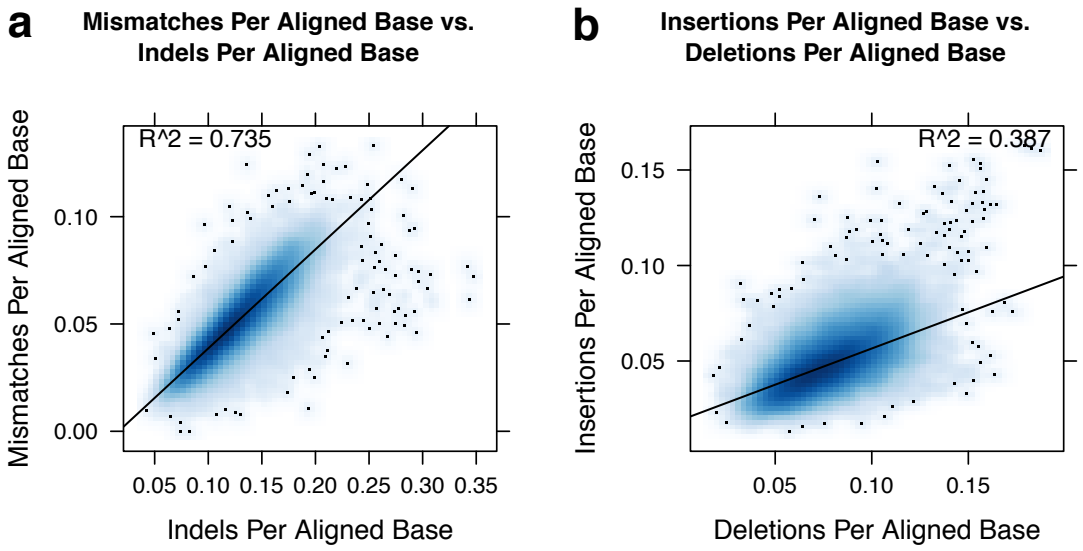


Fig. 12. Error profile analysis of 2D reads aligned using trained LAST realignments indicates a moderate correlation between mismatches and indels per aligned base, and a weak correlation between insertions per aligned base and deletions per aligned base.

10 Pipeline validation using *E. coli* data released by Quick *et al.*[7]

To assess if the analysis pipeline we designed would be suitable for larger, more complex genomes, we analyzed *E. coli* data released by Quick *et al.* that was obtained using R7.3 chemistry and Metrichor R7.3 2D Version 1.5. The most recent Metrichor update was not available to Quick *et al.* at the time of their data release. We analyzed full 2D reads, as defined by Quick *et al.*, and observed an improvement in average % identities with realignment. The results for this analysis are shown in Supplementary Note Table 7. The improvement in identity with EM demonstrates that the results were not specific to M13. Also, the MLEs for mismatches (0.0531 events/aligned base), insertions (0.0598 events/aligned base) and deletions (0.091

events/aligned base) were remarkably close to those found for the M13 data, suggesting that the errors were largely invariant to the source genome.

Table 7. Analyzing previously released *E. coli* data[7] with the UCSC analysis pipeline. Data obtained using R7.3 chemistry and Metrichor™ R7.3 2D Version 1.5.

Substrate	Average % identity	
	LAST Tuned	LAST EM
E. coli Full 2D	80.1	81.8
M13 *Full 2D*	70.0	80.7

11 Assessing MinION read coverage

We measured sequencing depth, termed coverage, across the M13mp18 reference. The coverage for template/complement/2D reads across three replicate experiments is shown in Supplementary Note Fig. 13a-c respectively. For all three read types coverage was largely consistent across the genome, apart from at the very ends of the genome (see below), and did not appear to fluctuate substantially based upon GC content - though the short length and relatively narrow fluctuation in GC across the M13mp18 genome precludes a thorough assessment of this issue.

Fitting a generalized extreme value distribution[15] (Supplementary Note Fig. 13d-f) to the 2D read coverage we identified 192 sites (2.6%) across M13 genome as under-represented using non-parametric statistical analysis. Briefly, we selected outliers based on positions where the observed coverage deviated beyond 2 standard deviations. We found the under-represented sites to be divisible into subsets. The first 49 and the last 43 nucleotides of the M13 reference were under-represented; we hypothesize these under-represented sites are the result of adaptor trimming by the base-calling software. A close examination of 5-mers overlapping the remaining 100 positions (four preceding nucleotides along with the nucleotide at the position of interest) revealed these sites to be rich in homopolymeric nucleotide runs (Supplementary Note Table 8).

Table 8. 5-mers observed at the 100 underrepresented positions in the M13 genome. These numbers do not consider positions at the beginning and end of M13 which are likely to be under-represented as a result of adaptor trimming.

K-mer	# Positions	K-mer	# Positions	K-mer	# Positions
AAAAA	13	CCTCT	1	GTCTA	1
AAAAC	1	CCTTT	1	GTTTT	2
AAAAG	1	CGCCC	1	TAAAA	2
AAAAT	1	CGTCA	1	TACAA	1
AAACA	1	CTGGT	1	TACAC	1
AAATT	1	CTTTC	1	TACAT	1
AAGTG	1	CTTTT	5	TAGAT	1
AATCG	1	GAGCC	1	TAGTG	2
ACTCT	1	GAGGA	1	TATAT	1
AGCCT	1	GCAAC	1	TGAAG	1
AGGCT	1	GCCAC	1	TGACC	1
AGTTA	1	GCCCT	2	TGCTA	1
ATTCA	1	GCCTT	1	TGTAC	1
ATTTG	1	GGGAT	1	TTATA	1
ATTTT	1	GGGGG	1	TTCAT	1
CAAAA	5	GGGTG	1	TTCGC	1
CAGCT	1	GGTAC	1	TTTCA	1
CCACC	2	GGTAT	1	TTTGA	1
CCCCA	1	GGTGA	1	TTTTA	2
CCCCC	1	GGTTA	1	TTTTT	13
CCCTA	1	GTAAC	1		

11.1 Homopolymer containing k-mers are under-represented in MinION reads

Coverage drops at homopolymeric sites was not unexpected because nanopore sequencers do not read individual bases, rather they measure a continuous change in current, with 5 bases within the pore at any time. To resolve this into a sequence of individual nucleotides, the base calling algorithm integrates the signal over 5-mer windows. To test whether any of the possible 1024 5-mers were under- or overrepresented we evaluated relative enrichment patterns in the M13 sequence datasets. We employed a sliding window analysis (spanning 5 bases with a slide of 1 base) to determine the frequency of all possible 5-mers in both forward and reverse complement orientation within both datasets. Briefly, enrichment/depletion significance was tested

through simulation. 5-mers were drawn 5,000 times across 1,000 replicates from the distributions counted from the data and then the Kolmogorov-Smirov test was used to compare these distributions, assigning a Bonferroni-corrected p-value to each comparison (not shown). Consistent with the observed coverage drops, the most under-represented 5-mers in the read set contain poly-dA or poly-dT, while the most enriched 5-mers are G/C rich and did not contain homopolymer repeats (Supplementary Note Table 9).

We also compared 5-mers spanning indels in alignments. For this experiment, indels were defined as any 5-mer which has an alignment gap of any size in the four internal positions. We found similar trends in these 5-mers as in the overall counts, with poly-dA and poly-dT 5-mers being under-represented in the read set. The similarity of these two comparisons was not surprising given the interspersed and highly common nature of 1-2 bp indels in these alignments (Supplementary Note Table 10).

In both comparisons, no systematic difference was seen between template, complement and 2D reads. Individual comparisons have different ordering of enriched and depleted 5-mers, but similar trends are found across each read type within each comparison.

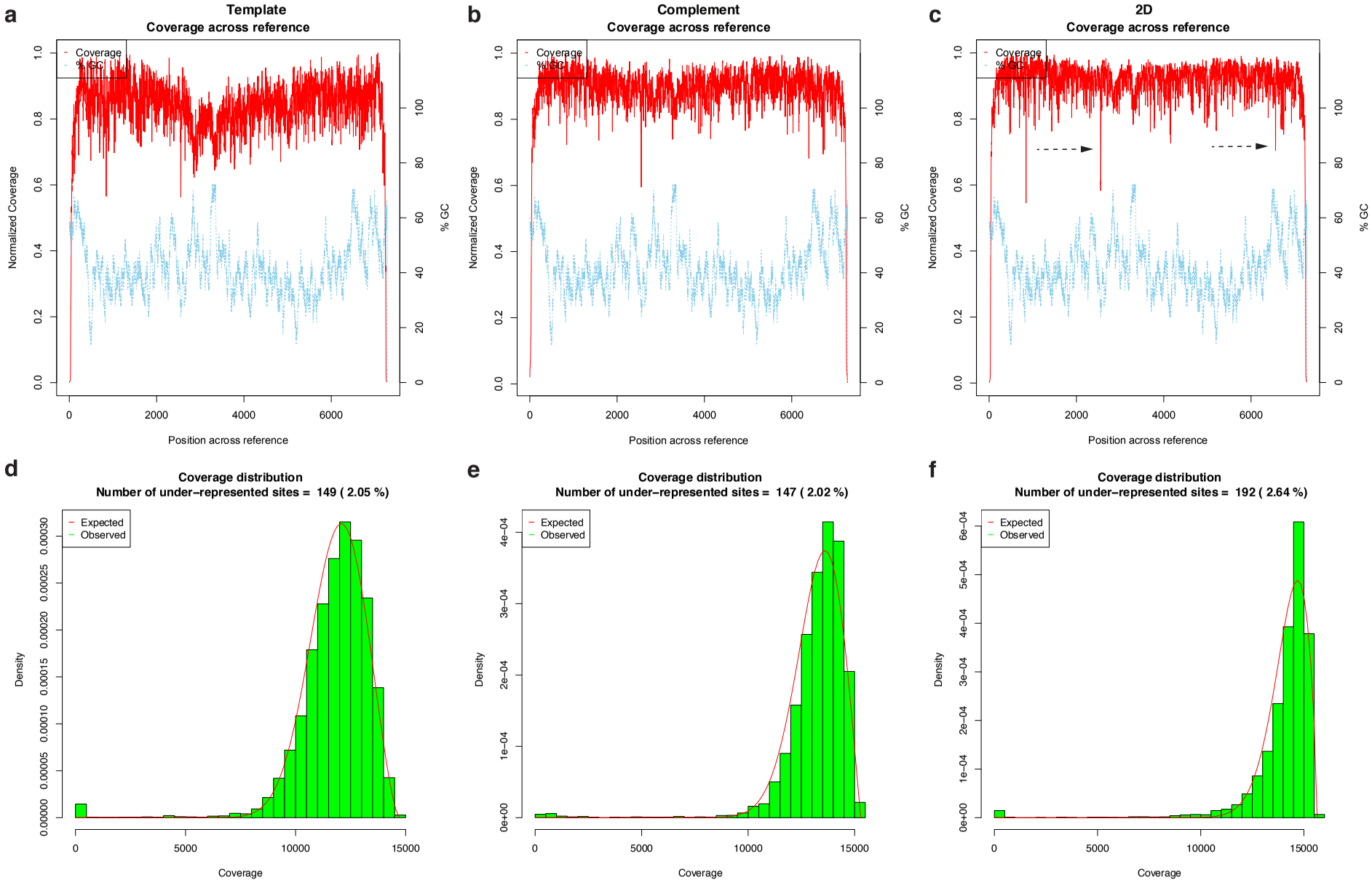


Fig. 13. (a-c) The coverage and GC% across the M13 genome. Coverage was smoothed by binning over a sliding 5 bp window, matching the k-mer length used in base calling. GC content was calculated by binning over a 50 bp sliding window, halving and doubling this window size did not drastically alter the result. (d-f) Coverage histograms for template, complement, and 2D reads across three M13 replicate experiments using R7.3 chemistry and aligned using trained LAST realignments. About 2.1%, 2.0%, and 2.6% of the M13 genome was under-represented in template, complement, and 2D reads, respectively.

