UNIVERSITY OF CALIFORNIA
SANTA CRUZ

HIGH-COVERAGE LONG READ DNA SEQUENCING WITH THE OXFORD NANOPORE MINION

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Miten Jain

June 2017

The Dissertation of Miten Jain
is approved:

Professor Mark Akeson, Chair

Benedict Paten, Assistant Professor

Karen H. Miga, PhD

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

High-coverage Long Read DNA sequencing with the Oxford Nanopore MinION

by

Miten Jain

Nanopore sequencing was conceived in 1989 by Dave Deamer (UCSC). Over two decades of development from research laboratories and, later on, Oxford Nanopore Technologies resulted in the MinION nanopore sequencer. This work describes the developments in MinION nanopore sequencing and software, and technical milestones achieved since the MinION's release in 2014. These developments include establishing DNA reads that exceed 200 kb+ lengths and direct, simultaneous detection of nucleotide modifications in DNA and RNA. Due to their portability and real-time aspect, MinIONs are poised to become a routine tool for genomics and biology.

To the only factors in my life,

Mom and Dad,

I Love You.

Acknowledgments

Success is always a combination of great mentorship and guidance. I want to thank all the members of my committee for providing their time, input, and feedback over my graduate career. Their mentorship has helped me become a better scientist.

Prof. Mark Akeson has had a great influence on my scientific career, and everything I do in science from this point on will always be in part attributed to him. His mentorship and support, personally and professionally, means a lot to me and I intend to continue this association.

Dr. Benedict Paten has taught me a majority of the bioinformatics I know, and has been a great teacher and friend along the way. I aspire to be as good of a teacher and mentor as he is. His patience and support was extremely helpful in my learning process.

Dr. Karen Miga has been a wonderful mentor, research collaborator, and friend since the beginning of my graduate career at UCSC. Her impeccable scientific thinking and writing is something I intend to learn and improve upon with time. She has been helpful with all aspects of my graduate career, and I am extremely grateful for the same.

A special word of acknowledgement for Dr. Hugh Olsen, who has been an exceptional mentor and friend. I work extremely closely with him on a daily basis. His experience and guidance has been influential in my development as a scientist, and a person.

I want to thank and acknowledge Ariah Mackie, who has been instrumental in helping with analysis, writing, and editing. She has been a great friend and resource, and a lot of timelines would have been missed without her help, collaboration, and support.

I want to thank the entire nanopore group for their help and support. My fellow graduate students, Andrew Smith, Art Rand, Jordan Eizenga, Logan Mulroney, have been great friends and scientific collaborators on various projects. Jeff Nivala, Jordan Hughey, and Joey Dahl (previous nanopore group students) were supportive in my years as a graduate student.

I want to also thank the entire BME department at UCSC, particularly Prof. Kevin Karplus and Dr. David Bernick, who were helpful during my initial years as a graduate student.

I want to thank Dr. John Tyson from the University of British Columbia, who I collaborated with for *C. elegans* assembly work as well direct RNA sequencing work.

I want to thank the consortiums that I have been a part of: the MinION Analysis and Reference Consortium (MARC) and the nanopore human genome consortium.

I also want to acknowledge key collaborators from Oxford Nanopore Technologies, Daniel Garalde, David Stoddart, and Daniel Turner, for their support.

I want to thank my friends in the BME department at UCSC, especially Dr. Yulia Newton and Dorothy Oliver, who have been immensely helpful and supportive since our first year together as graduate students.

The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community

Miten Jain[1], Hugh E Olsen[1], Benedict Paten[1] and Mark Akeson[1]

[1]UC Santa Cruz Genomics Institute and Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, USA. Correspondence should be addressed to M.A. (makeson@soe.ucsc.edu).

Abstract

Nanopore DNA strand sequencing has emerged as a competitive, portable technology. Reads exceeding 150 kilobases have been achieved, as have in-field detection and analysis of clinical pathogens. We summarize key technical features of the Oxford Nanopore MinION, the dominant platform currently available. We then discuss

pioneering applications executed by the genomics community.

Introduction

Nanopore sequencing was pioneered by David Deamer at the University of California Santa Cruz, and by George Church and Daniel Branton (both at Harvard University). Beginning in the early 1990s, academic laboratories reached a series of milestones towards developing a functional nanopore sequencing platform (reviewed in [1, 2]). These milestones included the translocation of individual nucleic acid strands in single file order [3], processive enzymatic control of DNA at single-nucleotide precision [4], and the achievement of single-nucleotide resolution [5, 6].

Several companies have proposed nanopore-based sequencing strategies. These involve either: the excision of monomers from the DNA strand and their funneling, one-by-one, through a nanopore (NanoTag sequencing (Genia), Bayley Sequencing (Oxford Nanopore)); or strand sequencing wherein intact DNA is ratcheted through the nanopore base-by-base (Oxford Nanopore MinION). To date, only MinION-based strand sequencing has been successfully employed by independent genomics laboratories. Where possible, this review focuses on peer-reviewed research performed using the MinION [1, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38]

DNA strand sequencing using the Oxford Nanopore MinION

Oxford Nanopore Technologies (ONT) licensed core nanopore sequencing patents in 2007, and began a strand sequencing effort in 2010 [2]. At the Advances in Genome Biology and Technology (AGBT) 2012 conference, Clive Brown (Chief Technical Officer of ONT) unveiled the MinION nanopore DNA sequencer, which was subsequently released to early-access users in April 2014 through the MinION Access Program (MAP).

The MinION is a 90-g portable device. At its core is a flow cell bearing up to 2048 individually addressable nanopores that can be controlled in groups of 512 by an application-specific integrated circuit (ASIC). Prior to sequencing, adapters are ligated to both ends of genomic DNA or cDNA fragments (Fig. 0.1). These adapters facilitate strand capture and loading of a processive enzyme at the $5'$-end of one strand. The enzyme is required to ensure unidirectional single-nucleotide displacement along the strand at a millisecond time scale. The adapters also concentrate DNA substrates at the membrane surface proximal to the nanopore, boosting the DNA capture rate by several thousand-fold. In addition, the hairpin adapter permits contiguous sequencing of both strands of a duplex molecule by covalently attaching one strand to the other. Upon capture of a DNA molecule in the nanopore, the enzyme processes along one strand (the 'template read'). After the enzyme passes through the hairpin, this process repeats for the complementary strand (the 'complement read').

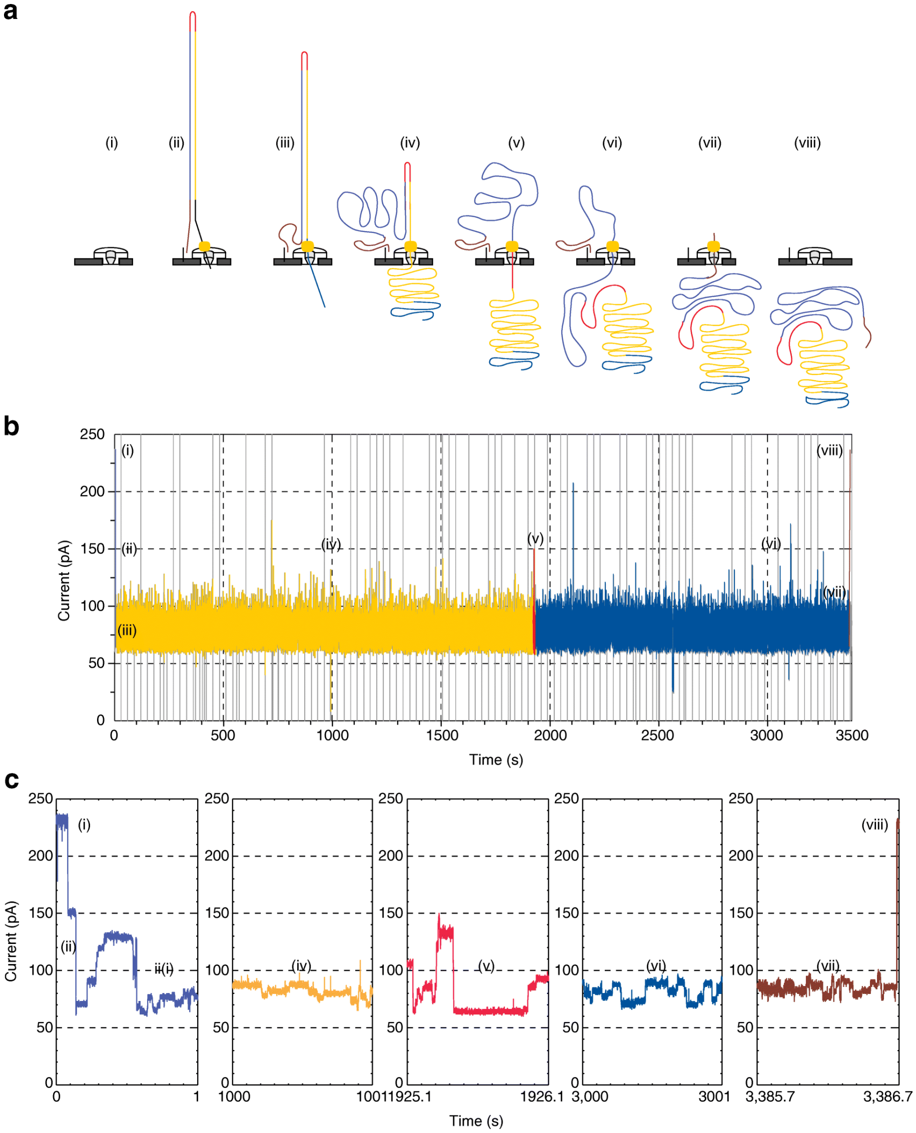


Figure 0.1: Data for a 2D read of a full-length λ phage dsDNA from the MinION nanopore sequencer. (a) Steps in DNA translocation through the nanopore: (i) open channel; (ii) dsDNA with lead adaptor (blue), bound molecular motor (orange) and hairpin adaptor (red) is captured by the nanopore; capture is 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 single 48-kb λ 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.

As the DNA passes through the pore, the sensor detects changes in ionic current caused by differences in the shifting nucleotide sequences occupying the pore. These ionic current changes are segmented as discrete events that have an associated duration, mean amplitude, and variance. This sequence of events is then interpreted computationally as a sequence of 3-6 nucleotide long kmers (‘words’) using graphical models. The information from template and complement reads is combined to produce a high-quality ‘2D read’, using a pairwise alignment of the event sequences.

An alternate library preparation method does not use the hairpin to connect the strands of a duplex molecule. Rather, the nanopore reads only one strand, which yields template reads. This allows for higher throughput from a flow cell, but the accuracy for these ‘1D reads’ is slightly lower than that of a ‘2D read’.

Benefits of MinION compared to other next generation sequencing platforms

Detection of base modifications

Next generation sequencing (NGS) technologies do not directly detect base modifications in native DNA. By contrast, single-molecule sequencing of native DNA and RNA with nanopore technology can detect modifications on individual nucleotides. Previously, Schreiber et al. [39] and Wescoe et al. [40] demonstrated that a single-channel nanopore system can discriminate among all five C-5 variants of cytosine (cytosine (C), 5-methylcytosine (5-mC), 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine

(5-fC), and 5-carboxylcytosine (5-caC)) in synthetic DNA. The discrimination accuracies ranged from 92 to 98% for a cytosine of interest in a background of known sequences [40].

In 2016, two research groups independently demonstrated that MinIONs can detect cytosine methylation in genomic DNA [41, 42]. Rand et al. [41] (see Chapter 3) developed a probabilistic method that combines a pair hidden Markov model (HMM) and a hierarchical Dirichlet process (HDP) mixture of normal distributions. They performed a three-way classification among C, 5-mC, and 5-hmC with a median accuracy of 80% in synthetic DNA [41]. Simpson et al. [42] performed a similar study in which they trained an HMM to perform a two-way classification among C and 5-mC, with 82% accuracy in human genomic DNA.

Real-time targeted sequencing

There are significant advantages to acquiring and analyzing DNA or RNA sequences in a few hours or less, especially for clinical applications. This is difficult using conventional NGS platforms, but relatively straightforward using the MinION because of its size, cost, simple library prep, and portability (see [14]). Beyond this, the MinION platform permits real-time analysis because individual DNA strands are translocated through the nanopore, allowing decisions to be made during the sequencing run.

This real-time utility of MinION was first demonstrated by Loose et al. [43] in a manuscript that described targeted enrichment ('Read Until') of 5 and 10 kb regions from phage lambda double-stranded DNA (dsDNA). Briefly, a mixture of DNA fragments is applied to the MinION flow cell. While a DNA strand is captured and processed in the nanopore, the resulting event levels are aligned against the expected pattern for a target sequence. If the pattern matches, the sequencing continues (Fig. 0.2a). If the pattern does not match, the DNA strand is ejected from the nanopore so that a subsequent DNA strand can be captured and analyzed (Fig. 0.2b). In doing this, reads of the targeted strand are rapidly accumulated relative to the DNA strand population as a whole. 'Read Until' demonstrates how MinION sequencing could significantly reduce the time required from biological sampling to data inference, which is pertinent for in-field and point-of-care clinical applications.