Table 9. Over and under represented 5-mers between reads and M13 reference. Lambda 5-mers were not counted in this comparison. Both strands are compared, but only one is represented in this table due to symmetry.

Top Kmers In Reads vs. M13 Reference

Reference	logFC	2D	logFC	Reference	logFC	complement	logFC	Reference	logFC	template	logFC
TGATC	-inf	TTTTT	1.871	TGATC	-inf	TTTTT	1.652	TGATC	-inf	TTTTT	1.158
GATCA	-inf	AAAAA	1.871	GATCA	-inf	AAAAA	1.652	GATCA	-inf	AAAAA	1.158
GTCCG	-inf	CAAAA	0.936	GTCCG	-inf	CAAAA	1.153	GTCCG	-inf	ATTTT	1.017
CGGAC	-inf	TTTTG	0.936	CGGAC	-inf	TTTTG	1.153	CGGAC	-inf	AAAAT	1.017
GGACC	-1.95	ATTTT	0.812	GGACC	-2.088	ATTTT	1.15	GGACC	-2.279	CAAAA	0.951
GGTCC	-1.95	AAAAT	0.812	GGTCC	-2.088	AAAAT	1.15	GGTCC	-2.279	TTTTG	0.951
CTAGG	-1.553	CTTTT	0.774	CTAGG	-1.85	ACCCT	1.055	CTAGG	-2.177	CCACC	0.878
CCTAG	-1.553	AAAAG	0.774	CCTAG	-1.85	AGGGT	1.055	CCTAG	-2.177	GGTGG	0.878
ACACG	-1.497	TATAT	0.727	TGTGC	-1.826	TTTTA	0.983	TGTGC	-1.641	ACCCT	0.822
CGTGT	-1.497	ATATA	0.727	GCACA	-1.826	TAAAA	0.983	GCACA	-1.641	AGGGT	0.822
TCGTG	-1.321	CCACC	0.726	ACACG	-1.783	CTTTT	0.901	ACACG	-1.638	TGAAA	0.794
CACGA	-1.321	GGTGG	0.726	CGTGT	-1.783	AAAAG	0.901	CGTGT	-1.638	TTTCA	0.794
TGTGC	-1.317	ACCCT	0.695	TCGTG	-1.658	GTTTT	0.9	CTTCG	-1.575	CCTCA	0.702
GCACA	-1.317	AGGGT	0.695	CACGA	-1.658	AAAAC	0.9	CGAAG	-1.575	TGAGG	0.702
CTTCG	-1.293	TTTTA	0.681	CTTCG	-1.599	ATATT	0.894	ACTAG	-1.54	CACCA	0.698
CGAAG	-1.293	TAAAA	0.681	CGAAG	-1.599	AATAT	0.894	CTAGT	-1.54	TGGTG	0.698
ACTAG	-1.183	CACCA	0.583	GTCCC	-1.565	TTTAA	0.858	GCTAG	-1.439	GAAAA	0.698
CTAGT	-1.183	TGGTG	0.583	GGGAC	-1.565	TTAAA	0.858	CTAGC	-1.439	TTTTC	0.698
ATCGA	-1.138	GTTTT	0.546	ACTAG	-1.357	GAAAA	0.856	TCGTG	-1.43	CGCCA	0.696
TCGAT	-1.138	AAAAC	0.546	CTAGT	-1.357	TTTTC	0.856	CACGA	-1.43	TGGCG	0.696

Table 10. Over and under represented 5mers that span indels in aligned reads across all three read types.

Top Enriched Kmers Spanning Aligned Indels

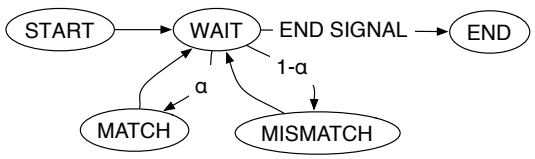
Reference	logFC	2D	logFC	Reference	logFC	complement	logFC	Reference	logFC	template	logFC
GATCA	-1.293	TTTTT	1.774	GATCC	-1.177	TTTTT	1.35	CAGAG	-1.14	GGTGG	0.99
GGATC	-1.226	ACTGG	1.196	GATCA	-0.984	AAAAA	1.01	GATCA	-1.074	TGGTG	0.889
GATCC	-1.223	TATAT	1.007	AACAG	-0.983	GCGGT	0.959	AGAGC	-1.021	ACTGG	0.831
TTTGA	-1.123	AGTTT	0.957	ACAGC	-0.978	AGTTT	0.85	GAAGC	-1.007	GGACT	0.829
GAACA	-1.095	AAAAA	0.954	CGTCA	-0.951	TGCAA	0.844	TGATC	-1.0	GCCTT	0.826
AGAGC	-1.093	TCGGT	0.949	GGATC	-0.914	AGTAA	0.828	GAGAT	-0.988	TGGCG	0.805
TGATC	-1.025	GCGGT	0.947	ATCCA	-0.887	AGTCT	0.821	AAGAG	-0.943	AAAAA	0.782
AGGGG	-1.023	AGTCT	0.944	GAACA	-0.885	ACTGG	0.812	GGAAG	-0.914	CGGTG	0.777
CTGTG	-1.005	GTTTC	0.913	CAGAG	-0.87	ATCTT	0.775	GAACC	-0.898	GGAGT	0.766
AAGAG	-0.987	TTGTC	0.846	AGAGC	-0.843	TAAAA	0.77	GAACA	-0.879	AGTCT	0.722
TGAGA	-0.934	CCAGT	0.83	TGAAC	-0.819	TCGGT	0.756	AGGGG	-0.878	GCGGT	0.714
GAGCC	-0.903	TGCAA	0.807	GAGCC	-0.806	TTTTG	0.751	GACCC	-0.85	TTTTT	0.696
GAAGC	-0.874	TGGTG	0.795	CGATC	-0.801	GGTGG	0.751	CAGGG	-0.846	TTAGT	0.694
GGAAG	-0.845	GGAAA	0.793	TGATC	-0.766	TTGTC	0.75	CTAGG	-0.844	TTGCA	0.685
GAGAG	-0.84	TAATA	0.793	CTACG	-0.766	AATCT	0.743	ACAGC	-0.818	GGTTA	0.672
AAGCA	-0.837	CGGTG	0.772	CTGTG	-0.764	GTTTT	0.734	ATCAC	-0.816	TAGTT	0.658
GACCC	-0.836	CTTGG	0.763	CATCC	-0.733	TAATA	0.726	CAGAT	-0.81	GTGAC	0.654
ATCAC	-0.835	CTCTC	0.758	ATAAC	-0.73	GACAA	0.725	GCCGC	-0.795	GGTGA	0.645
CAAAG	-0.83	CGAAA	0.751	GAAGC	-0.719	TATAT	0.701	GAGAG	-0.779	TCGGT	0.641
GCCGC	-0.824	CCTTG	0.744	ACGTC	-0.717	CGGTG	0.696	GCAGG	-0.776	GTGGT	0.629

12 Single Nucleotide Variant Calling with MinION™ reads as a demonstration of alignment accuracy

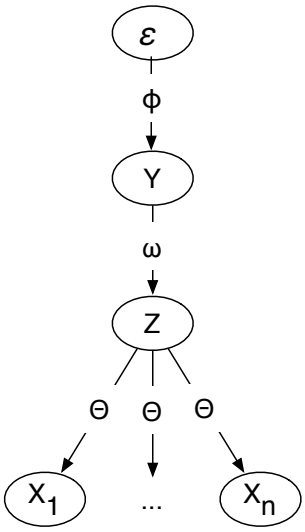
An important use of current next generation sequencing is single nucleotide variant (SNV) discovery, however the relatively high error rates of MinION™ reads make this potentially challenging (see S. Fig. 14). To establish how useful MinION™ reads are for simple SNV discovery in monoploid genomes we took the M13mp18 reference sequence and randomly introduced substitutions at a frequency of 1, 5, 10 and 20%, picking the alternate allele with equal probability for each possible alternate base. We call 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 (see below). In addition to exploring simple SNV discovery with MinION™ reads, the "held out" known differences between the mutated reference sequence and the DNA being sequenced can be used to assess read alignment accuracy, because correct alignments should improve recovery of the introduced substitutions while avoiding issues of reference allele bias. Reference allele bias being the tendency for consensus sequences derived from read alignments to resemble the reference sequence to which they are aligned because of the alignment bias towards creating matches between identical bases.

12.1 Approach to SNV detection

Let a sequence $S = S_1, \ldots, S_m$ be a finite string over the alphabet of nucleotide characters $\pi = \{A, C, T, G\}$, termed *bases*. Let $X = \{X^1, \ldots, X^n\}$ be the set of read sequences, Y the given mutated reference sequence, Z the true M13mp18 reference sequence, θ a read error model that can be used to calculate $P(X|Z, \theta)$, ω a substitution model that can be used to calculate $(Z|Y, \omega)$, and ϕ a generator model that can be used to calculate $(Y|\phi)$. Each of θ, ω and ϕ can be described as forms of branch transducer model, which are a subtype of graphical model that receive input symbols (here individual bases) from an input sequence and output symbols (again, here individual bases) to an output sequence conditional on the input symbols[16]. Branch transducers can be composed together to form evolutionary HMMs, which give HMM models for arbitrary phylogenies. Here ω is very simple, having a single parameter, α, corresponding to substitution frequency:



In the above representation of ω the *WAIT* state is a silent state that receives bases from the input sequence until it receives the END-SIGNAL at which it transitions to the end state. For each input base it chooses with probability α to emit the input base (*MATCH* state), else a different base (*MISMATCH* state). The transducers ϕ and θ composed together, $\phi \circ \theta$, are equivalent to the 5-state HMM described earlier, i.e. $P(X, Z|\phi \circ \theta) = P(X|Z, \theta)P(Z|\phi)$. Composing the branch transducers together we get an evolutionary HMM modeling the reads and reference sequences (where ϵ is the empty string):



A simple way to define the variant calling problem is that of finding a member of

$$f(X, Y) = \arg\max_{Z'} = P(Z'|Y, \omega)P(Y|\phi) \prod_i P(X^i|Z', \theta), \tag{1}$$

a maximum likelihood (ML) prediction of the true reference sequence, Z, given the mutated reference sequence and the reads. Unfortunately this optimization, corresponding to the multiple sequence alignment problem, is NP-hard [17], though exact dynamic programming algorithms that are exponential in the cardinality of X exist, and a number of principled heuristics have been proposed [18].

Let $\sim$ represent a pairwise alignment of each read sequence to the mutated reference Y. We write $Y_i \sim X^j_k$ to indicate element i of the mutated reference sequence Y is aligned to element k of read sequence X^j. As the alignment allows for only indels and matches, for each read sequence X^j, $\sim$ defines a strictly increasing relationship between the indices of aligned bases in Y and X^j. A probability calculated using an HMM can be conditioned on such an alignment by restricting the state space investigated to a subspace of the overall space. Here we define this restriction as requiring the HMM to emit the sets of aligned bases in the order defined by the sequences. While computing f is intractable, it is straightforward, given the simple definition

of ω, to compute a member of

$$f'(X, Y, \sim) = \arg\max_{Z'} P(Z'|Y, \sim, \omega)P(Y, \phi) \prod_i P(X^i|Z', \sim, \theta), \tag{2}$$

a ML estimate of the true reference sequence conditional on a fixed alignment, because, it is easy to show, this corresponds to calculating the ML base independently for each column i containing one or more aligned read positions:

$$\arg\max_{Z'_i} P(Z'_i|Y_i, \omega)P(Y_i|\psi) \prod_{X^j_k \sim Y_i} P(X^j_k|Z'_i, \theta), \tag{3}$$

concatenating the resulting ML bases together in order to form Z'.

To generate an alignment $\sim$ we used one of the mapping programs described earlier, or the composed transducer $\phi \circ \omega \circ \theta$ (see below), which combines the five-state HMM error model described earlier with the simple model for substitutions between Y and Z and the sequencing generating transducer ϕ. The parameters for the error model were determined using the EM training described earlier, the substitution parameter for ω was set by manual, empirical investigation.

A simple improvement over using the fixed alignment algorithm is to use the posterior match probabilities between bases in the alignments to replace (3) with

$$\arg\max_{Z'_i} P(Z'_i|Y_i, \omega)P(Y_i|\psi) \prod_j \sum_k P(X^j_k|Z'_i, \theta)P(X^j_k \sim Y_i|\phi \circ \omega \circ \theta), \tag{4}$$

where $P(X^j_k \sim Y_i|\phi \circ \omega \circ \theta)$ is the posterior probability that the element k of sequence X^j is aligned to element i in sequence Y given the composed transducer $\phi \circ \omega \circ \theta$. Note this is not the same as evaluating f directly, but instead is equivalent to the column calculation in 3 marginalising over the probability of all pairwise alignments between each read and the mutated reference sequence.

Instead of calculating 4 we can alternatively calculate the related *posterior base calling probability* that the base at given index of Z is equal to a given base, and so obtain the likelihood of each alternate base (bases not the same as the given mutated reference base) for our chosen parameters. We can then assess the number of non-reference true positive and false positive predictions with a posterior probability greater than or equal to a given value. We define a *false positive* for an index i and posterior probability p as a base x not equal to either Y_i or Z_i and with posterior base calling probability $\geq p$. Conversely, we define a *true positive* to be when x is equal to Z_i, not equal to Y_i (because we are interested in sites that have changed between the true and mutated reference), and the posterior base calling probability is $\geq p$. Given

these definitions, summing over all columns, we use standard the information theoretic measures of precision, recall and F-score to judge performance for a given posterior probability threshold.

12.2 MinION^{TM} reads can call SNVs with high recall and precision.

The described SNV calling algorithm has two steps: computing posterior alignment match probabilities and then calculating posterior base calling probabilities. Starting with 2D reads aligned with tuned LAST (as described earlier; for this task LAST was found to work slightly better than using BLASR (data not shown)), to compute the posterior match probabilities we constructed a band around the guide alignment, exactly as in the EM-training described earlier, and computed the forward-backward algorithm within the band. The model $\phi \circ \omega \circ \theta$ was composed by combining an EM trained HMM model ($\phi \circ \theta$) on 2D reads using tuned LAST as the guide alignment (as described earlier) with the substitution model ω, setting $\alpha = 0.8$, which was found to work well and which corresponds to a mismatch rate of 20%.

Supplementary Fig. 15 and Supplementary Table 11 shows the results. Note the numbers in the table (and subsequent tables) are the avg. precision/recall/F-scores over all replicates, where for each replicate the precision/recall/F-score value shown is for the optimal F-score for that replicate. In the figure (and subsequent figures), the precision and recall value pairs which define the curves are the avg. over all replicates as a function of the posterior base calling probability threshold.

In short, at a mutation frequency of 1% using all the data and choosing a posterior base calling threshold that gives the optimal avg. F-score for each replicate we achieve, in this best case scenario, an avg. recall of $\geq 99\%$ and precision of $\geq 99\%$. Reducing the coverage down to a more reasonable 60x we achieve a recall and precision of 97%. Increasing the mutation frequency decreases the F-score progressively, presumably because the alignment between reads and the mutated reference becomes even harder.

To demonstrate the methods and parameters we chose were reasonable we compared to a number of parameter and algorithm variations.

In calculating the posterior match probabilities setting $\alpha = 0.6$ (a mismatch rate of 40%) we see a decrease in F-score for a 1% mutation frequency (avg. across all coverages), but a gain for 5% and greater mutation frequencies (Supplementary Fig. 16 and Supplementary Table 12). This suggests, as might be expected, that α should be set lower when the expected divergence between the reference and sample is greater. With $\alpha = 0.6$ we achieve an avg. precision and recall of 98% for a 5% mutation frequency.

For $\alpha = 1.0$ (equivalent to not modeling mismatches) we see very significantly lower performance (Supplementary Fig. 17 and Supplementary Table 13). We speculate the relatively large α values work well because the trained model strongly prefers to avoid certain matches - e.g. adenosine to thymine, but such matches should be made when aligning the reads to a mutated reference sequence rather than the true reference

sequence. The higher substitution rates therefore allows the model to overcome this bias, rather than giving weight to likely alternative scenarios, e.g. the creation of additional indels to avoid these matches.

In calculating the posterior base calling probabilities switching θ from the EM trained model to a model which treats all substitutions as having equal probability (and which is therefore equivalent to picking the base with highest posterior match probability expectation) we find a very small decrease in performance (Supplementary Fig. 18 and Supplementary Table 14), suggesting the trained substitution model performs better than a naive strategy.

Switching from using posterior match probabilities to a fixed input alignment in the calculation of the posterior base calling probability we find significantly lower performance (Supplementary Fig. 19 and Supplementary Table 15). This is unsurprising given that the modal posterior match probability is less than 90% (Fig. 5(C)).

As might be expected, switching to using template or complement reads instead of 2D reads we find substantially poorer performance (Supplementary Fig. 20-21 and Supplementary Tables 16-17), however, this may be somewhat down to using an alignment model trained for 2D reads.

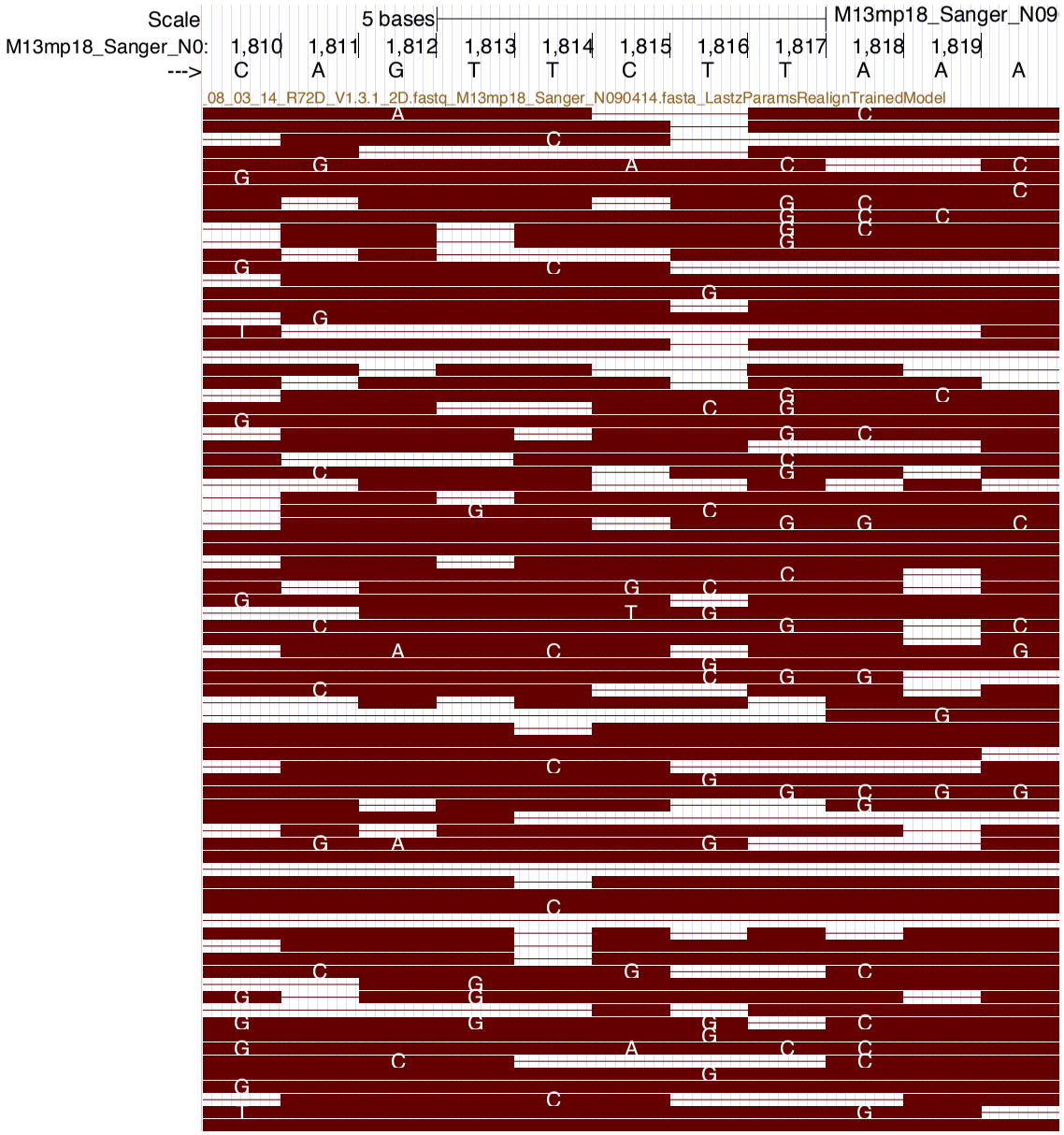


Fig. 14. Visualization of an alignment of 2D reads with M13 using trained LAST realignments on the UCSC Genome Browser. The high indel and mismatch rate are clearly evident.