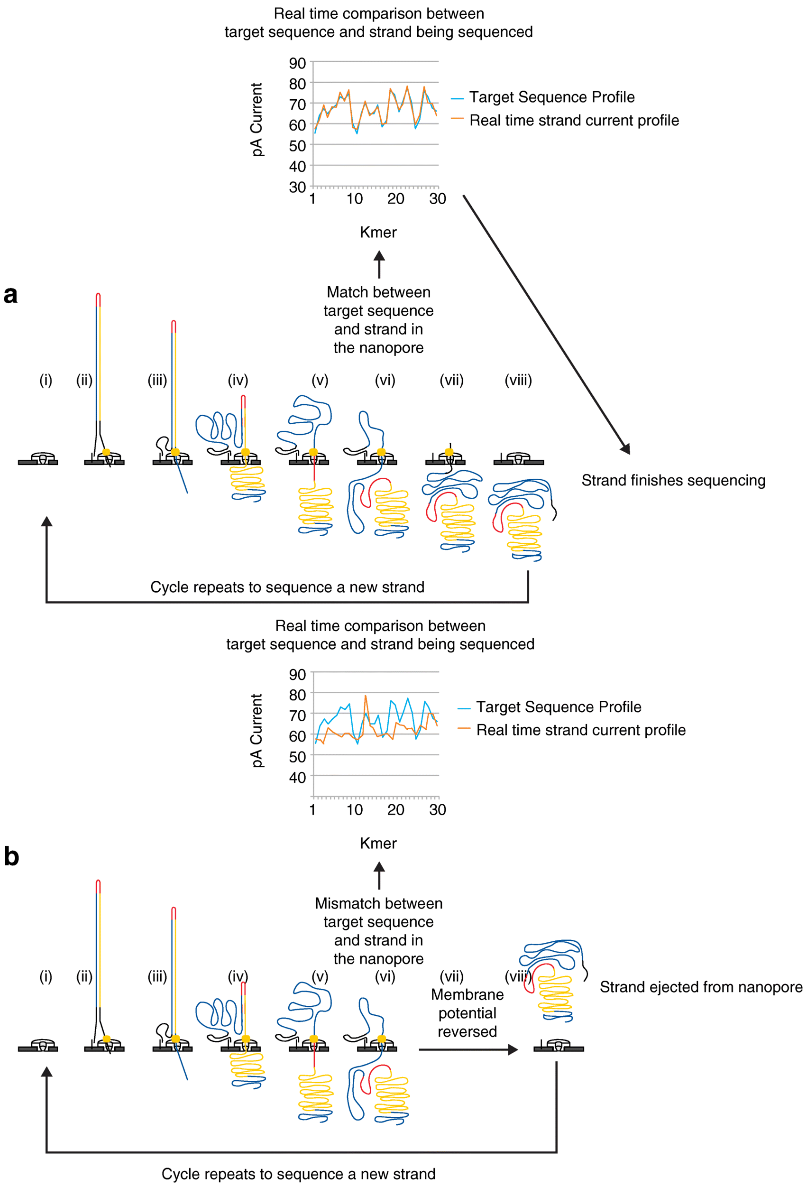

Real time comparison between
target sequence and strand being sequenced
pA Current
90
80
70
60
50
40
30
1
10
20
30
Kmer
Target Sequence Profile
Real time strand current profile
a
Match between
target sequence
and strand in
the nanopore
(i)
(ii)
(iii)
(iv)
(v)
(vi)
(vii)
(viii)
Strand finishes sequencing
Cycle repeats to sequence a new strand
Real time comparison between
target sequence and strand being sequenced
pA Current
90
80
70
60
50
40
1
10
20
30
Kmer
Target Sequence Profile
Real time strand current profile
b
Mismatch between
target sequence
and strand in
the nanopore
(i)
(ii)
(iii)
(iv)
(v)
(vi)
(vii)
Membrane
potential
reversed
(viii)
Strand ejected from nanopore
Cycle repeats to sequence a new strand

Figure 0.2: ‘Read Until’ strategy for selective sequencing of dsDNA molecules. The ionic current profile obtained during translocation of a DNA strand through the nanopore is compared in real time to the ionic current profile of a target sequence. a As sequencing of the template strand of DNA proceeds (during step iv), the measured current is compared to the reference current profile. If there is a match, sequencing of that strand continues to completion (steps v-vii). A new strand can now be captured. b Alternatively, if the measured current does not match the reference current profile, the membrane potential is reversed, sequencing of that strand stops, and the strand is ejected (at stage v). A new strand can now be captured. (Image based on the strategy of Loose et al. [43]).

Extending read lengths using the MinION

A virtue of nanopore DNA strand sequencing is read lengths that substantially exceed those of dominant NGS platforms. For example, 1D reads over 300 kb in length and 2D reads up to 60 kb in length have been achieved using *Escherichia coli* genomic DNA [44]. To demonstrate utility, Jain et al. [9] (see Chapter 1) used 36-kb+ MinION reads to resolve a putative 50-kb gap in the human Xq24 reference sequence. Previously, this gap in the reference sequence could not be completed because it contained a series of 4.8-kb-long tandem repeats of the cancer-testis gene CT47. This work established eight CT47 repeats in this region (Fig. 0.3).

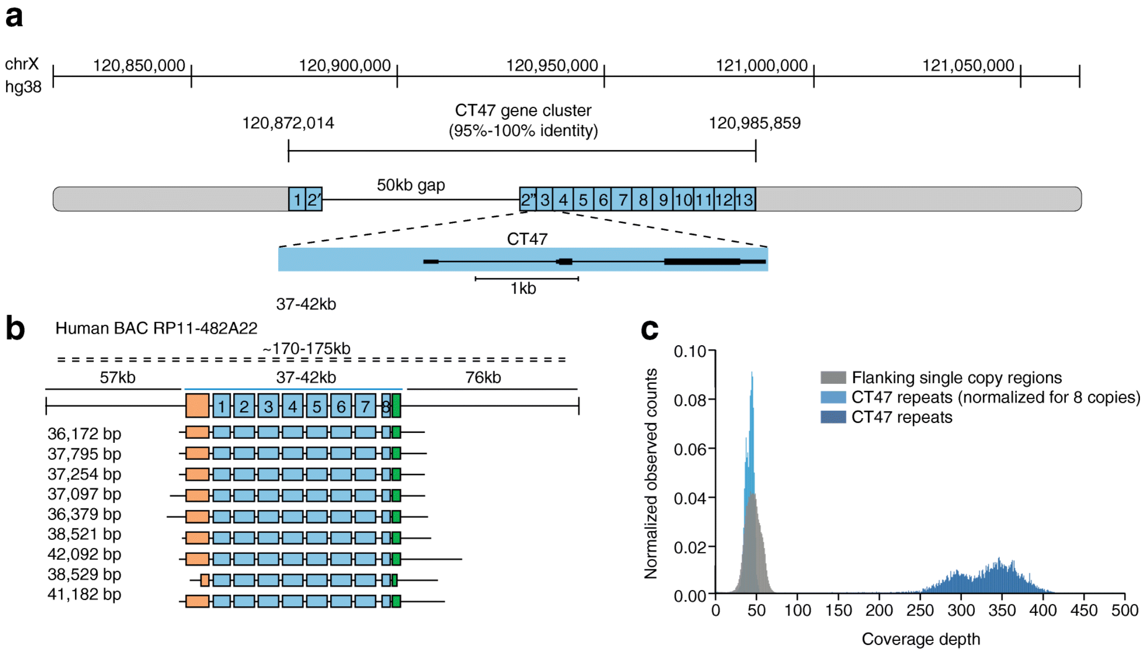


Figure 0.3: Estimate CT47-repeat copy-number on human chromosome Xq24. (a) BAC end sequence alignments (RP11-482A22: AQ630638 and AZ517599) span a 247-kb region, including 13 annotated CT47 genes [45] (each within a 4.8-kb tandem repeat), and a 50-kb scaffold gap in the GRCh38/hg38 reference assembly. (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. The insert (dashed line), whose size is estimated from pulse-field gel electrophoresis, 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 of 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 of 329.3-base coverage) and was similar to single-copy estimates when normalized for eight copies (light blue distribution; mean of 41.15-base coverage). (Figure reproduced from Jain et al. [9]).

Detection of structural variants

Mistakes arising in assemblies of 450-base-long NGS reads are also problematic when characterizing structural variants in human genomes. The problem is acute in cancer, where examples of copy number variants, gene duplications, deletions, insertions, inversions, and translocations are common. For reads that averaged 8 kb in length, Norris et al. [46] used the MinION to detect structural variants in a pancreatic cancer cell line. These authors concluded that the MinION allowed for reliable detection of structural variants with only a few hundred reads compared to the millions of reads typically required when using NGS platforms.

RNA expression analysis

RNA expression analysis is most often performed by NGS sequencing of cDNA copies. A drawback of this strategy is that the reads are relatively short, thus requiring assembly of cDNA reads into full-length transcripts. This is an issue for the accurate characterization of RNA splice isoforms because there is often insufficient information to deconvolute the different transcripts properly. Full-length cDNA reads would avoid this problem and can be executed with either the PacBio or MinION platforms.

To illustrate, Bolisetty et al. [8] used the MinION to determine RNA splice variants and to detect isoforms for four genes in Drosophila. Among these is Dscam1, the most complex alternatively spliced gene known in nature, with 18,612 possible isoforms ranging in length from 1806 bp to 1860 bp [8]. They detected over 7000 isoforms for

Dscam1 with >90% alignment identity. Identifying these isoforms would be impossible with 450-base-long NGS reads.

Bioinformatics and platform advances

The first manuscript to discuss MinION performance was based on limited data and ill-suited analysis, and thus yielded misleading conclusions about the platform's performance [24]. Over the subsequent 9-month period, ONT optimized MinION sequencing chemistry and base-calling software. Combined with new MinION-specific bioinformatics tools (Table 0.1), these refinements improved the identity of sequenced reads, that is, the proportion of bases in a sequencing 'read' that align to a matching base in a reference sequence, from a reported 66% in June 2014 [9] to 92% in March 2015 [44]. Links to these tools are provided in Table 0.1 and highlighted in the sections that follow.

Table 0.1: Software tools developed specifically for MinION sequence data; there are existing tools that can also be made to work with nanopore data (not shown)

Name	**Applications**	**Link**
Poretools [22]	Sequence data extraction and statistics	`https://github.com/arq5x/poretools`
poRe [37]	Sequence extraction and basic statistics	`https://sourceforge.net/projects/rpore/`
BWA MEM [47]	Sequence alignment	`https://github.com/lh3/bwa`
LAST [48]	Sequence alignment	`http://last.cbrc.jp/`

NanoOK [20]	Sequence alignment, statistics, and visualization	https://documentation.tgac.ac.uk/display/NANOOK/
marginAlign [9]	Sequence alignment, SNV calling, and statistics	https://github.com/benedictpaten/marginAlign
Nanopolish [49]	Signal alignment and SNV calling	https://github.com/jts/nanopolish
GraphMap [12]	Sequence alignment and SNV calling	https://github.com/isovic/graphmap
minimap	Fast approximate mapping	https://github.com/lh3/minimap
miniasm	De novo assembly	https://github.com/lh3/miniasm
CANU [50]	De novo assembly	https://github.com/marbl/canu
Nanocorrect [49]	De novo assembly	https://github.com/jts/nanocorrect
PoreSeq [51]	De novo assembly and SNV calling	https://github.com/tszalay/poreseq
NaS [23]	De novo assembly	https://github.com/institut-de-genomique/NaS
Nanocorr [13]	De novo assembly	https://github.com/jgurtowski/nanocorr

Mash [52]	Species identification and fast approximate alignments	`https://github.com/marbl/mash`
minoTour [53]	Real-time data analysis	`https://github.com/minoTour/minoTour`
Read Until [43]	Selective sequencing	`https://github.com/mattloose/RUscripts`
Nanocall [54]	Local base-calling	`https://github.com/mateidavid/nanocall`
DeepNano [55]	Recurrent neural network (RNN)-based base-calling	`https://bitbucket.org/vboza/deepnano`

SNV single nucleotide variant

De novo base-calling

The base-calling for MinION data is performed using HMM-based methods by Metrichor, a cloud-based computing service provided by ONT. Metrichor presently requires an active internet connection [54, 55] and is a closed source. However, its base-calling source code is now available to registered MinION users under a developer license. To create a fully open-source alternative, earlier in 2016, two groups independently developed base-callers for MinION data. Nanocall [54] is an HMM-based base-caller that performs efficient 1D base-calling locally without requiring an internet connection at accuracies comparable to Metrichor-based 1D base-calling. DeepNano

[55], a recurrent neural network framework, performs base-calling and yields better accuracies than HMM-based methods. Being able to perform local, offline base-calling is useful when performing in-field sequencing with limited internet connectivity [30].

Sequence alignment

When the MAP began, the first attempts at aligning MinION reads to reference sequences used conventional alignment programs. Most of these are designed for short-read technologies, such as the 250-nucleotide highly accurate reads produced by the Illumina platform. Not surprisingly, when applied to lower accuracy 10-kb MinION reads, these aligners disagreed in their measurement of read identity and sources of error, despite parameter optimization (Fig. 0.4). MarginAlign was developed to improve alignments of MinION reads to a reference genome by better estimating the sources of error in MinION reads [9]. This expectation-maximization-based approach considerably improves mapping accuracy, as assayed by improvements in variant calling, and yielded a maximum likelihood estimate of the insertion, deletion, and substitution errors of the reads (Fig. 0.4). This was later used by a MAP consortium to achieve a 92% read accuracy for the E. coli k12 MG1655 genome [44].