SNV detection using 2D reads

Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	94.59	97.72	99.00	100.00
	5	94.77	96.14	96.26	96.66
	10	94.52	95.25	95.68	96.16
	20	91.68	92.27	92.51	93.19
Precision	1	96.29	97.79	99.43	99.58
	5	98.03	98.80	98.66	99.04
	10	96.79	97.57	98.30	98.14
	20	93.85	94.90	95.73	96.12
F-score	1	95.40	97.73	99.21	99.79
	5	96.37	97.45	97.44	97.83
	10	95.63	96.40	96.97	97.14
	20	92.74	93.56	94.09	94.63

Table 11. Variant calling on M13 using 2D reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

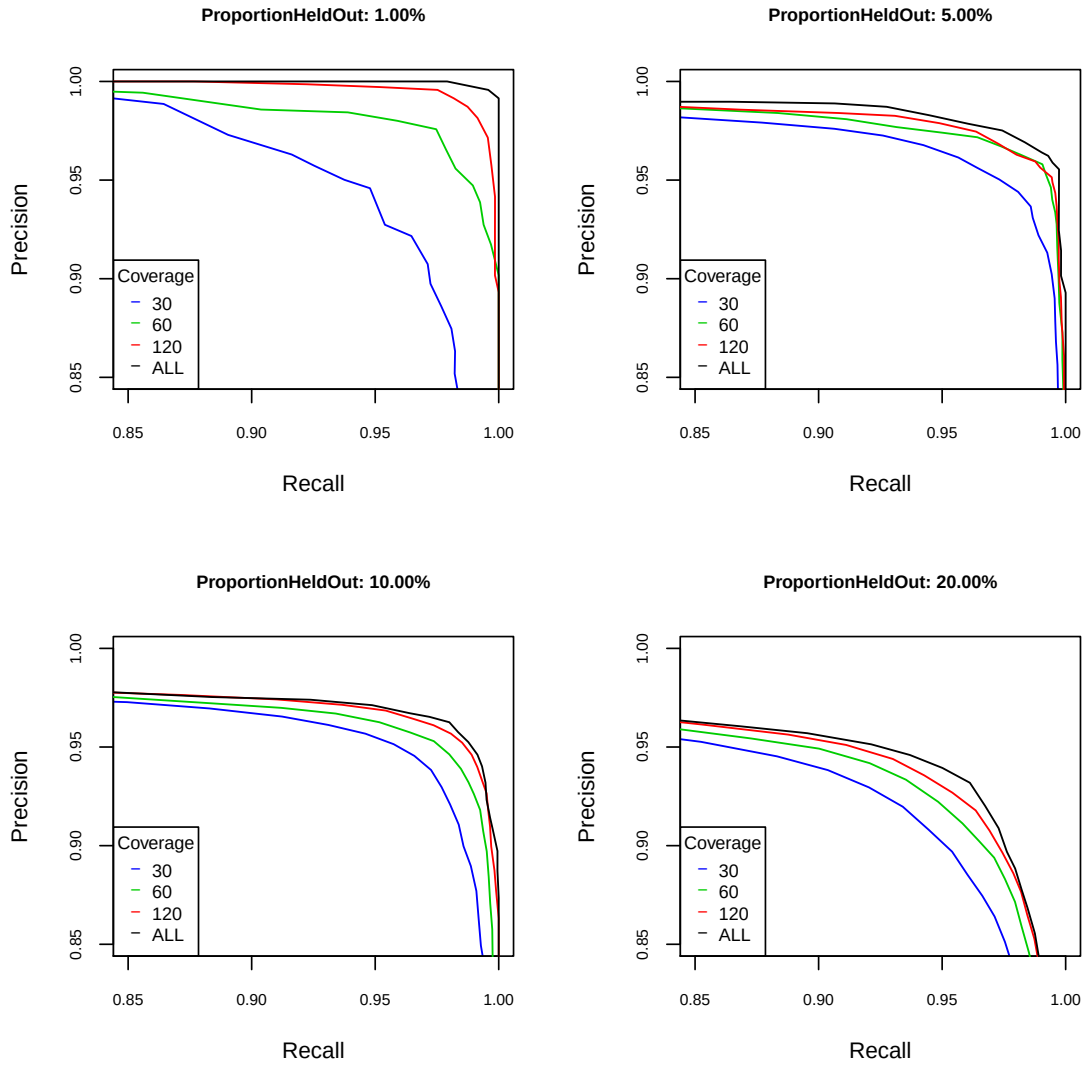


Fig. 15. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using 2D reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

SNV detection using 2D reads					
Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	94.87	96.72	97.58	98.72
	5	96.03	97.23	97.63	98.71
	10	95.49	96.29	96.30	96.44
	20	94.13	94.28	95.18	95.00
Precision	1	95.25	96.77	98.28	99.15
	5	97.43	98.23	98.31	97.96
	10	96.84	98.20	98.64	99.25
	20	95.86	97.06	97.01	97.71
F-score	1	95.02	96.72	97.92	98.93
	5	96.72	97.72	97.97	98.34
	10	96.16	97.23	97.46	97.82
	20	94.98	95.65	96.08	96.33

Table 12. Variant calling on M13 using 2D reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 40% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

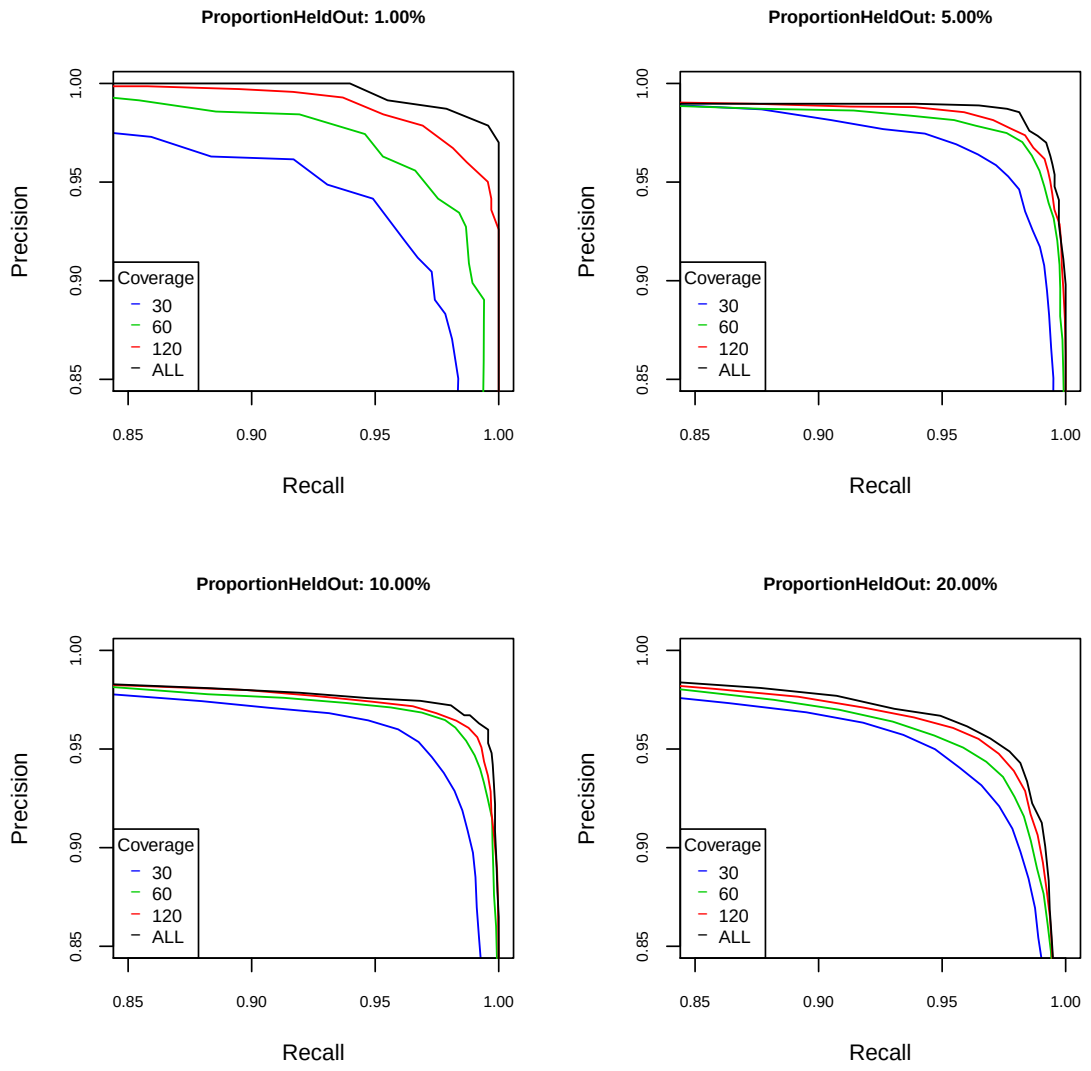


Fig. 16. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using 2D reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 40% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

SNV detection using 2D reads					
Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	89.60	89.74	91.60	91.88
	5	83.86	84.92	84.86	85.52
	10	74.58	74.87	75.60	77.35
	20	67.12	67.11	67.05	68.10
Precision	1	92.98	96.80	97.90	98.64
	5	94.88	95.21	97.08	96.90
	10	88.10	88.53	89.62	88.24
	20	82.25	82.88	84.35	83.02
F-score	1	91.23	93.12	94.63	95.13
	5	89.01	89.76	90.55	90.85
	10	80.76	81.11	82.00	82.41
	20	73.89	74.15	74.70	74.82

Table 13. Variant calling on M13 using 2D reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, without accounting for substitution differences between the given reference and true underlying reference. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

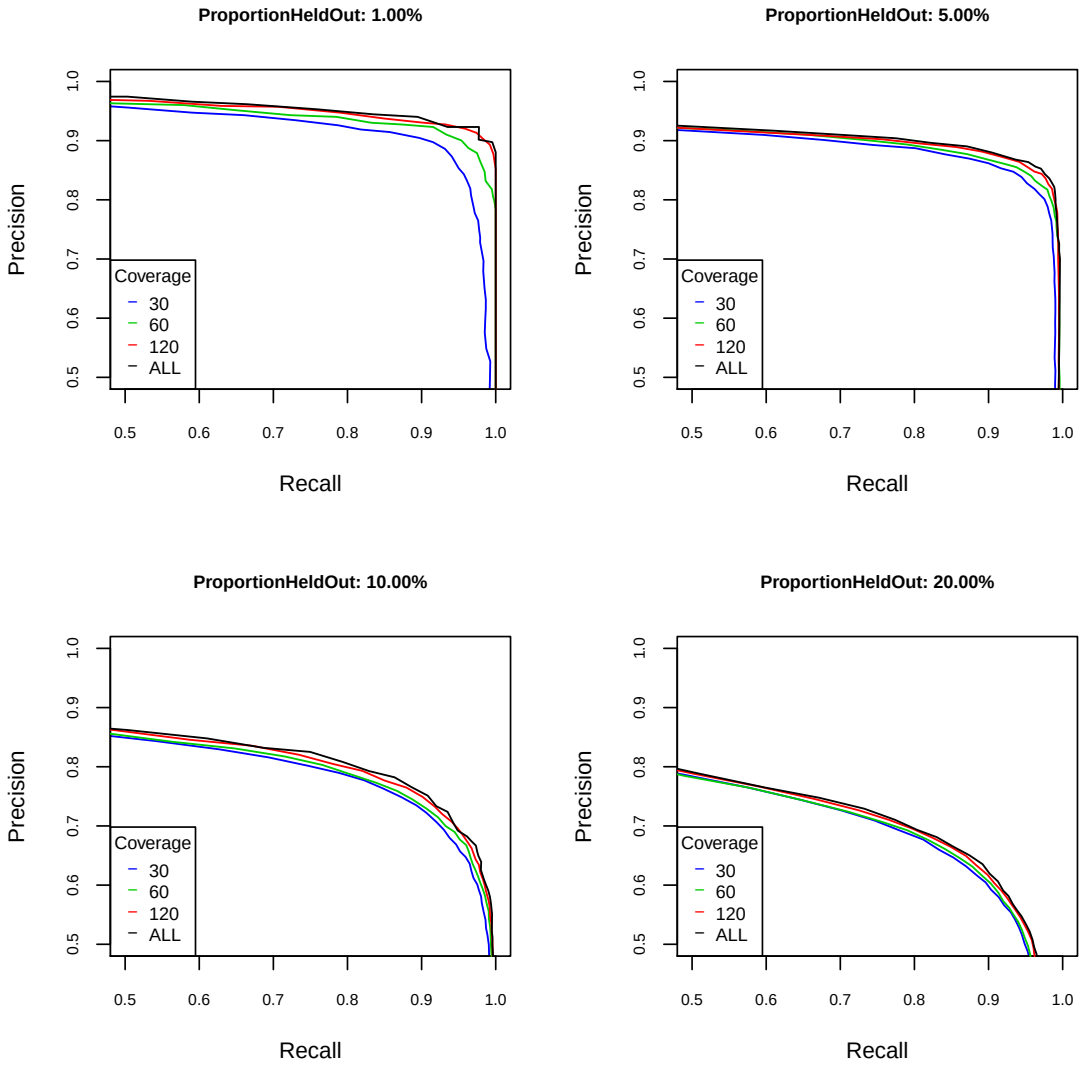


Fig. 17. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using 2D reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, without accounting for substitution differences between the given reference and true underlying reference. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

SNV detection using 2D reads

Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	95.16	98.15	99.15	100.00
	5	94.83	96.32	95.74	97.00
	10	94.37	95.04	95.57	96.07
	20	91.56	92.04	92.98	93.52
Precision	1	95.05	97.23	99.01	100.00
	5	97.75	98.54	99.01	98.44
	10	97.02	97.86	98.40	98.23
	20	94.06	95.12	95.40	95.70
F-score	1	95.08	97.67	99.08	100.00
	5	96.26	97.41	97.34	97.71
	10	95.67	96.42	96.97	97.14
	20	92.78	93.55	94.17	94.59

Table 14. Variant calling on M13 using 2D reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling corresponds to choosing the maximum-frequency/expectation of a non-reference base. Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

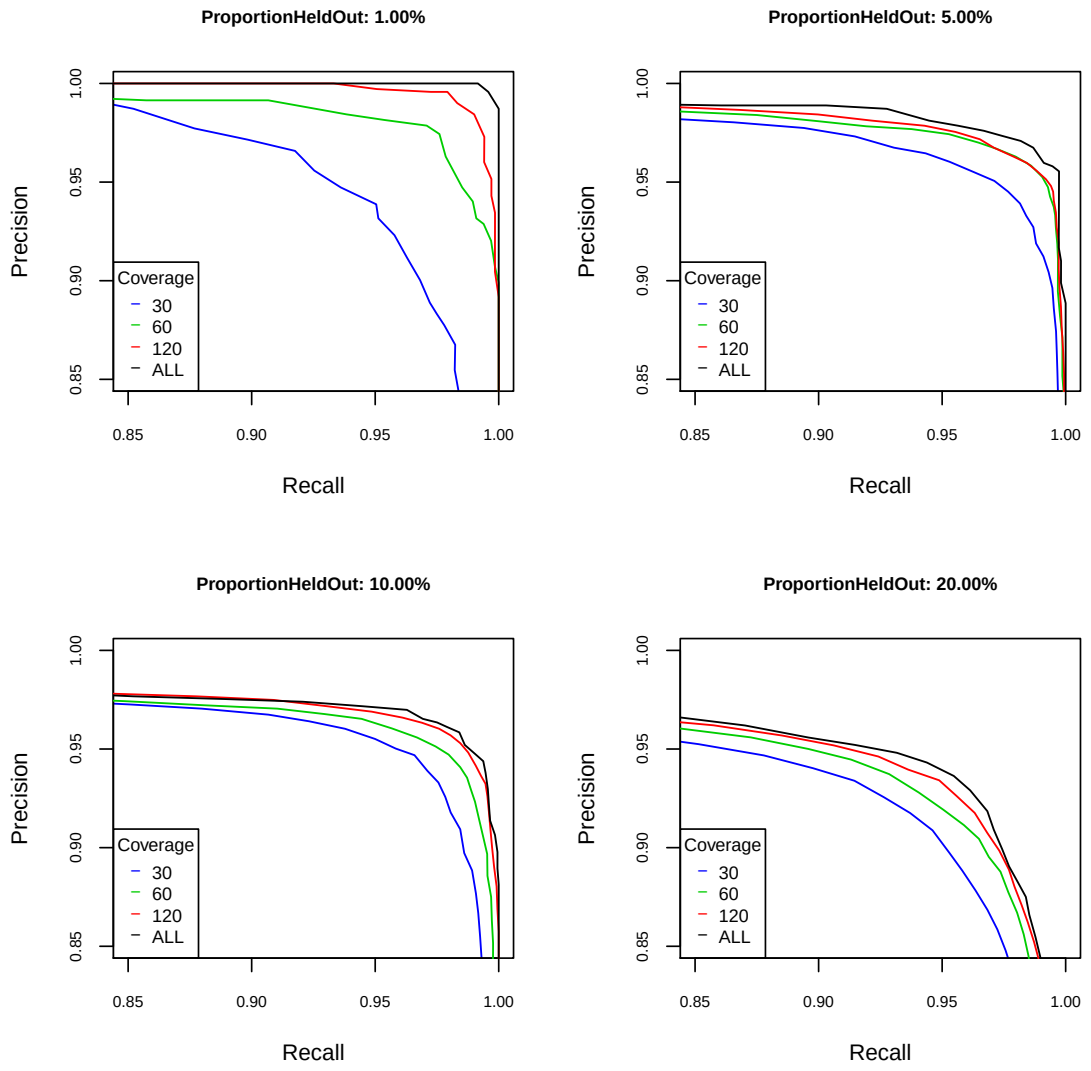


Fig. 18. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using 2D reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling corresponds to choosing the maximum-frequency/expectation of a non-reference base. Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

SNV detection using 2D reads

Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	92.02	92.74	95.87	96.58
	5	93.29	95.06	94.80	95.63
	10	93.18	94.58	95.62	95.53
	20	90.36	91.61	92.25	92.81
Precision	1	91.46	92.85	92.14	97.84
	5	96.22	96.11	96.50	98.50
	10	96.60	96.59	97.00	97.99
	20	94.43	95.53	96.04	96.73
F-score	1	91.67	92.72	93.95	97.20
	5	94.72	95.57	95.64	97.04
	10	94.86	95.57	96.30	96.74
	20	92.34	93.52	94.10	94.72

Table 15. Variant calling on M13 using 2D reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed conditioned on the fixed input alignment. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