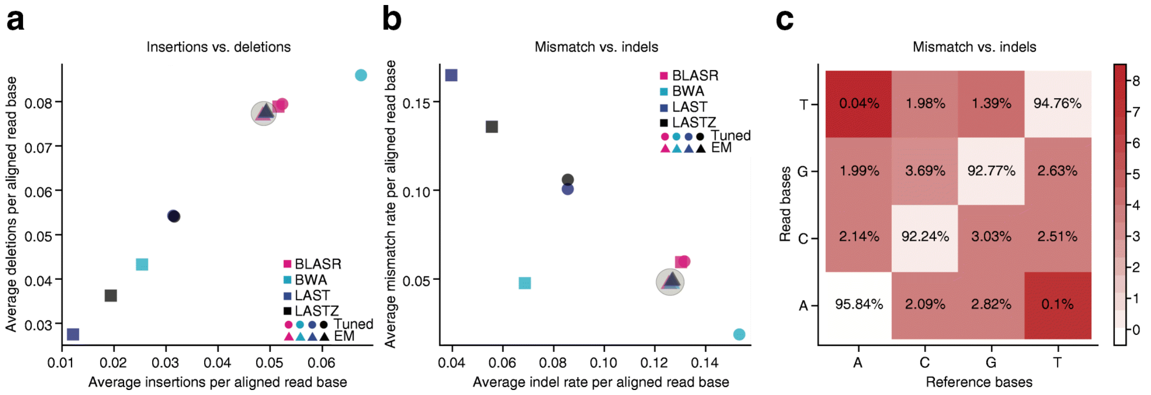


Figure 0.4: Maximum-likelihood alignment parameters derived using expectation-maximization (EM). The process starts with four guide alignments, each generated with a different mapper using tuned parameters. Squares denote error estimates derived from different mappers when used without tuning; circles denote error estimates post-tuning; and triangles denote error estimates post-EM. (a) Insertion versus deletion rates, expressed as events per aligned base. (b) Indel events per aligned base versus rate of mismatch per aligned base. Rates varied strongly between different guide alignments; but 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. (Figure reproduced from Jain et al. [9]).

MarginAlign refines alignments generated by a mapping program, such as

LAST [48] or BWA mem [47], and is therefore reliant on the accuracy of the initial alignment. GraphMap [12] is a read mapper that employs heuristics that are optimized for longer reads and higher error rates. In their study, Sovic et al. [12] demonstrated that GraphMap had high sensitivity (comparable to that of BLAST) and that GraphMap's estimates of error rates were in close agreement with those of marginAlign.

De novo assembly

The current error profile of MinION reads makes them largely unsuitable for use with de novo assembly methods that are designed for short reads, such as de Bruijn graph-based methods. This is principally for two reasons. First, these methods rely on a sufficient fraction of all possible k-mers sequenced being reconstructed accurately; the overall indel and substitution error rates produced by MinION are unlikely to meet this demand. Second, de Bruijn graphs, in their structure, do not exploit the longer-read information generated by the MinION. Instead, nanopore sequencing is helping to mark a return to overlap-consensus assembly methods [49], a renaissance that largely started with the earlier advent of SMRT sequencing [56]. Overlap-consensus methods were principally developed for lower-error-rate Sanger-based sequencing, and so novel strategies are required to error correct the reads before they are assembled. The first group to demonstrate this approach achieved a single contig assembly of the E. coli K-12 MG1655 genome at 99.5% base level accuracy using only MinION data [49]. Their pipeline, 'nanocorrect', corrected errors by first aligning reads using the graph-based, greedy partial order aligner method [57], and then by pruning errors that were apparent given the alignment graph. The error-corrected reads were then assembled using the Celera Assembler. This draft assembly was then further improved using Loman and co-worker's polishing algorithm, 'nanopolish' [49].

Single-nucleotide variant calling

Reference allele bias, the tendency to over-report the presence of the reference allele and under-report non-reference alleles, becomes more acute when the error rate of the reads is higher, because non-reference variants are more likely to be lost in noisy alignments. To overcome this problem for MinION reads, several academic laboratories have developed MinION-specific variant calling tools.

The marginCaller module in marginAlign [9] uses maximum-likelihood parameter estimates and marginalization over multiple possible read alignments to call single nucleotide variants (SNVs). At a substitution rate of 1% (in silico), marginCaller detected SNVs with 97% precision and 97% recall at 60x coverage. Similarly, by optimizing read level alignments, Sovic et al. [12] used their GraphMap approach, for accurate mapping at high identity, to detect heterozygous variants from difficult-to-analyze regions of the human genome with over 96% precision. They also used in silico tests to demonstrate that GraphMap could detect structural variants (insertions and deletions of different lengths) with high precision and recall.

Nanopolish [49] uses event-level alignments to a reference for variant calling. This algorithm iteratively modifies the starting reference sequence to create a consensus of the reads by evaluating the likelihood of observing a series of ionic current signals given the reference nucleotide sequence. At each iteration, candidate modifications to the consensus sequence are made and the sequence with the highest likelihood is chosen. At termination of iteration, the alignment of the final consensus to the final reference sequence defines the variants (differences) between the reads and the reference. This approach was used to demonstrate the feasibility of real-time surveillance as part of a study in West Africa in which Quick et al. [30] identified ebola virus sub-lineages using the MinION with ∼80% mean accuracy.

PoreSeq [51] is a similar algorithm to Nanopolish, published around the same time, that also iteratively maximizes the likelihood of observing the sequence given a model. Their model, which like Nanopolish uses MinION event-level data, accounts for the uncertainty that can arise during the traversal of DNA through the nanopore. PoreSeq can achieve high precision and recall SNV-calling at low coverages of sequence data. Using a 1% substitution rate in the M13 genome, Szalay and Golovchenko [51] demonstrated that PoreSeq could detect variants with a precision and recall of 99% using 16x coverage. This is around the same accuracy as marginAlign on the same data, but at a substantially lower coverage, demonstrating the power of the event-level, iterative approach.

Consensus sequencing for high accuracy

The read accuracy of 92% currently achieved by MinION is useful for some applications, but at low coverage it is insufficient for applications such as haplotype phasing and SNV detection in human samples, where the number of variants to be detected is smaller than the published variant-detection error rates of algorithms using MinION data. One method previously used to improve the quality of single-molecule sequence employed rolling circle amplification [56]. In a parallel method for the MinION, Li et al. [58] used rolling circle amplification to generate multiple copies of the 16S ribosomal RNA (rRNA) gene in one contiguous strand. MinION nanopore sequencing of each contiguous strand gave a consensus accuracy of over 97%. This allowed sensitive profiling in a mixture of ten 16S rRNA genes.

Current applications of the MinION

Analysis of infectious agents at point-of-care

Next-generation sequencing can detect viruses, bacteria, and parasites present in clinical samples and in a hospital environment [11, 14, 27, 34]. These pathogen sequences enable the identification and surveillance of host adaptation, diagnostic targets, response to vaccines, and pathogen evolution [30]. MinIONs are a new tool in this area that provide substantial advantages in read length, portability, and time to pathogen identification, which is documented to be as little as six hours from sample collection [14]. Pathogen identification can be performed in as little as four minutes once the sample is loaded on the MinION [14]. The breadth of clinical applications demonstrated to date include studies of chikungunya virus [14], hepatitis virus C [14], *Salmonella enterica* [28], and *Salmonella typhimurium* [7], as well as work on antibiotic resistance genes in five Gram-negative isolates and on the mecA gene in a methicillin-resistant

Staphylococcus aureus (MRSA) isolate [17].

Arguably, the most inspired clinical use of the MinION to date involved teams of African and European scientists who analyzed Ebola samples on-site in West Africa [30, 59]. The recent viral epidemic was responsible for over 28,599 Ebola cases and more than 11,299 deaths [60]. In the larger of the two studies, Quick and colleagues [30] transported a MinION field sequencing kit (weighing <50 kg, and fitting within standard suitcases) by commercial airline to West Africa. Once there, they sequenced blood samples from 142 Ebola patients in a field laboratory. Ebola virus sequence data were generated within 24 h after sample delivery, with confirmation of Ebola sequences taking as little as 15 min of MinION run time. To our knowledge, these studies by Quick et al. [30] and by Hoenen et al. [59] are the first applications of any sequencing device for real-time on-site monitoring of an epidemic.

Teaching and citizen science

The low cost of entry and portability of the MinION sequencer also make it a useful tool for teaching. It has been used to provide hands-on experience to undergraduate students as part of a recently taught course at Columbia University [61] and to teach graduate students at the University of California Santa Cruz. Every student was able to perform their own MinION sequencing. Similarly, the short and simple process of preparing a sequencing library allowed researchers at Mount Desert Island Biological Laboratory in Maine to train high school students during a summer course and have them run their own MinION experiments. Their Citizen Science initiative intends to address questions pertaining to health and environment that would otherwise be implausible [62].

Aneuploidy detection

One of the immediate applications of the MinION is aneuploidy detection in prenatal samples. The typical turnaround time for aneuploidy detection in such samples is 1-3 weeks when using NGS platforms [63]. Wei and Williams [38] used the MinION to detect aneuploidy in prenatal and miscarriage samples in under 4 h. They concluded that the MinION can be used for aneuploidy detection in a clinical setting.

MinIONs in space

At present, it is hard to detect and identify bacteria and viruses on manned space flights. Most of these analyses, along with understanding the effects of space travel on genomes, occur when the samples are brought back to Earth. As a first step to resolve this shortcoming, NASA plans to test MinION-based real-time sequencing and pathogen identification on the International Space Station (ISS) [64, 65]. In a proof-of-concept experiment, Castro-Wallace et al. [66] demonstrated successful sequencing and de novo assembly of a lambda phage genome, an *E. coli* genome, and a mouse mitochondrial genome. They noted that there was no significant difference in the quality of sequence data generated on the ISS and in control experiments that were performed in parallel on Earth [66].

Outlook

PromethION

The MinION allows individual laboratories to perform sequencing and subsequent biological analyses, but there is a part of the research community that is interested in high-throughput sequencing and genomics. Realizing this need, ONT has developed a bench-top instrument, PromethION, that is projected to provide high-throughput and is modular in design. Briefly, it will contain 48 flow cells that could be run individually or in parallel. The PromethION flow cells contain 3000 channels each, and are projected to produce up to 6 Tb of sequencing data each day. This equates to over 60 human genomes per day at 30x coverage.

Read accuracy

Single read accuracy is 92% for the current MinION device [44], which is often sufficient for applications such as the identification of pathogens or mRNA (cDNA) splice variants. However, some medical applications, such as the detection of individual nucleotide substitutions or base adducts in a single mitochondrial genome, would require read accuracies exceeding 99.99%. Given prior experience, it is reasonable that ONT will continue to improve their chemistry and base-calling software. Nevertheless, it is probable that Q40 nanopore sequencing will entail a single strand re-read strategy [2].

As is true for all sequencing platforms, MinION's base-call accuracy is improved using consensus-based methods. For example, for an *E. coli* strain where single reads averaged ∼80% accuracy, consensus accuracy improved to 99.5% at 30x coverage [49]. The remaining 0.5% error appears to be non-random. This improvement is in part due to the inability of the present MinION platform to resolve homopolymers longer than the nanopore reading head (six nucleotides), and to the absence of training in the detection of base modifications. It is plausible that resolving these two issues will push nanopore consensus accuracy to ≥99.99%.

Read length

With the advent of single-molecule sequencing technologies (PacBio and MinION), the average read lengths increased from 250 nucleotides to 10 kb. More recently, reads of more than 150 kb have routinely been achieved with the MinION (Akeson, unpublished findings), and this is expected to improve in the next few months. Achieving long reads will allow progress in understanding highly complex and repetitive regions in genomes that are otherwise hard to resolve.

Direct RNA sequencing

Sequencing of direct RNA with nanopore technology is an active area of development at ONT and in academic research groups. Single-molecule detection of tRNA has been previously demonstrated in single-channel and solid-state nanopores [67, 68]. Nanopore sensing can also detect nucleotide modifications in both DNA [39, 40, 41, 42]] and tRNA [69]. Direct RNA sequencing will reveal insights in RNA biology that presently can get lost due to issues with reverse transcription and PCR amplification.

Single-molecule protein sensing

At present, mass spectrometry is the preferred technique for performing a comprehensive proteomics analysis [70], but there are limitations to the sensitivity, accuracy, and resolution of any one analytical technique [70]. In 2013, Nivala et al. [71] demonstrated enzyme-mediated translocation of proteins through a single-channel nanopore. Their study showed that sequence-specific features of the proteins could be detected. They then engineered five protein constructs bearing different mutations and rearrangements, and demonstrated that these constructs could be discriminated with accuracies ranging from 86 to 99%. Protein sequencing will allow studies of complex interactions among cells in different tissues [72].

Conclusions

Nanopore DNA strand sequencing is now an established technology. In the short interval since the ONT MinION was first released, performance has improved rapidly, and the technology now routinely achieves read lengths of 50 kb and more and single-strand read accuracies of better than 92%. Improvement in read lengths, base-call accuracies, base modification detection, and throughput is likely to continue. Owing to its portability, the MinION nanopore sequencer has proven utility at the point-of-care in challenging field environments. Further miniaturization of the platform (SmidgION) and associated library preparation tools (Zumbador, VolTRAX) promise an age of ubiquitous sequencing. Parallel applications, including direct RNA sequencing, are on the horizon.

Acknowledgements

The authors thank the National Human Genome Research Institute of the US National Institutes of Health for funding their research under award numbers HG006321 (MA), HG007827 (MA), and U54HG007990 (BP). The authors also thank Ariah Mackie for proofreading the manuscript.

Contributions

All authors contributed to the writing, editing, and completion of the manuscript. All authors read and approved the final manuscript.

Chapter 1

Improved data analysis for the MinION nanopore sequencer

Miten Jain[1,2], Ian T Fiddes[1,2], Karen H Miga[1,2], Hugh E Olsen[1,2], Benedict Paten[1,2] & Mark Akeson[1,2]

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

Abstract

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. Subsequently, we 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.

Introduction

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 long read lengths to help overcome some of the drawbacks of short-read sequencing. As part of the MinION Access Program (MAP), we evaluated the sequencing platform and then developed it to call single-nucleotide variants (SNVs) and to resolve the repeat structure of highly repetitive regions. Our open-source analysis tools are available online (Software 1 and 2; `https://github.com/mitenjain/nanopore` and `https://github.com/benedictpaten/marginAlign` 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. Metrichor, a cloud-based software provide by ONT, calls DNA bases by using 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 (see 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 (see Methods). A DNA construct analyzed on the MinION (Fig. 1.1) 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.

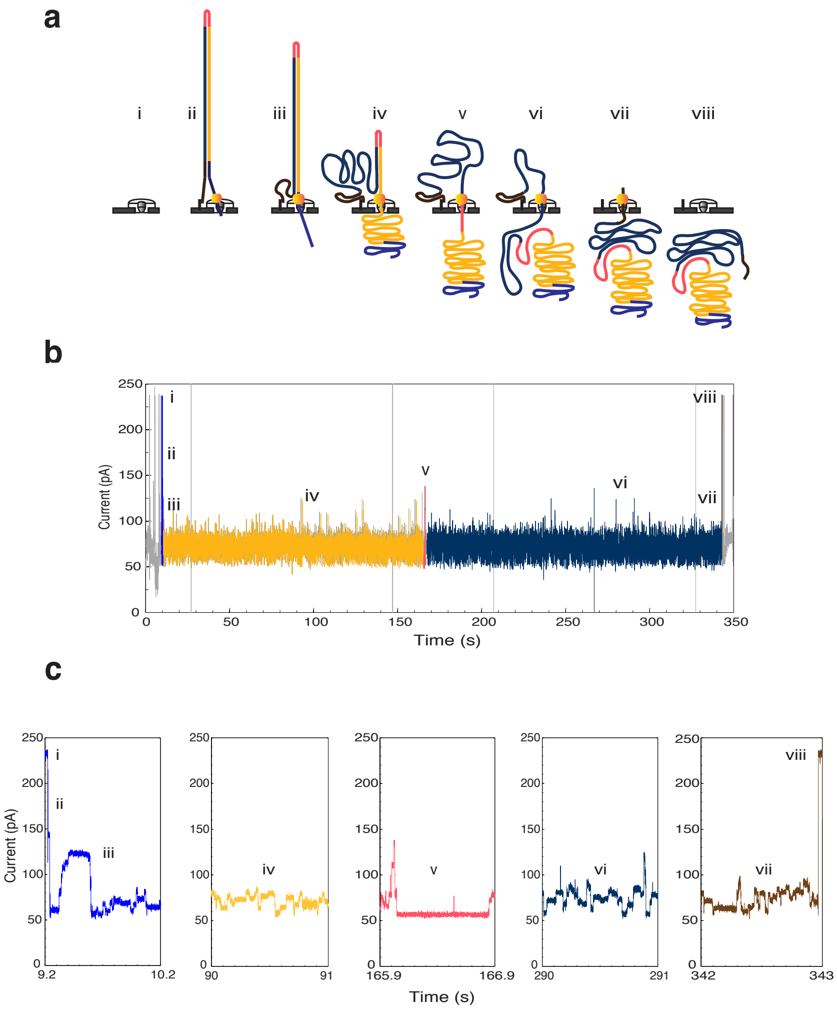


Figure 1.1: Molecular events and ionic current trace for a 2D read of a 7.25 kb M13 phage dsDNA molecule. (a) Steps in DNA translocation through the nanopore: (i) Open channel; (ii) , bound molecular motor (orange) and hairpin adaptor (red) is captured by the nanopore; capture is followed by translocation 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.

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. 1.1). 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 the 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 (Fig. 1.2). The average observed identity (the proportion of bases in a read that align to a matching base in a reference sequence) for 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.

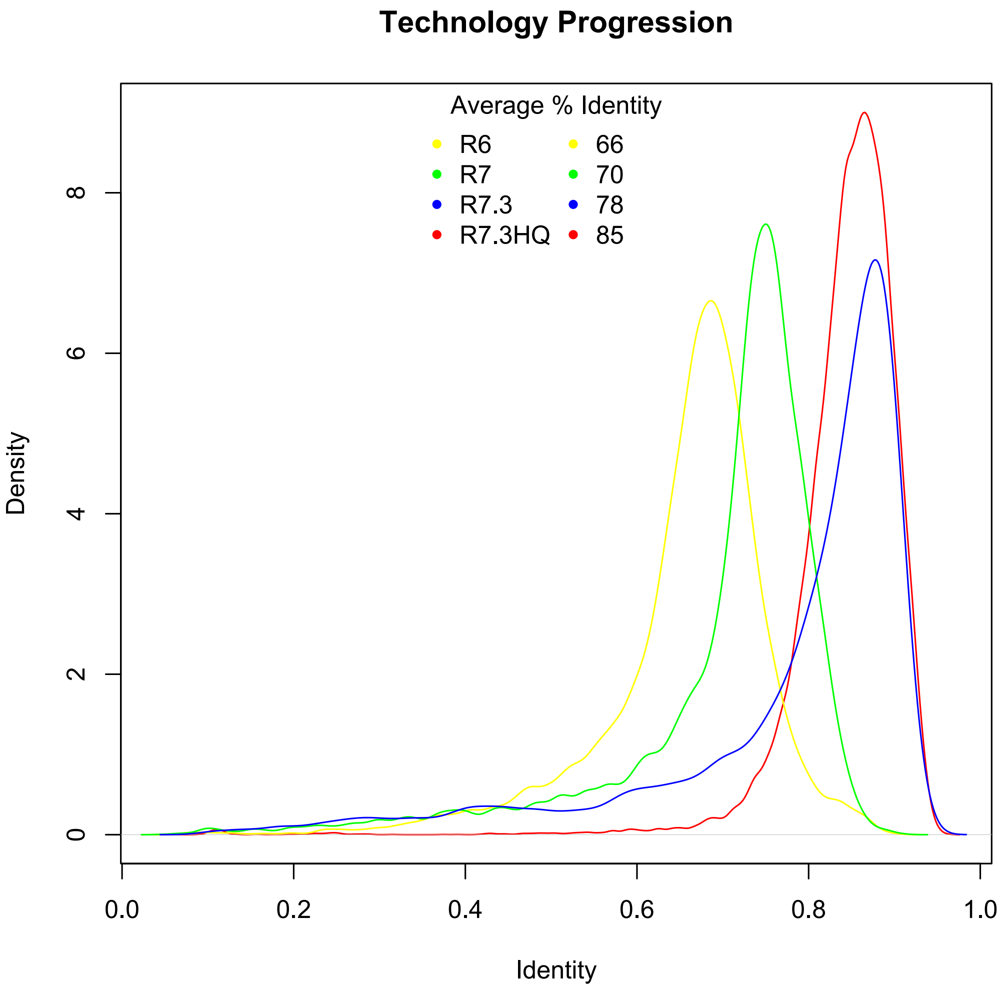


Figure 1.2: MinION technology progression. Progression of read identity distributions with MinION versions since June 2014.

MinION throughput

We sequenced intact replicative-form M13 phage dsDNA using three MinION flow cells that contained 337-473 functional channels (see Methods). Reads were characterized as 'template', 'complement', or '2D'. '2D' represented 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 (Table 1.1). 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.

Table 1.1: 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

Establishing a mapping pipeline for MinION reads

To evaluate the quality of these MinION reads, we experimented with four different alignment programs [73, 74, 75, 76] (see Methods). Each was run with its default parameters and with tuned parameters that were selected either by experimentation or by expert advice from other MAP participants (Table 1.2).

Table 1.2: 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$^{\text{TM}}$ long reads
LAST	-s 2 -T 0 -Q 0 -r 1 -a 1 -b 1 -q 1	Quick *et al* [77], MAP participants
LASTZ	–hspthresh=1800 –gap=100,100	Oxford Nanopore

The proportion of reads that mapped to reference sequences (M13 or ONT phage λ DNA control) varied by aligner (Fig. 1.3). LAST [75] with tuned parameters was the most inclusive alignment program, and stringency analysis indicated that few of its alignments were false positives (Fig. 1.4). For data pooled from the three M13 experiments, tuned LAST mapped 95.26% of template, 98.31% of complement, and 98.96% of 2D reads. Most unmapped reads were homologous to *E. coli*, indicating minor contamination [78, 79] (see Methods, Fig. 1.5a-c, Table 1.3).

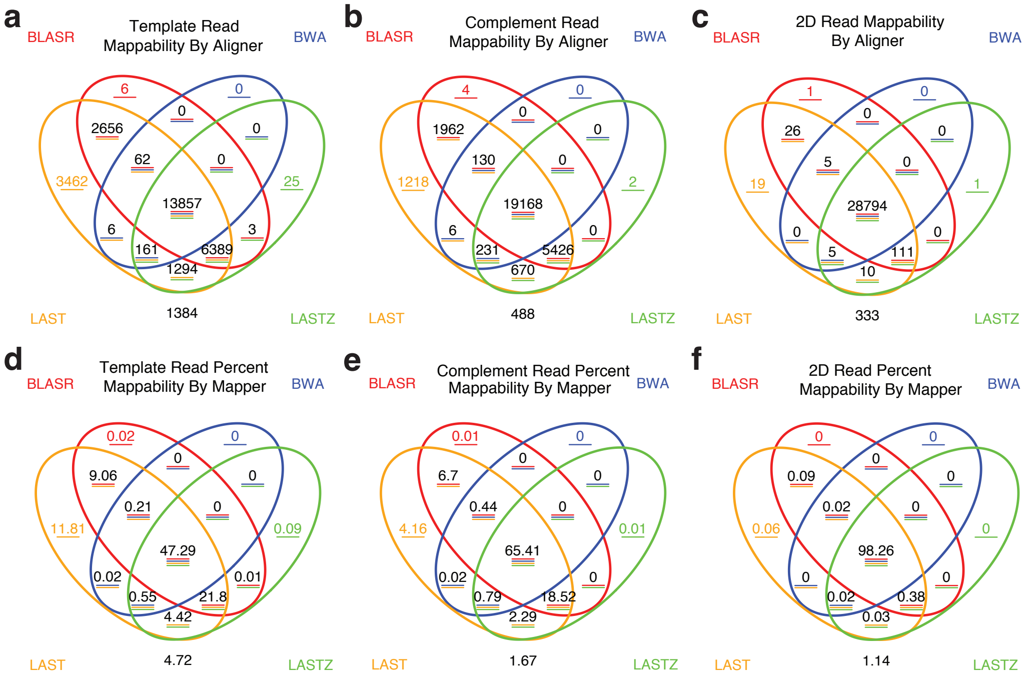


Figure 1.3: Venn diagram representing read mappability for MinION reads across three replicate M13 experiments using R7.3 chemistry.
Mappability represents the proportion of reads that can be aligned to either the M13 or the phage λ DNA control using the tuned parameters for each mapper. In our analysis, 2D reads had the highest mappability, with 99% of reads being mappable, followed by complement and template reads, with 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 greatest proportion of mappable reads on its own.

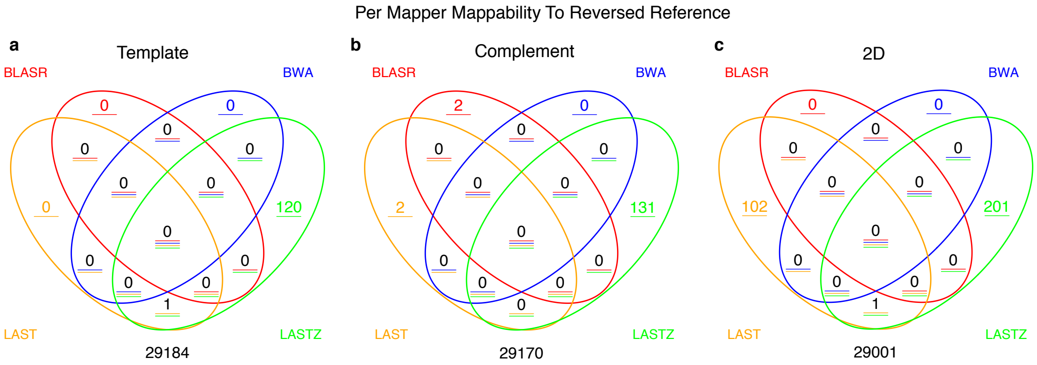


Figure 1.4: Venn diagram representing read mappability to a reversed reference for MinION reads from three replicate M13 experiments using R7.3 chemistry. Because the reference was reversed, effectively no reads should map; this is thus a proxy measure of specificity. Results were obtained using the tuned alignment parameters.