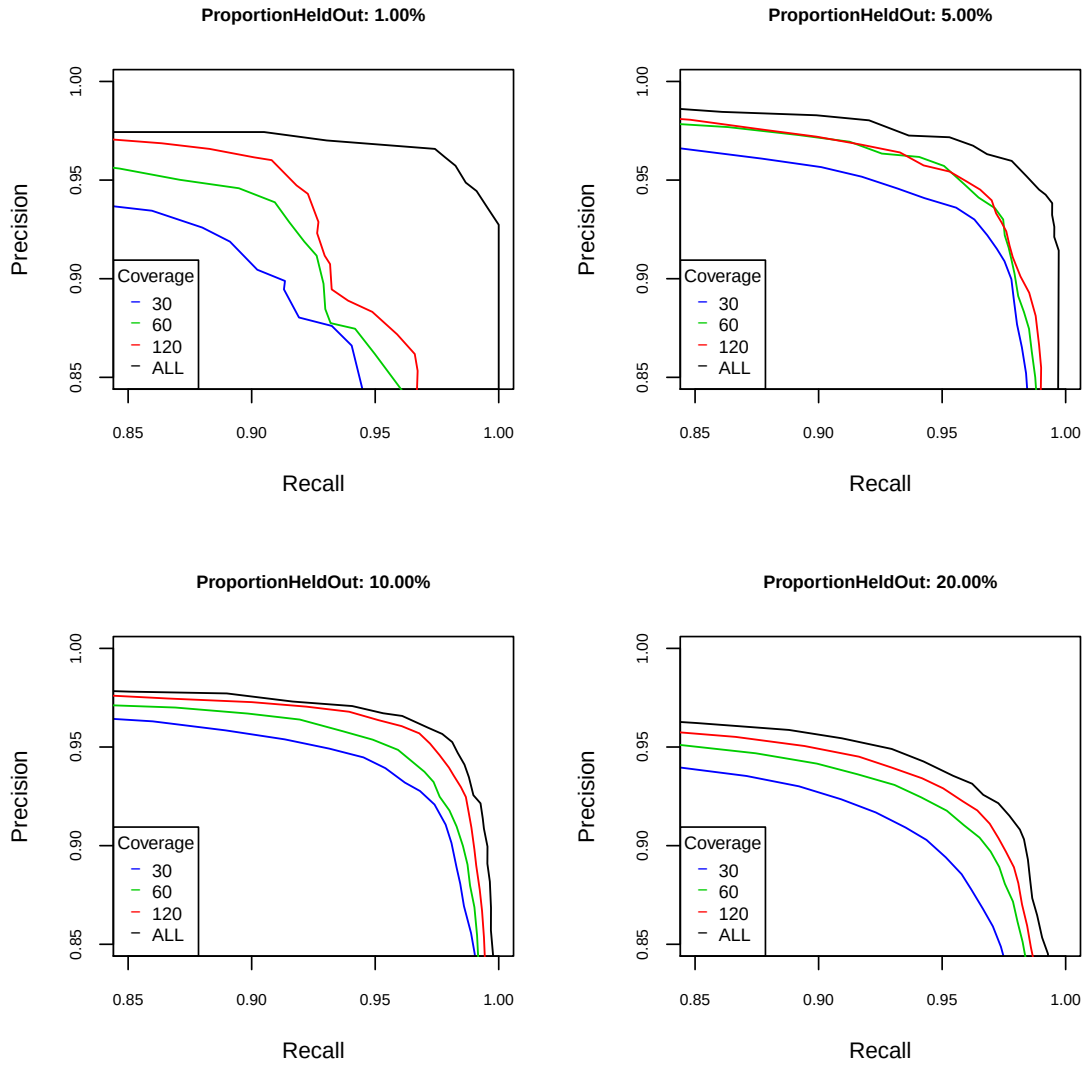


Fig. 19. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using 2D reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed conditioned on the fixed input alignment. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

SNV detection using complement reads

Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	66.24	70.80	74.64	75.64
	5	78.75	82.98	85.38	88.52
	10	75.56	79.36	80.18	82.92
	20	72.84	76.07	77.42	78.72
Precision	1	81.69	88.86	90.28	91.66
	5	83.87	87.83	88.99	90.00
	10	83.95	85.15	87.95	88.26
	20	80.03	82.21	83.78	84.98
F-score	1	72.95	78.66	81.47	82.76
	5	81.09	85.30	87.09	89.25
	10	79.45	82.09	83.86	85.50
	20	76.23	78.95	80.42	81.72

Table 16. Variant calling on M13 using complement reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

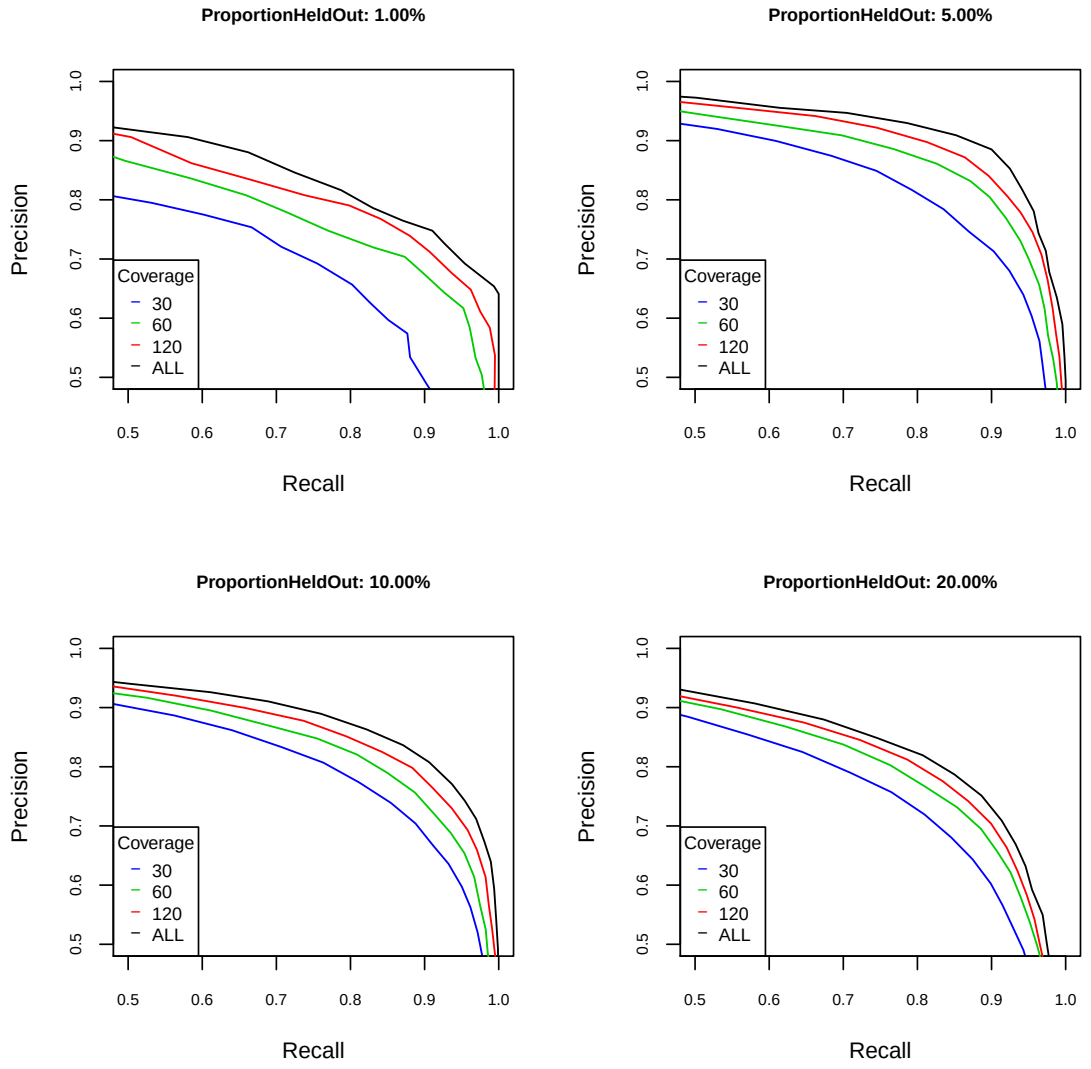


Fig. 20. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using complement reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

SNV detection using template reads					
Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	52.56	61.25	62.39	64.10
	5	69.15	75.24	76.46	78.32
	10	69.47	74.66	75.31	77.26
	20	67.40	71.69	73.84	75.38
Precision	1	68.25	68.05	70.64	78.85
	5	75.92	79.99	83.64	84.73
	10	74.33	78.66	81.36	83.74
	20	74.44	77.69	78.12	80.57
F-score	1	59.10	63.92	66.02	70.48
	5	72.31	77.49	79.79	81.36
	10	71.70	76.56	78.16	80.27
	20	70.67	74.54	75.87	77.89

Table 17. Variant calling on M13 using template reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling corresponds to choosing the maximum-frequency/expectation of a non-reference base. Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

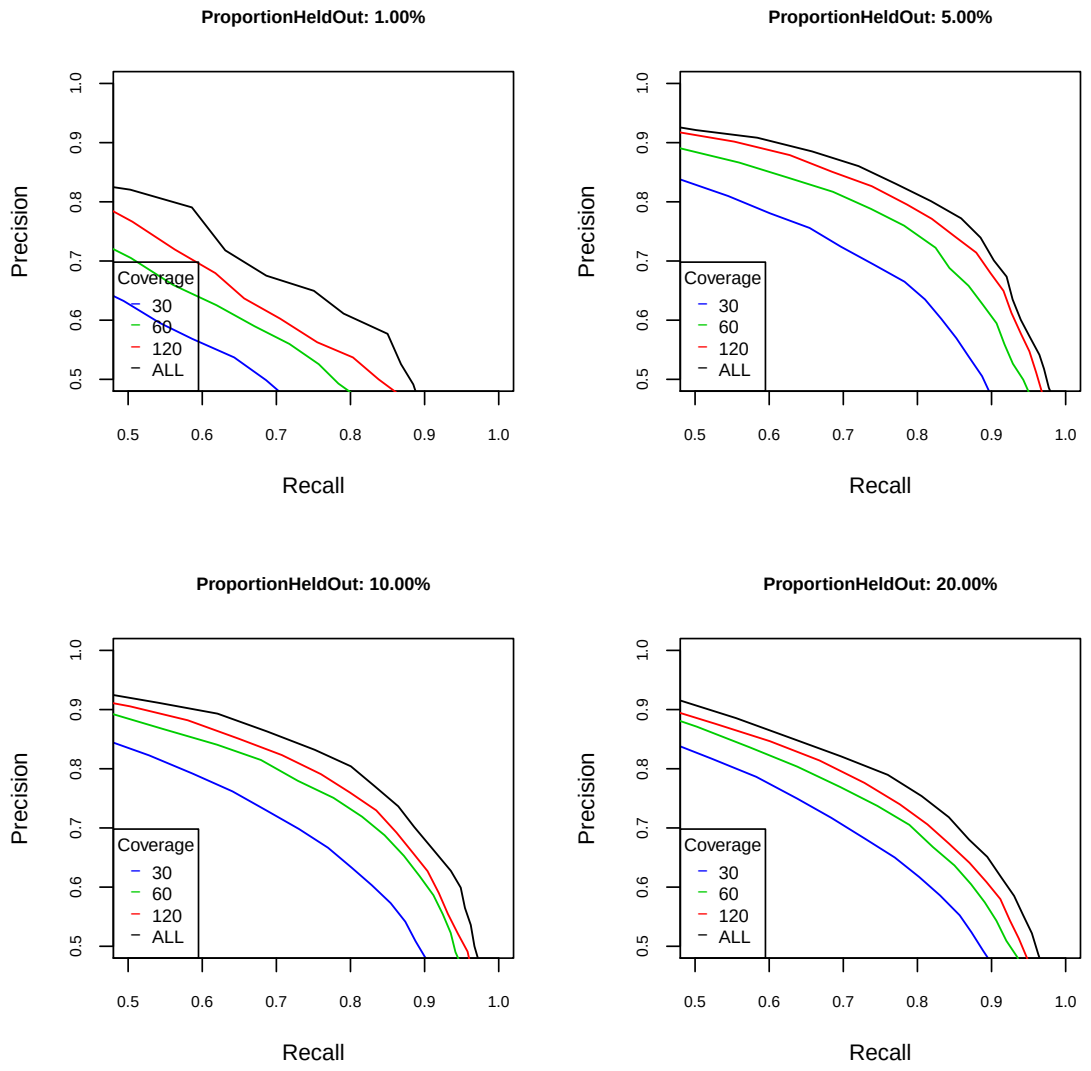


Fig. 21. Precision/recall curves showing variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling performed using template reads starting with the tuned Last (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference sequence, using the initial guide alignment to band the calculations. Variant calling corresponds to choosing the maximum-frequency/expectation of a non-reference base. Posterior match probabilities calculated using the EM trained HMM model, accounting for substitution differences between the mutated reference and true underlying reference, assuming 20% divergence. Variant calling results shown for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference to which reads where aligned, and for which variants were called. Coverage is the total length of reads sampled divided by the length of the reference. ALL corresponds to using all the reads for a given experiment. Results shown are across three replicate experiments, and, at each coverage value, three different samplings of the reads. Raw results are available in the supplementary spread-sheet.