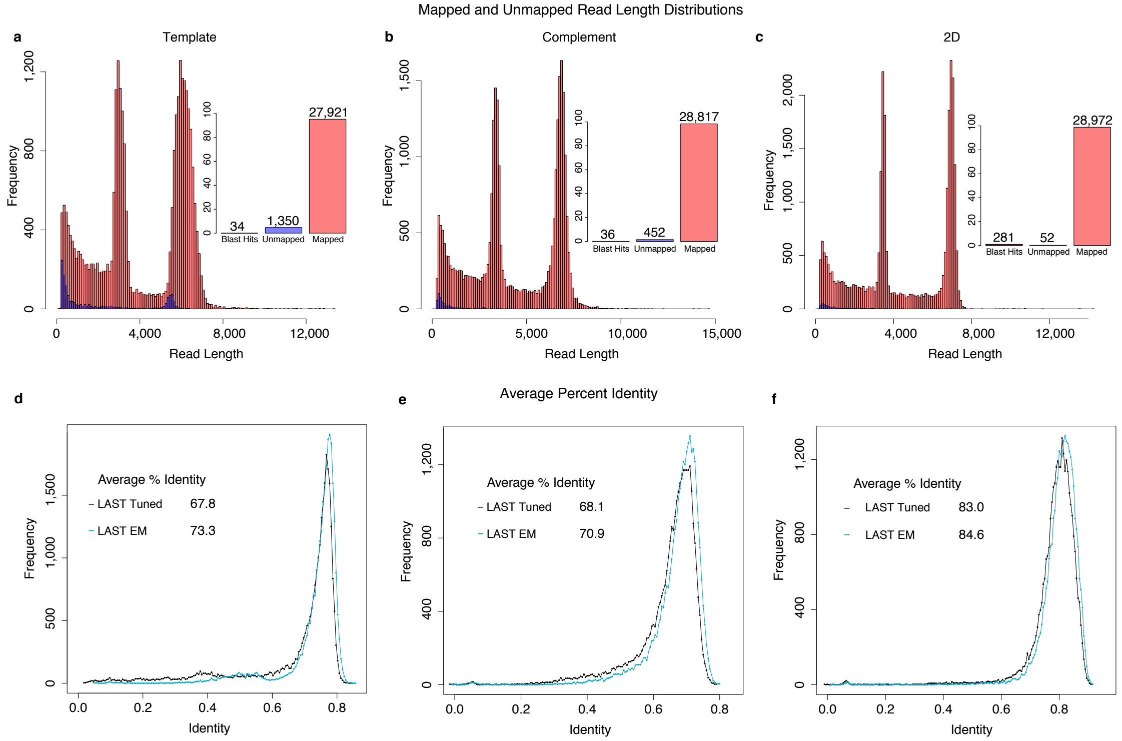


Figure 1.5: Read-length distributions and identity plots for M13. (a-c) Read-length histograms for mapped versus unmapped reads across three replicate M13 experiments for (a) template, (b) complement and (c) 2D reads. Most reads mapped to phage λ DNA control or M13 reference sequences (peaks at 3.8 kb 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.

Table 1.3: BLAST hits of unmapped reads

Sequence Name	**Counts**
2D reads unmapped by any mapper	
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

Complement reads unmapped by any mapper

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

Template reads unmapped by any mapper

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

We observed two distinct peaks for reads, one at about 7.2 kb, corresponding to full-length M13 DNA, and one at 3.8 kb, corresponding to the ONT λ phage DNA control (Fig. 1.5a-c). A large number of reads spanned the full length of the M13 genome, whereas unmappable reads made up a small proportion (<0.2% of all 2D reads) and were generally shorter than mappable reads.

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. 1.6a-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 (see Methods, [80]). Using expectation maximization to train an HMM Fig. 1.7 and alignment-banding heuristics for efficiency [81], we obtained robust convergence of parameter MLEs across all replicate experiments, guide alignments, and random starting parameterizations (Fig 1.6a-b, Fig. 1.8).

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 (Fig. 1.9). Substitutions varied from 0.21 (for template reads) to 0.05 (for 2D reads) events per aligned base (Figs. 1.6c, 1.10, and 1.11). Substitutions 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 (see Supplementary Note 1).

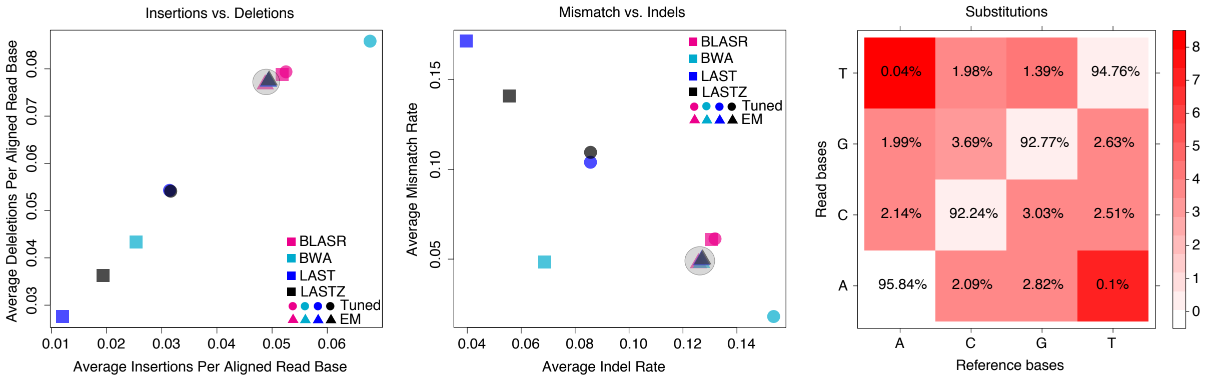


Figure 1.6: Maximum-likelihood alignment parameters derived using expectation maximization (EM). The process starts with four guide alignments, each generated with a different mapper using tuned parameters. (a) Insertion versus deletion rates, expressed as events per aligned base. (b) Indel events per aligned base versus rate of mismatch per aligned base (see 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.

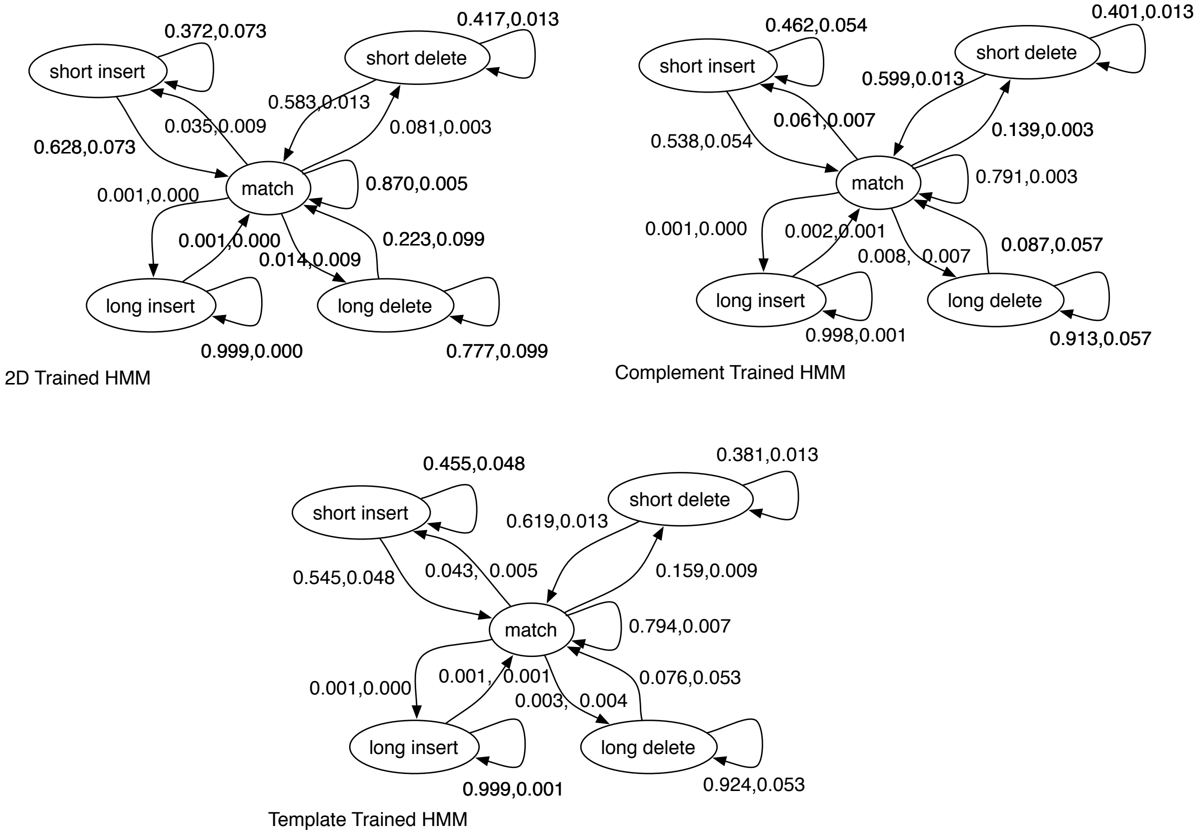


Figure 1.7: Structure for the hidden Markov model (HMM) used for expectation maximization (EM).
Structure of HMM used for EM, along with the estimated parameters for transition probabilities for template, complement and 2D reads. For each transition in order, the mean estimates and standard error across all experiments for that read type are shown.

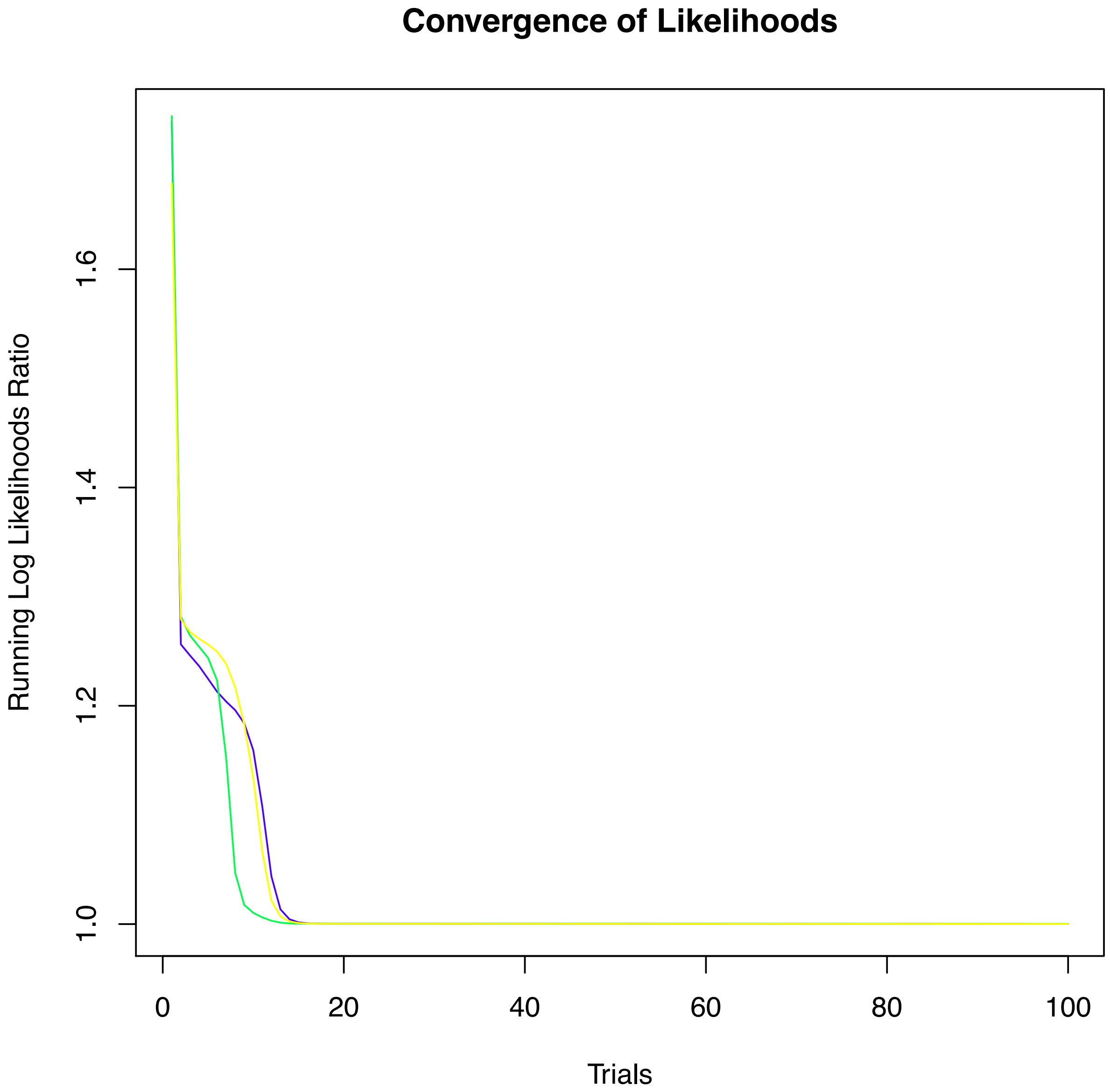


Figure 1.8: Convergence of log-likelihood ratios achieved using expectation maximization.

Convergences of log-likelihood for three independent runs of expectation maximization, each from a randomly parameterized 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 LAST.