13 High Molecular Weight Sequence Scaffolding across tandemly-duplicated CT47 repeat cluster using MinION reads

High molecular weight BAC DNA (RP11-482A22) was isolated using standard methods for purification of large constructs (QIAGEN Large-Construct Kit, cat 12462). To avoid DNA shearing for high-molecular weight sequencing, we performed NotI-HF (NEB Cat. No. R3189S) restriction digest (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 then proceeded for rest of the steps as the standard ONT recommended steps, mentioned above. The device was operated using ONT's MinKNOW software, according to the provided instructions. The flowcells used were chemistry version R6.0 and R7.0. The read files were base called using ONT's Metrichor software, version 2D basecalling v1.2 and v1.3.1.

Long reads spanning the CT47-repeat cluster were identified using three sequence models[19]: a single copy sequence directly upstream of the repeat array (6.6 kb, hg38 chrX:120865735 120872351), 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). Nine reads were identified to contain both upstream and downstream models with each supporting the estimate of eight CT47-repeat copies (see Supplementary Data 1 below). Reads were trimmed to the only present sequences involved in the repeat classification models (Data available in ENA; primary accession number is PRJEB8230, and the secondary accession number is ERP009289). Pecan software was used to generate multiple alignment of reads (Data available in ENA; primary accession number is PRJEB8230, and the secondary accession number is ERP009289)[12].

Supplementary Data 1: MinION long read CT47-repeat characterization

Rd No.	Read ID	Total Read Size	HMM Model Prediction	Trim Start	Trim End	Span through CT47-Rpts (+Upstream and Downstream HMM Models)	HMM Model Prediction Start	HMM Model Prediction End	Trim Read Start	Trim Read End	HMM Model Prediction Base Span Trim Read
1	channel_278_read_20	38375	Upstream	**36**	36208	**36172**	27	6611	5	5513	5509
1	channel_278_read_20	38375	Rpt1	36	36208	**36172**	1121	4859	5247	8569	3323
1	channel_278_read_20	38375	Rpt2	36	36208	**36172**	42	4859	8571	12650	4080
1	channel_278_read_20	38375	Rpt3	36	36208	**36172**	12	4858	12653	16678	4026
1	channel_278_read_20	38375	Rpt4	36	36208	**36172**	1	4819	16679	20779	4101
1	channel_278_read_20	38375	Rpt5	36	36208	**36172**	20	4857	20783	24875	4093
1	channel_278_read_20	38375	Rpt6	36	36208	**36172**	35	4635	24880	28864	3985
1	channel_278_read_20	38375	Rpt7	36	36208	**36172**	11	4815	28872	32989	4118
1	channel_278_read_20	38375	Rpt8	36	36208	**36172**	17	1164	32983	34017	1035
1	channel_278_read_20	38375	Downstream	36	36208	**36172**	1	2686	33901	36169	2269
2	channel_198_read_22	40110	Upstream	21	37816	**37795**	28	6596	3	5740	5738
2	channel_198_read_22	40110	Rpt1	21	37816	**37795**	1044	4853	5547	8908	3362
2	channel_198_read_22	40110	Rpt2	21	37816	**37795**	17	4606	8913	13183	4271
2	channel_198_read_22	40110	Rpt3	21	37816	**37795**	42	4858	13218	17460	4243
2	channel_198_read_22	40110	Rpt4	21	37816	**37795**	1	4856	17461	21675	4215
2	channel_198_read_22	40110	Rpt5	21	37816	**37795**	9	4859	21677	25938	4262
2	channel_198_read_22	40110	Rpt6	21	37816	**37795**	1	4849	25941	30183	4243
2	channel_198_read_22	40110	Rpt7	21	37816	**37795**	3	4819	30185	34478	4294
2	channel_198_read_22	40110	Rpt8	21	37816	**37795**	24	1271	34488	35723	1236
2	channel_198_read_22	40110	Downstream	21	37816	**37795**	1	2703	35424	37793	2370
3	channel_227_read_5	39526	Upstream	39	37293	**37254**	50	6611	5	5601	5597
3	channel_227_read_5	39526	Rpt1	39	37293	**37254**	949	4858	5334	8777	3444
3	channel_227_read_5	39526	Rpt2	39	37293	**37254**	6	4811	8780	12994	4215
3	channel_227_read_5	39526	Rpt3	39	37293	**37254**	5	4832	12998	17166	4169
3	channel_227_read_5	39526	Rpt4	39	37293	**37254**	1	4825	17167	21371	4205
3	channel_227_read_5	39526	Rpt5	39	37293	**37254**	1	4813	21375	25570	4196
3	channel_227_read_5	39526	Rpt6	39	37293	**37254**	13	4842	25572	29819	4248
3	channel_227_read_5	39526	Rpt7	39	37293	**37254**	5	4841	29816	33949	4134
3	channel_227_read_5	39526	Rpt8	39	37293	**37254**	4	1171	33950	35008	1059
3	channel_227_read_5	39526	Downstream	39	37293	**37254**	79	2723	34931	37254	2324
4	channel_277_read_0	39384	Upstream	2260	39357	**37097**	32	6613	10	5621	5612
4	channel_277_read_0	39384	Rpt1	2260	39357	**37097**	1050	4848	5450	8806	3357
4	channel_277_read_0	39384	Rpt2	2260	39357	**37097**	19	4859	8807	12977	4171
4	channel_277_read_0	39384	Rpt3	2260	39357	**37097**	1	4857	12979	17204	4226
4	channel_277_read_0	39384	Rpt4	2260	39357	**37097**	10	4820	17207	21402	4196
4	channel_277_read_0	39384	Rpt5	2260	39357	**37097**	6	4153	21413	25055	3643
4	channel_277_read_0	39384	Rpt6	2260	39357	**37097**	1339	4791	26300	29571	3272
4	channel_277_read_0	39384	Rpt7	2260	39357	**37097**	1	4857	29594	33838	4245
4	channel_277_read_0	39384	Rpt8	2260	39357	**37097**	20	1174	33844	35077	1234
4	channel_277_read_0	39384	Downstream	2260	39357	**37097**	6	2668	34763	37096	2334
5	channel_433_read_0	39384	Upstream	4141	40520	**36379**	4735	6617	2	1762	1761
5	channel_433_read_0	39384	Rpt1	4141	40520	**36379**	902	4858	1338	5174	3837
5	channel_433_read_0	39384	Rpt2	4141	40520	**36379**	1	4859	5180	9772	4593
5	channel_433_read_0	39384	Rpt3	4141	40520	**36379**	1	4857	9775	14300	4526
5	channel_433_read_0	39384	Rpt4	4141	40520	**36379**	1	4831	14302	18907	4606
5	channel_433_read_0	39384	Rpt5	4141	40520	**36379**	1	4859	18910	23573	4664

Rd No.	Read ID	Total Read Size	HMM Model Prediction	Trim Start	Trim End	Span through CT47-Rpts (+Upstream and Downstream HMM Models)	HMM Model Prediction Start	HMM Model Prediction End	Trim Read Start	Trim Read End	HMM Model Prediction Base Span Trim Read
5	channel_433_read_0	39384	Rpt6	4141	40520	**36379**	1	4859	23576	28138	4563
5	channel_433_read_0	39384	Rpt7	4141	40520	**36379**	1	4859	28141	32799	4659
5	channel_433_read_0	39384	Rpt8	4141	40520	**36379**	1	1169	32802	34173	1372
5	channel_433_read_0	39384	Downstream	4141	40520	**36379**	1	2713	33850	36378	2529
6	channel_456_read_11	50527	Upstream	11	38532	**38521**	4719	6617	2	1873	1872
6	channel_456_read_11	50527	Rpt1	11	38532	**38521**	773	4816	1404	5536	4133
6	channel_456_read_11	50527	Rpt2	11	38532	**38521**	6	4858	5554	10471	4918
6	channel_456_read_11	50527	Rpt3	11	38532	**38521**	1	4823	10474	15298	4825
6	channel_456_read_11	50527	Rpt4	11	38532	**38521**	1	4859	15308	20151	4844
6	channel_456_read_11	50527	Rpt5	11	38532	**38521**	1	4857	20154	25000	4847
6	channel_456_read_11	50527	Rpt6	11	38532	**38521**	1	4848	25003	29828	4826
6	channel_456_read_11	50527	Rpt7	11	38532	**38521**	1	4859	29832	34684	4853
6	channel_456_read_11	50527	Rpt8	11	38532	**38521**	1	1170	34687	35915	1229
6	channel_456_read_11	50527	Downstream	11	38532	**38521**	7	2715	35770	38520	2751
7	channel_462_read_4	44672	Upstream	68	42160	**42092**	36	6617	4	6441	6438
7	channel_462_read_4	44672	Rpt1	68	42160	**42092**	906	4859	6016	9979	3964
7	channel_462_read_4	44672	Rpt2	68	42160	**42092**	1	4859	9982	14850	4869
7	channel_462_read_4	44672	Rpt3	68	42160	**42092**	1	4859	14854	19640	4787
7	channel_462_read_4	44672	Rpt4	68	42160	**42092**	1	4829	19643	24262	4620
7	channel_462_read_4	44672	Rpt5	68	42160	**42092**	1	4859	24265	29004	4740
7	channel_462_read_4	44672	Rpt6	68	42160	**42092**	1	4848	29007	33739	4733
7	channel_462_read_4	44672	Rpt7	68	42160	**42092**	1	4859	33742	38422	4681
7	channel_462_read_4	44672	Rpt8	68	42160	**42092**	1	1170	38425	39801	1377
7	channel_462_read_4	44672	Downstream	68	42160	**42092**	2	2716	39461	42091	2631
8	channel_506_read_6	41355	Upstream	2794	41323	**38529**	7	4901	1	4391	4391
8	channel_506_read_6	41355	Rpt1	2794	41323	**38529**	5320	6613	4550	5750	1201
8	channel_506_read_6	41355	Rpt2	2794	41323	**38529**	947	4857	5447	9025	3579
8	channel_506_read_6	41355	Rpt3	2794	41323	**38529**	7	4820	9026	13421	4396
8	channel_506_read_6	41355	Rpt4	2794	41323	**38529**	7	4857	13424	17838	4415
8	channel_506_read_6	41355	Rpt5	2794	41323	**38529**	1	4846	17840	22233	4394
8	channel_506_read_6	41355	Rpt6	2794	41323	**38529**	20	4851	22238	26739	4502
8	channel_506_read_6	41355	Rpt7	2794	41323	**38529**	4	4800	26740	31239	4500
8	channel_506_read_6	41355	Rpt8	2794	41323	**38529**	39	4809	31255	35589	4335
8	channel_506_read_6	41355	Downstream	2794	41323	**38529**	1	2156	36538	38510	1973
9	channel_94_read_4	43785	Upstream	84	41266	**41182**	22	6617	5	6178	6174
9	channel_94_read_4	43785	Rpt1	84	41266	**41182**	828	4857	5746	9701	3956
9	channel_94_read_4	43785	Rpt2	84	41266	**41182**	2	4858	9706	14355	4650
9	channel_94_read_4	43785	Rpt3	84	41266	**41182**	1	4859	14357	18898	4542
9	channel_94_read_4	43785	Rpt4	84	41266	**41182**	1	4859	18901	23527	4627
9	channel_94_read_4	43785	Rpt5	84	41266	**41182**	11	4859	23530	28250	4721
9	channel_94_read_4	43785	Rpt6	84	41266	**41182**	2	4857	28253	32890	4638
9	channel_94_read_4	43785	Rpt7	84	41266	**41182**	6	4859	32896	37482	4587
9	channel_94_read_4	43785	Rpt8	84	41266	**41182**	16	1160	37487	38909	1423
9	channel_94_read_4	43785	Downstream	84	41266	**41182**	39	2709	38528	41180	2653