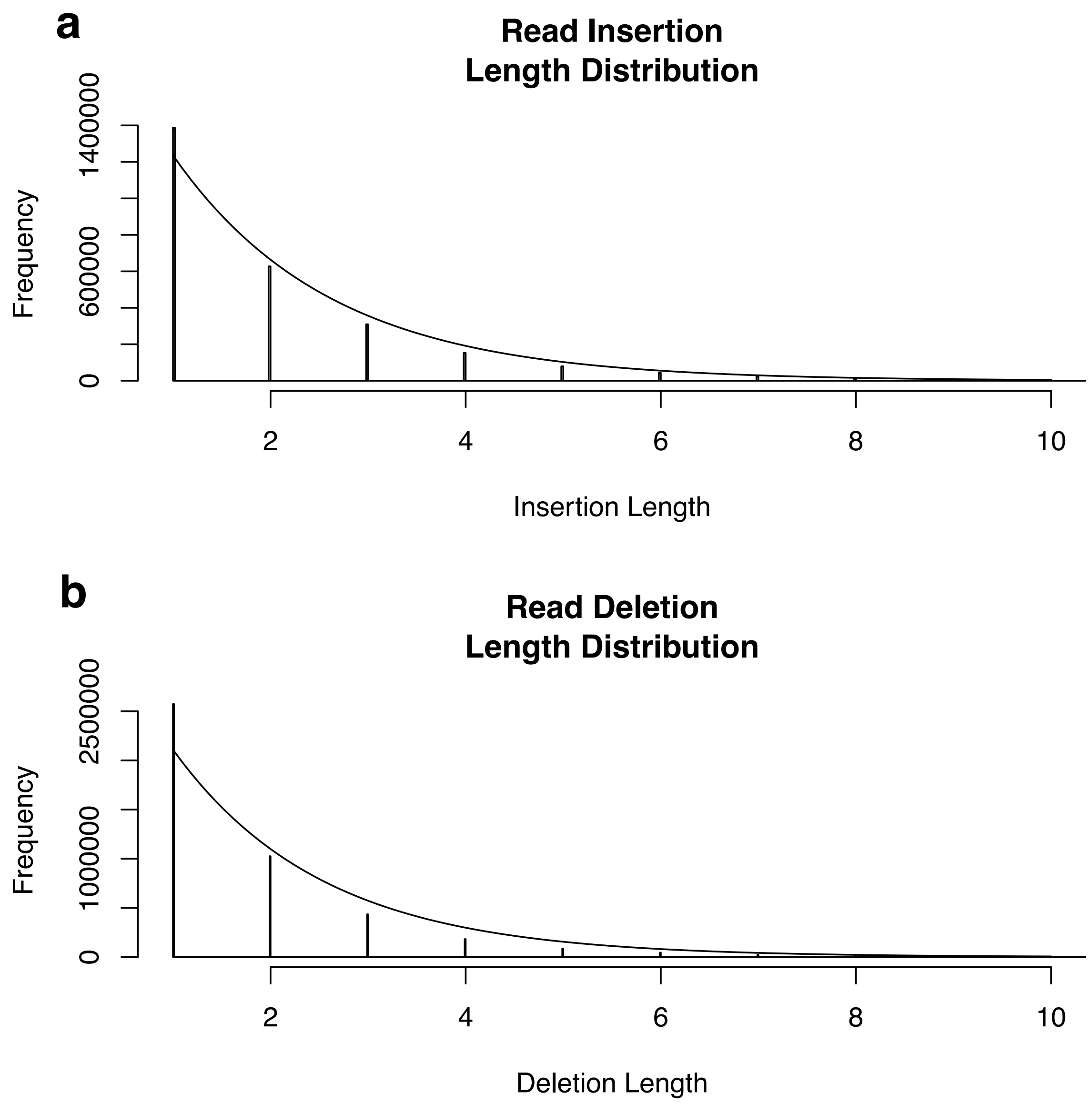


Figure 1.9: Frequency plots for insertions and deletions in MinION read alignments. Representative insertion and deletion plot for reads (fitted with an exponential distribution) from one M13 experiment using R7.3 chemistry, aligned using expectation maximization-trained LAST.

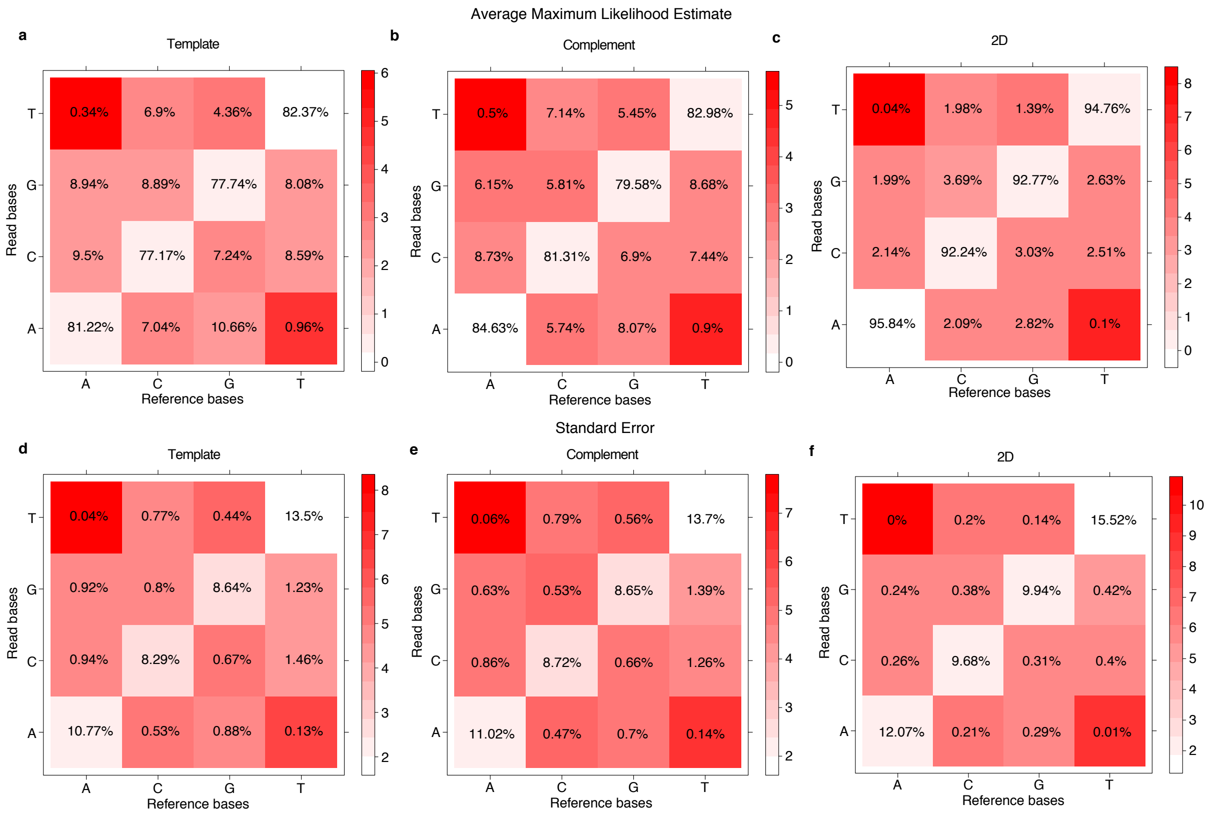


Figure 1.10: Substitution matrices from alignments using expectation maximization-trained model.

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 row illustrates the average maximum-likelihood estimates for these substitutions, with the standard error represented in the lower row. For all aligners, thymine-to-adenosine and adenosine-to-thymine substitution rates were low, indicating that the device rarely miscalled one as the other. The color scheme is fitted on a log scale, and the substitution values are on an absolute scale.

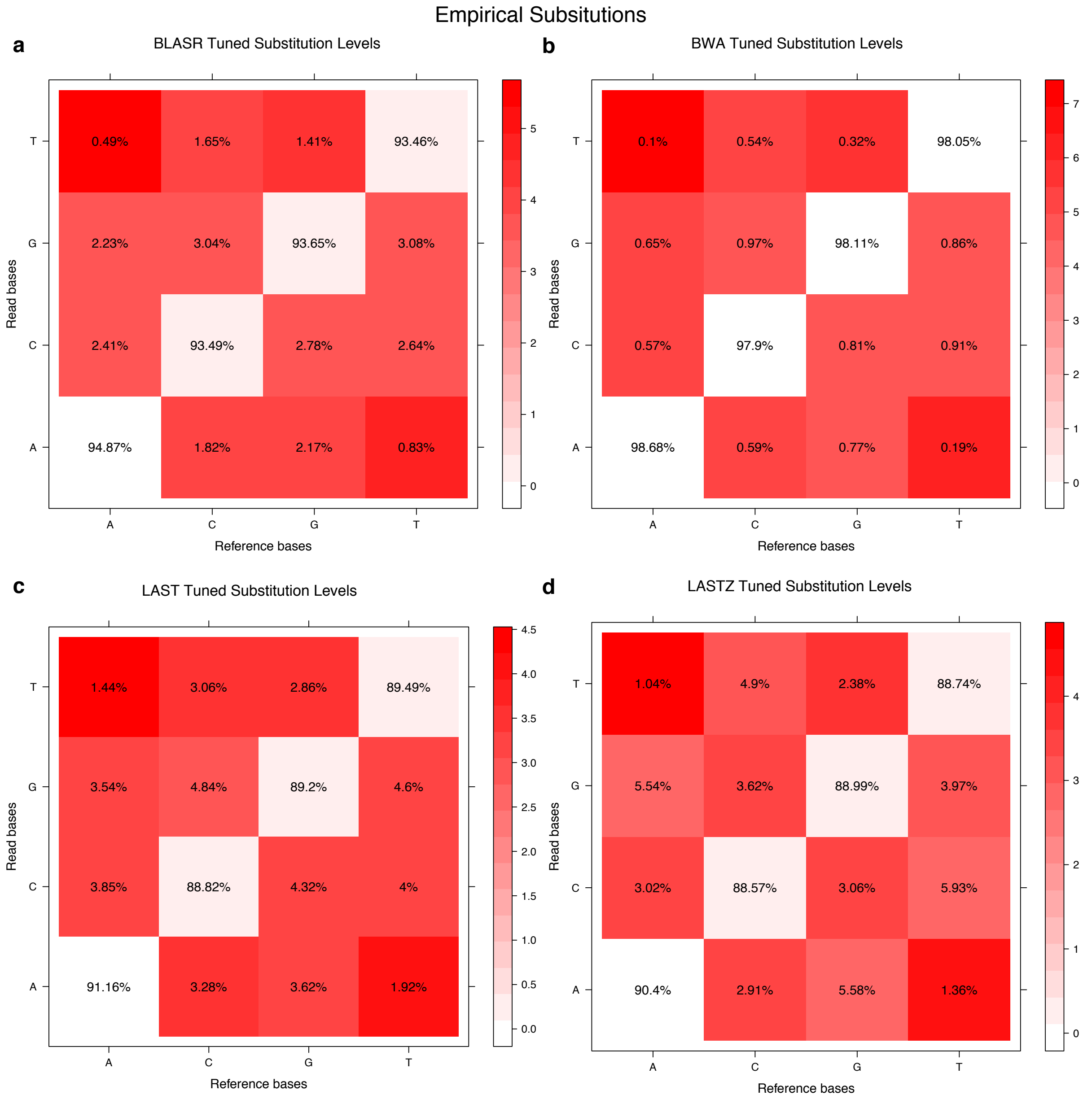


Figure 1.11: Substitution matrices from alignments using tuned parameters. Substitution matrices for each of the four tuned aligners across three M13 experiments using R7.3 chemistry. For all aligners, thymine-to-adenosine and adenosine-to-thymine substitution rates were low, indicating that the device rarely miscalled one as the other. The color scheme is fitted on a log scale, and the substitution values are on an absolute scale.

Realigning reads using the MLE parameters and the AMAP objective function [82] yielded substantial improvements over the initial alignments for every tuned program (see Methods, Figs. 1.5d-f and 1.12). For high-confidence alignments, there were no clear correlations between read length and errors (Fig. 1.13). However, there were positive correlations among the rates of insertions, deletions and substitutions in 2D reads (1.14; Supplementary Note 2).

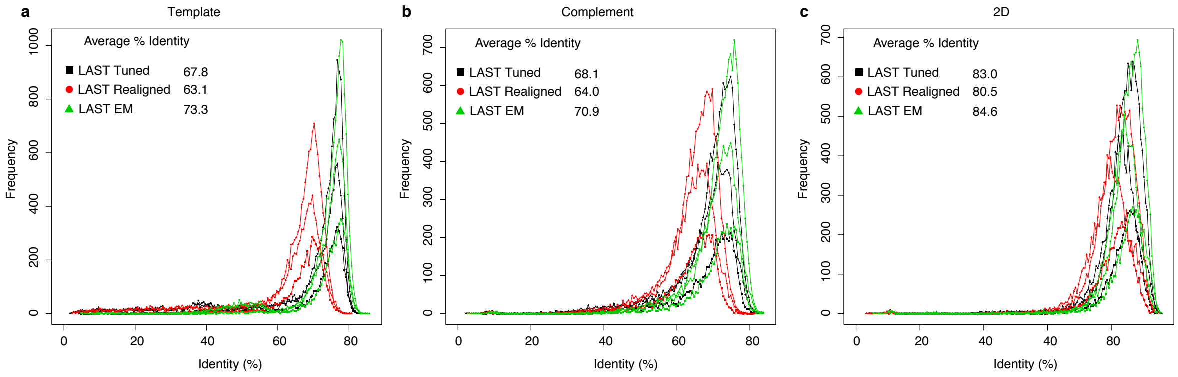


Figure 1.12: Realignment improves read identity.
Read identity for template, complement and 2D reads for three M13 replicate experiments using R7.3 chemistry, aligned using LAST. Three versions of the LAST alignment are shown: tuned LAST, trained LAST realignments and naive LAST realignments.

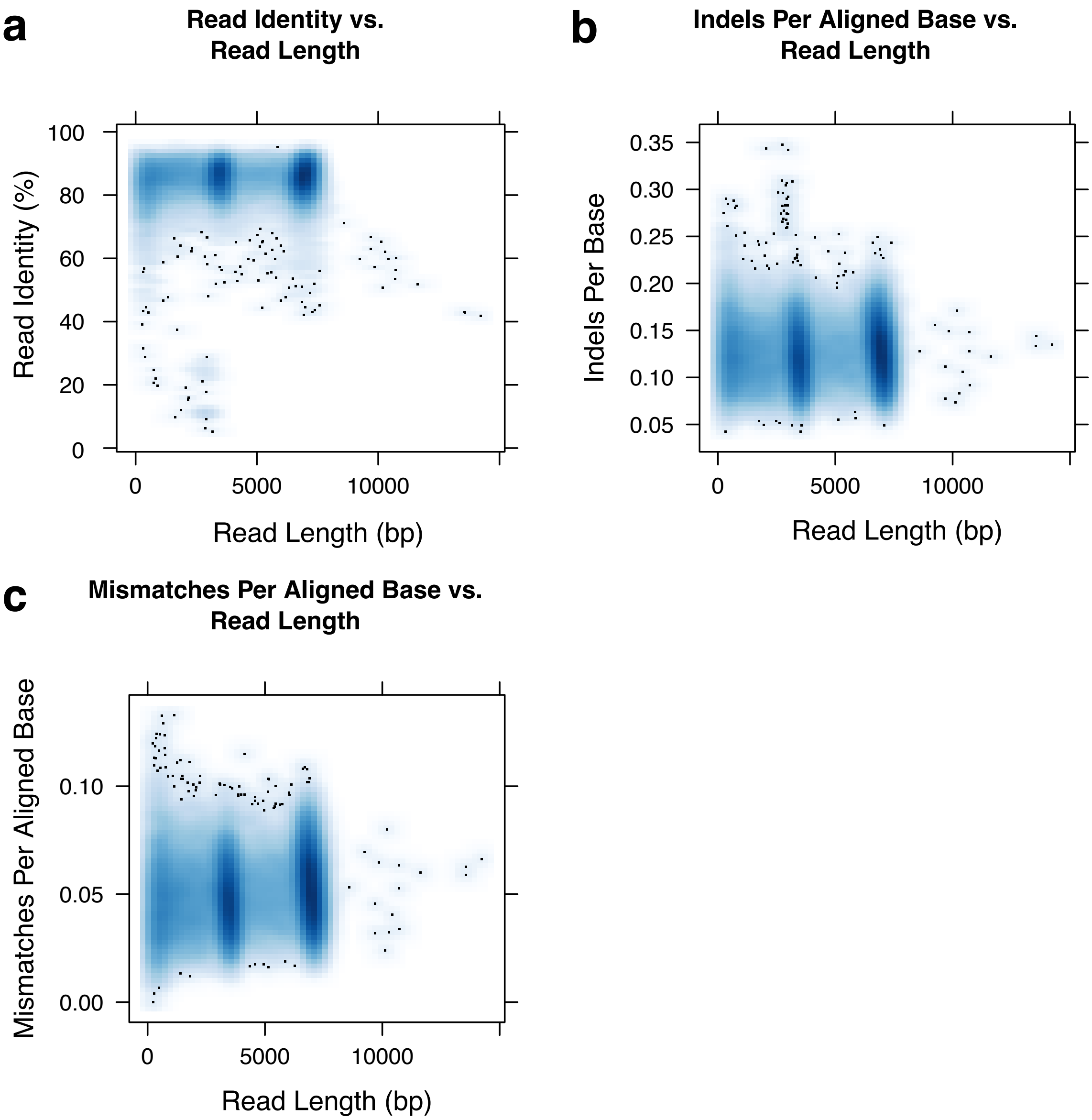


Figure 1.13: An alignment quality measurement for 2D reads across three M13 replicate experiments.

Alignments were obtained using expectation maximization-trained LAST realignments. The two density clusters correspond to M13 and phage λ DNA control.

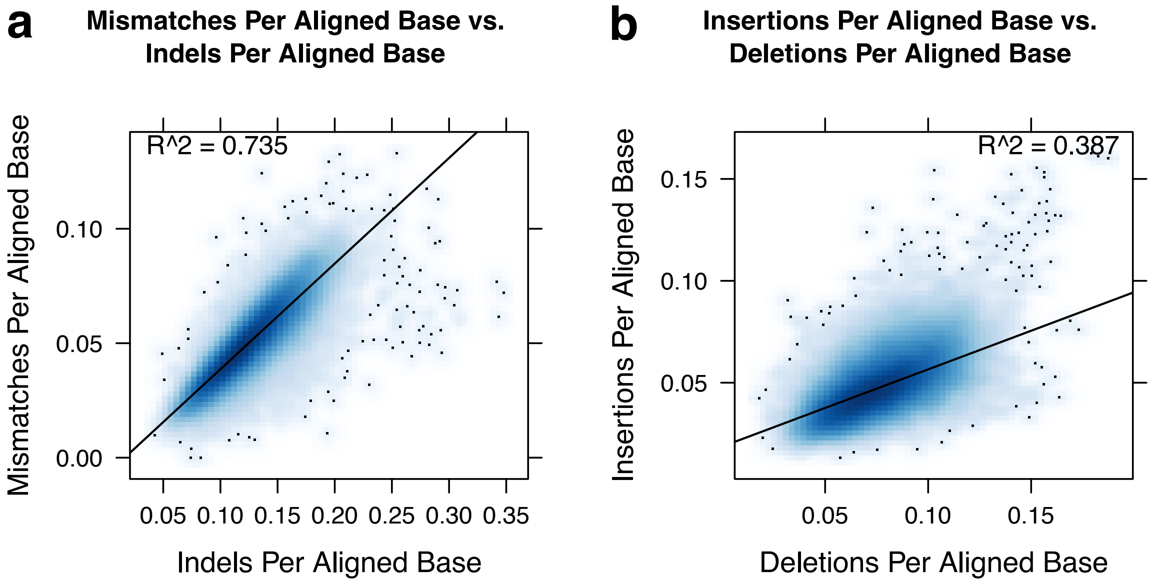


Figure 1.14: Error profiles for 2D reads after realigning using expectation maximization-trained model.
Error profile analysis of 2D reads aligned using expectation maximization-trained LAST realignments indicates a moderate correlation between mismatches and indels per aligned base, and a weak correlation between insertions per aligned base and deletions per aligned base.

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 (Table 1.4); however, error rates were substantially closer to the MLE parameters estimated by expectation maximization. This suggests that ont2d is an improvement over the pacbio mode (for Pacific Biosciences) that we used originally.

Table 1.4: 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

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. [77], 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. [77] released their data set. We observed an improvement in average identity from 80.1% with tuned LAST to 81.8% after re-alignment 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 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 (Figs. 1.15 and 1.16). However, 192 positions (2.6%) were underrepresented (see 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 with 5-mers rich in polymeric nucleotide runs (Table 1.5). 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 (see Supplementary Note 3; Table 1.6). These findings are consistent with observations from Ashton *et al.* [83].

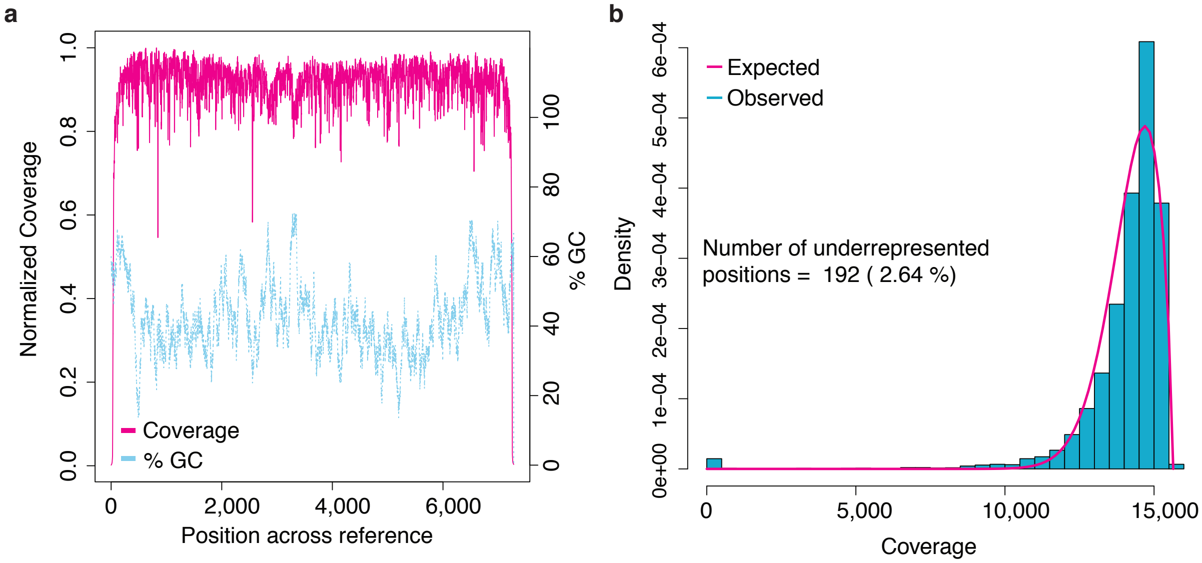


Figure 1.15: M13 sequencing depth. (a) The magenta line denotes coverage by position in the genome (binned over a sliding 5-bp window), and the blue line depicts the local percentage of GC for that position (binned over a 50-bp sliding window). (b) Coverage-depth distribution fitted with a generalized extreme-value distribution.

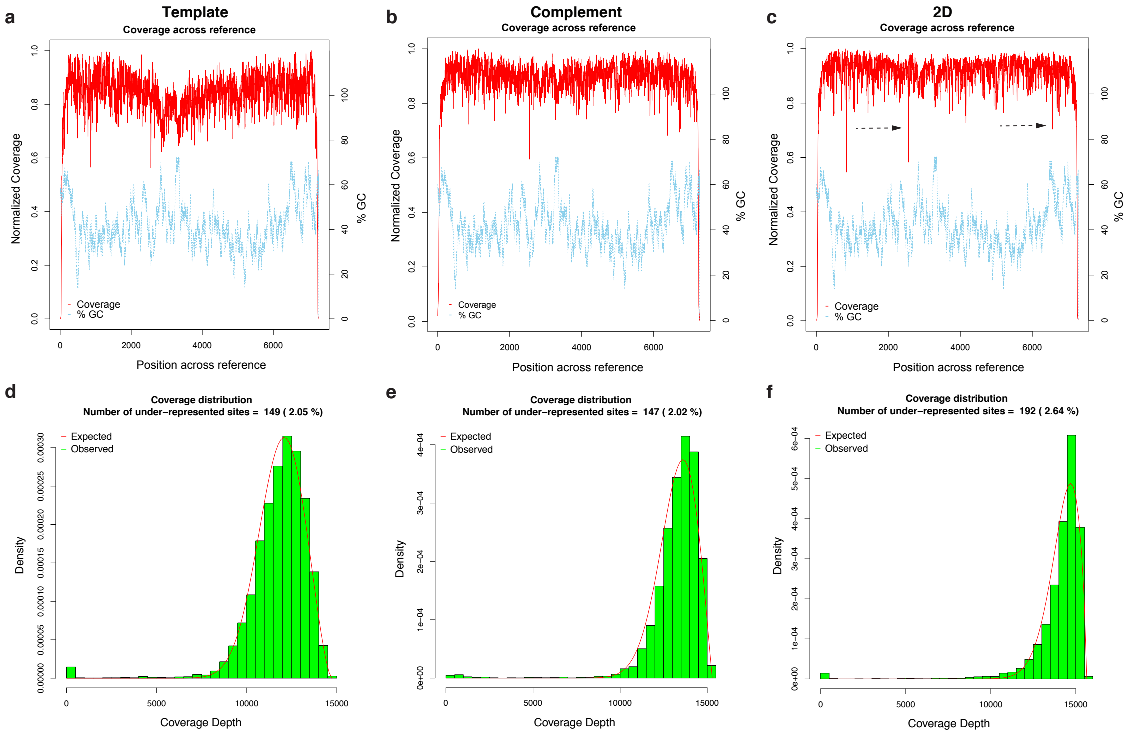


Figure 1.16: The coverage and percentage of GC across the M13 genome. (a-c) Coverage, smoothed by binning over a sliding 5-bp window, matching the k-mer length used in base calling. The GC content was calculated by binning over a 50-bp sliding window. Halving and doubling this window size did not drastically alter the result. (d-f) Coverage histograms across three M13 replicate experiments using R7.3 chemistry and aligned using expectation maximization-trained LAST realignments. About 2.1%, 2.0% and 2.6% of the M13 genome was underrepresented in template, complement and 2D reads, respectively.

Table 1.5: 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 by Metrichor.

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		

Table 1.6: Over and under represented 5-mers between reads and M13 reference. Lambda 5-mers were not counted in this comparison. Both strands are compared and represented in this table. Below, over and under represented 5mers that span indels in aligned reads across all three read types. *(T - template; C - complement)*

Top Kmers In Reads vs. M13 Reference

Ref	logFC	2D	logFC	Ref	logFC	C	logFC	Ref	logFC	T	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

Top Enriched Kmers Spanning Aligned Indels

Ref	logFC	2D	logFC	Ref	logFC	C	logFC	Ref	logFC	T	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

MinION reads can call SNVs with high recall and precision

SNV detection is important for metagenomics and microbial strain detection [84, 85, 86]. To determine if MinION reads could be used for SNV discovery in monoploid genomes, we computationally introduced random substitutions into the M13 reference sequence at 1-to-20% frequency. Using this altered sequence as an alignment reference we attempted to recover these substitutions using a Bayesian transducer framework [87] (see Methods; Supplementary Note 4) and assessed performance in terms of precision, recall, and F-score. These experiments also addressed the accuracy of alignment and error models while avoiding reference-allele bias. Reference-allele bias can skew simple metrics like alignment identity.

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. 1.17a). When we reduced the sequencing depth down to a more reasonable 60x 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. 1.17b)

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. 1.17c). 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. 1.17a-b).