14 CT47 repeat 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 Cat. No. 520079). By alignment to the hg38 reference sequence (hg38 chrX:120,814,747-121,061,920, omitting 50 kb scaffold gap), using BLASR Tuned (as described above) we identified 2006 2D reads that mapped to the RP11-482A22 DNA. Base coverage was determined from sorted alignment RP11-482A22 bam file using bedtools genomecov (bedtools genomecov -d -ibam mapping.sorted.bam)[20]. Coverage estimates were converted to a bed file with each row entry defining coverage at a single base and base + 1, and then subdivided into bases that overlapped with the CT47 repeat region and those that did not overlap with the repeats, labeled as flanking regions (bedtools intersect -woa and -v, respectively)[20]. Histogram of base coverage was determined across all flanking bases, and determined to have a mean coverage value of 46.2 bases. Base coverage estimates across the CT47 repeats were combined represent a combined depth over a single 4.8 kb repeat unit (mean observed base coverage of 329.3). Normalization of read depth for 8 copies of the repeat predicted an average read depth of 41 bases. The distribution of normalized read depth was provided by dividing by 8 across all base positions of the repeat with combined sequence depth.

15 Pulse-field gel electrophoresis validation of RP11-482A22 insert length

BAC insert length estimate of NotI-HF (NEB, cat R3189S) or AatII (NEB, cat R0117S) digested DNA (1 μg) was determined by pulse field gel electrophoresis (PFGE) using a CHEF-DRII system (BioRad). Length estimates were determined using standard PFGE markers: Low-range (NEB, cat N0350S) and MidRange I (NE551S). Samples were run for 15 hrs (gradient 6.0V/cm, in angle 120 degrees, switch time linear, with initial ramping 0.2 seconds and finishing at 26 seconds) in 1% pulsed field certified agarose (BioRad) and 0.5x TBE at 4^{o}C. Banding was identified using standard SYBR Gold (LifeTechnologies) staining.

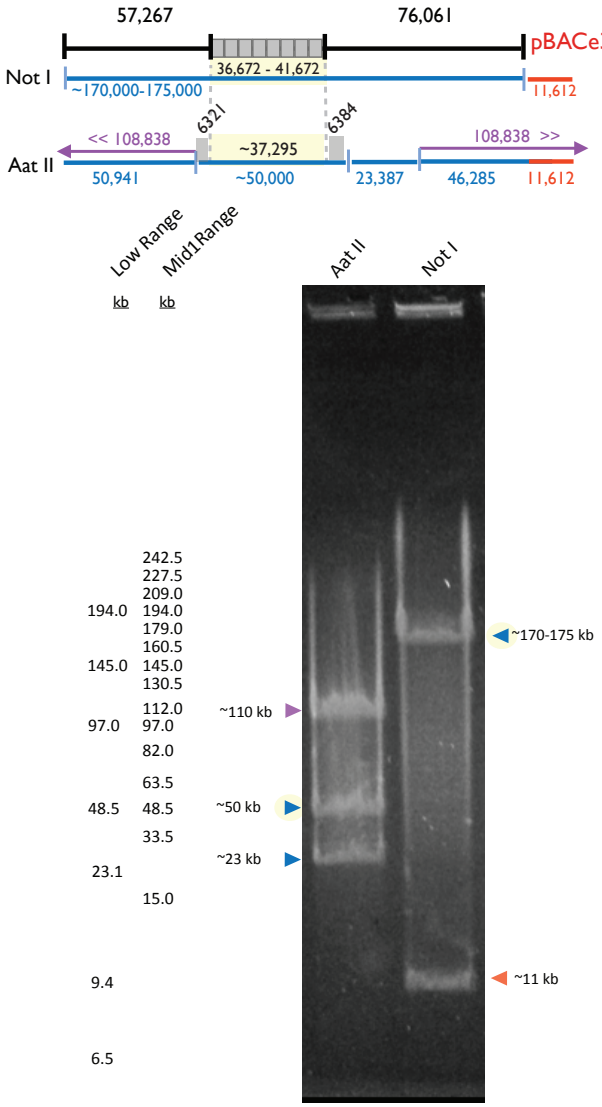


Fig. 22. Pulse-field gel electrophoresis of RP11-482A22 BAC DNA to determine insert length. Span of BAC end sequences relative to GRCh38 reference assembly provides an estimate of 57 kb to the right of the repeats and 76 kb to the left of the repeats (depicted in black). To determine the length of the repeats NotI and AatII digests were performed on RP11-482 DNA. The NotI digest isolates the insert DNA in its entirety from the cloning vector insert, pBACe3.6, providing evidence for a cloned insert in the range of 170-175 kb band (blue) and a 11.6 kb cloning vector band (red). After subtracting the known flanking region sizes this estimate provides a range of 36.7 - 41.7 kb repeat region, or 7.5-8.5 copies of the CT47 repeat. The AatII digest was expected to cut the BAC three times, as illustrated in the schematic, providing three resulting fragments: (a) 108 kb including the upstream flanking region (50kb), downstream flanking region (46 kb) and the cloning vector insert (11.6 kb), shown in purple; (b) a 23 kb region directly downstream from the repeat array (blue), and a region observed by PFGE to be ∼50 kb that spans the CT47 repeat cluster (providing evidence for a 37 kb repeat region after subtracting 12 kb of known flanking sequence, marked with grey shading). Regions providing evidence for repeat copy number are highlighted in yellow shading.

Bibliography

[1] Chaisson, M. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC bioinformatics* **13**, 238 (2012). URL http://www.biomedcentral.com/1471-2105/13/238/.

[2] Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM **00**, 3 (2013). URL http://arxiv.org/abs/1303.3997. 1303.3997.

[3] Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. URL https://github.com/lh3/bwa/blob/master/NEWS.md\#release-079-19-may-2014/.

[4] Frith, M. C., Wan, R. & Horton, P. Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic acids research* **38**, e100 (2010). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2853142\&tool=pmcentrez\&rendertype=abstract.

[5] Frith, M. C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. URL http://last.cbrc.jp/.

[6] Harris, R. S. *Improved pairwise alignment of genomic DNA*. Ph.D. thesis, The Pennsylvania State University (2007).

[7] Quick, J., Quinlan, A. & Loman, N. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *GigaScience* 1–6 (2014). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4226419/.

[8] Altschup, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **215**, 403–410 (1990).

[9] Benson, D. A. *et al.* GenBank. *Nucleic acids research* **41**, D36–42 (2013). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531190\&tool=pmcentrez\&rendertype=abstract.

[10] Quick, J. L. N. Bacterial whole-genome read data from the Oxford Nanopore Technologies MinION nanopore sequencer (2014). URL http://dx.doi.org/10.5524/100102.

[11] 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). URL http://link.springer.com/chapter/10.1007/978-1-4614-1347-9_14http://books.google.com/books?hl=en\&lr=\&id=R5P2GlJvigQC\&oi=fnd\&pg=PR9\&dq=Biological+Sequence+Analysis:+Probabilistic+Models+of+Proteins+and+Nucleic+Acids\&ots=hpBPoFnh6v\&sig=yGVckNE2kuie_3wkjtaYmODYVewhttp://books.google.com/books?hl=en\&lr=\&id=

R5P2GlJvigQC\&oi=fnd\&pg=PR9\&dq=Biological+sequence+analysis:+probabilistic+models+of+proteins+and+nucleic+acids\&ots=hpBPoFnh6D\&sig=rvNxCHHnjr2_F0sKOzRlSIMCei0.

[12] Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome research* **18**, 1814–28 (2008). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2577869\&tool=pmcentrez\&rendertype=abstract.

[13] Paten, B. *et al.* Cactus: Algorithms for genome multiple sequence alignment. *Genome research* **21**, 1512–28 (2011). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3166836\&tool=pmcentrez\&rendertype=abstract.

[14] Schwartz, A. S. & Pachter, L. Multiple alignment by sequence annealing. *Bioinformatics (Oxford, England)* **23**, e24–9 (2007). URL http://www.ncbi.nlm.nih.gov/pubmed/17237099.

[15] You, F., Huo, N., Deal, K. & Gu, Y. genome-wide SNP discovery in the large and complex Aegilops tauschii genome using next-generation sequencing without a reference genome sequence. *BMC genomics* **12**, 59 (2011). URL http://www.biomedcentral.com/1471-2164/12/59/.

[16] Holmes, I. & Bruno, W. J. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics* **17**, 803–820 (2001). URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/17.9.803.

[17] Elias, I. Settling the intractability of multiple alignment. *Journal of Computational Biology* **13**, 1323–1339 (2006). URL http://dx.doi.org/10.1089/cmb.2006.13.1323.

[18] Westesson, O., Lunter, G., Paten, B. & Holmes, I. Phylogenetic automata, pruning, and multiple alignment (2011). URL http://arxiv.org/abs/1103.4347. 1103.4347.

[19] Eddy, S. Profile hidden Markov models. *Bioinformatics* 755–763 (1998). URL http://bioinformatics.oxfordjournals.org/content/14/9/755.short.

[20] Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841–2 (2010). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2832824\&tool=pmcentrez\&rendertype=abstract.