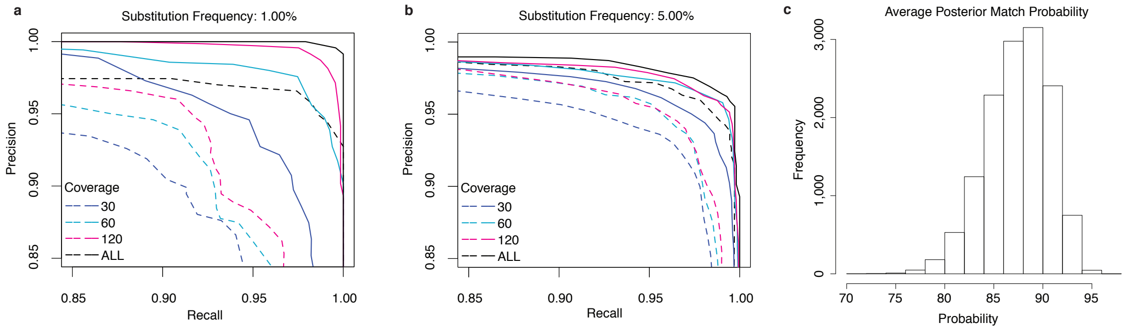


Figure 1.17: Exploring SNV calling with MinION reads.
(a,b) Variant calling with substitution frequencies of (a) 1% and (b) 5%. Dashed lines in both a and b represent results from variant calling using a transducer model conditioned on a fixed, tuned LAST alignment. Different sampled read coverages are 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.

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. 1.5), we found substantially 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 λ DNA mapped back to the reference with 87% identity (Fig. 1.18). 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.

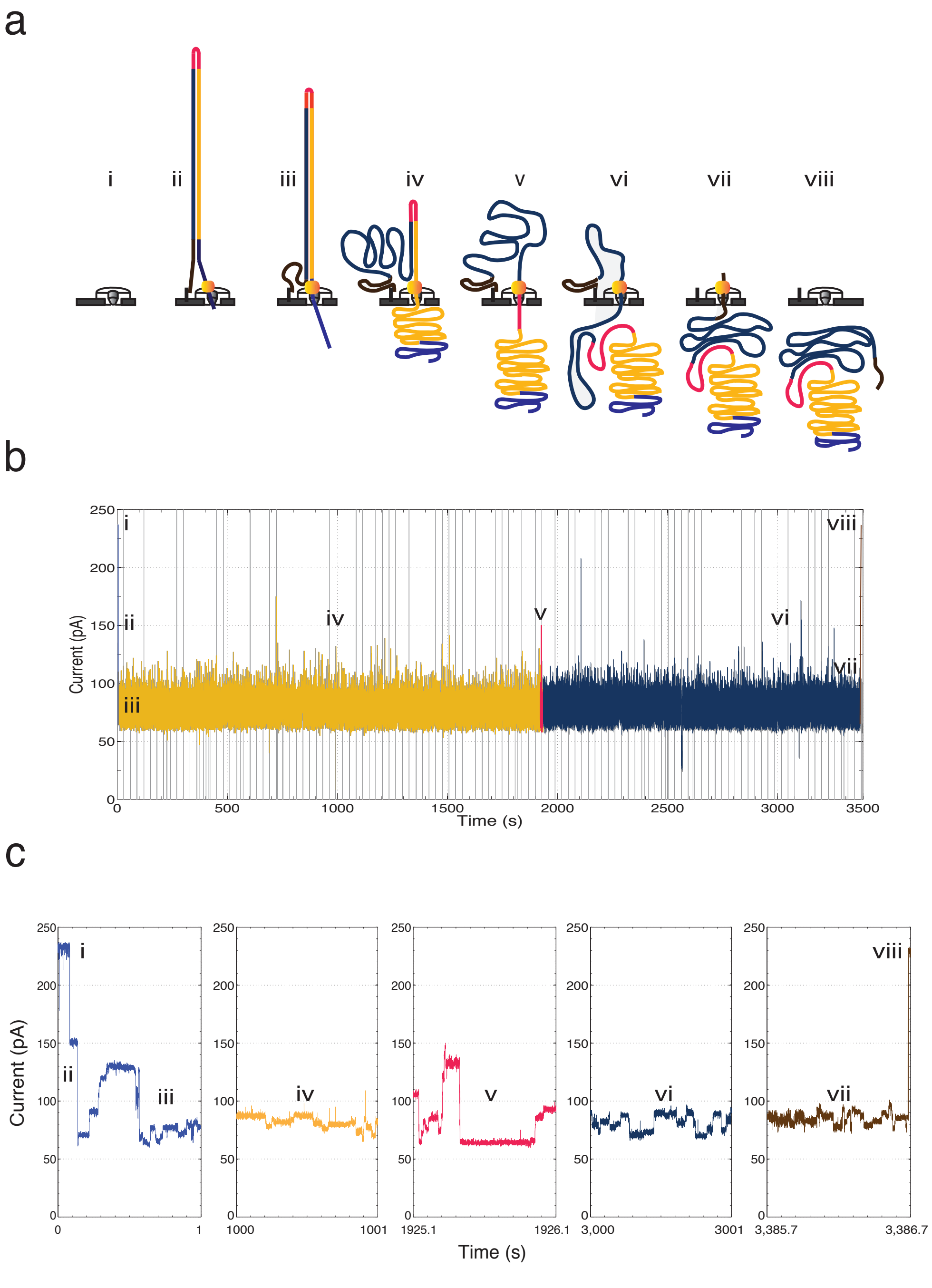

a
i
ii
iii
iv
v
vi
vii
viii
b
250
200
150
100
50
0
Current (pA)
0
500
1000
1500
2000
2500
3000
3500
Time (s)
i
ii
iii
iv
v
vi
vii
viii
c
Current (pA)
i
ii
iii
iv
v
vi
vii
viii
0
1
1000
1001
1925.1
1926.1
3,000
3001
3,385.7
3,386.7
Time (s)

Figure 1.18: MinION data for full-length (48-kb) λ phage dsDNA. Data for a 2D read of a full-length λ phage dsDNA from the MinION. (a) Molecular events for translocation of a single 48-kb λ dsDNA molecule through the MinION 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 the DNA construct through the nanopore (approximately 2,789 s). Regions of the ionic current trace corresponding to steps i-viii are labeled. (c) Expanded 1-s 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. The FASTA sequence is available at `http://figshare.com/articles/UCSC_Full_Length_Lambda_2D_Read/1209636`.

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. 1.19a). 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 [88]. 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 [89]. Furthermore, copy-number expansion and contraction involving genes contribute to variability in gene expression, epigenetic regulation and association with human disease [90, 91].

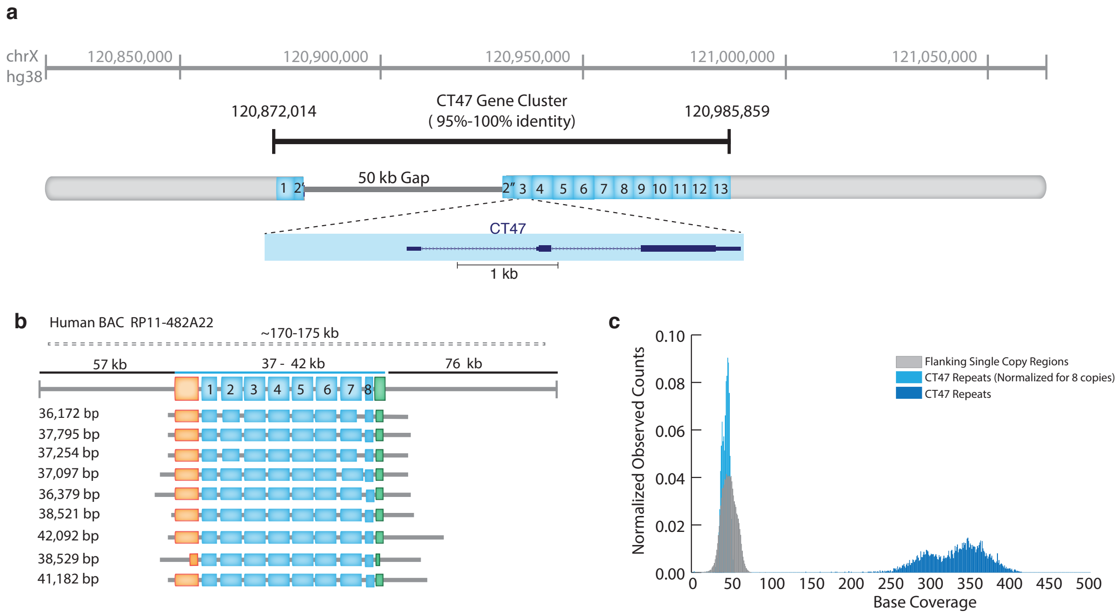


Figure 1.19: Resolution of CT47 repeat copy-number estimate on human chromosome Xq24. (a) BAC end sequence alignments (RP11-482A22: AQ630638 and AZ517599) span a 247-kb region, including 13 annotated CT47 genes16 (each within a 4.8-kb tandem repeat), and a 50-kb scaffold gap in the GRCh38/hg38 reference 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).

We used the MinION to acquire 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 (see Methods, Fig. 1.19b, and Table 1.7). 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 (Fig. 1.20). As an additional test, we obtained 40x-60x sequence coverage of the unresolved Xq24 segment using shorter (∼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. 1.19c)

Table 1.7: 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
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

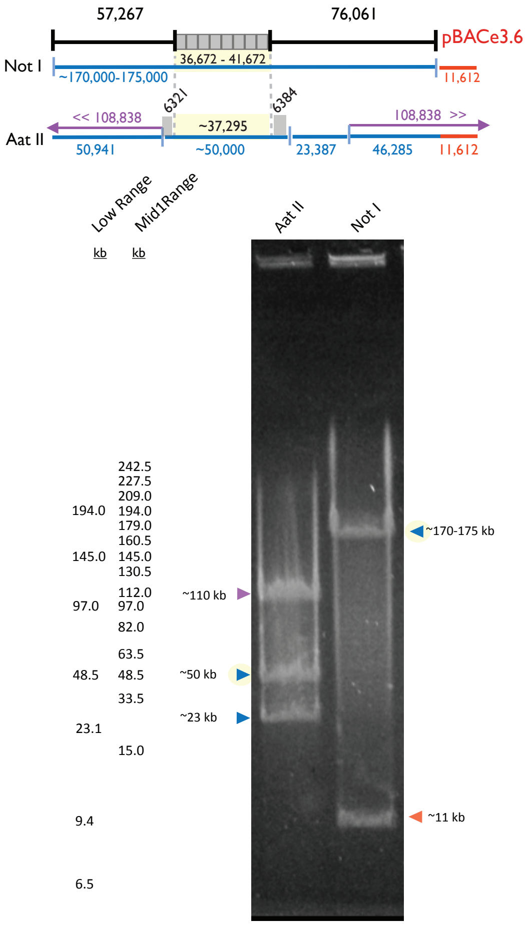


Figure 1.20: Pulse-field gel electrophoresis of RP11-482A22 BAC DNA to determine insert length. The span of BAC end sequences relative to GRCh38 reference assembly provides estimates of 57 kb to the right of the repeats and 76 kb to the left of the repeats (depicted in black). To determine the length of the repeats, we performed NotI and AatII digests 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 (blue) and an 11.6-kb cloning vector band (red). After subtraction of the known flanking region sizes, this estimate provides a repeat region in the range of 36.7-41.7 kb, or 7.5 to 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 (50 kb), the 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 subtraction of 12 kb of known flanking sequence, marked with gray shading). Regions providing evidence for repeat copy number are highlighted in yellow.

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 (~14.4 kb total) were routinely achieved. Approximately 99% of 2D reads mapped to a reference (M13 or phage λ 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 [83, 92] 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 [93, 94, 95], abasic residues [96, 97], DNA adducts [98], 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

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 (10 mM Tris, 1 mM EDTA, pH 8) buffer-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 TAAGGTAATTCACAATGATTAAAGTTG, CTGTGGAATGCTACAGGC, CACCTTTAATGAATAATTTCCGTC, CATGCTCGTAAATTAGGATGG, GTTTTACGTGCTAATAATTTTGATATG, CAAGGCCGATAGTTTGAGT, CACTGGCCGTCGTTTTA, GAGGCTTTATTGCTTAATTTTGC, AGGTCTTTACCCTGACTATTATAG, AGGCTTTGAGGACTAAAGAC, AATGGATCTTCATTAAAGCCAG, CAGCCTTTACAGAGAGAATAAC, TCCGGCTTAGGTTGGG, GTGAGGCGGTCAGTATTAAC, GAGATAGGGTTGAGTGTTGT 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 λ 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 [73] (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 ([74, 99]) (H. Li's popular adaptation of the BWA mapper altered for handling long reads), LAST Version 490 ([75, 100]) (a fast, sensitive, adaptable and popular pairwise-alignment tool) and LASTZ Release 1.02.00 ([76]) (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 λ phage DNA supplied by ONT 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 (Fig. 1.4).

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 [78, 79]. The majority of unmapped 2D reads had BLAST hits (Fig. 1.5 and Table 1.3), 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[101] that has two sets of insertion-deletion states (Fig. 1.7), 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.[101]. To train the model, we used a hybrid form of the Baum-Welch algorithm (a type of expectation maximization). For each read, this hybrid algorithm works within an alignment band around a fixed guide alignment[81]. The band is constructed as described by Paten et al.[81] using code adapted from the Cactus alignment program[102]. The guide alignment comes from a mapping program. 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. Figure 1.8 shows the results of one training experiment, in which there is convergence of log-likelihood for all three runs to essentially the same value. Figure 1.8 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. This indicated that our parameter estimates were robust.

Figure 1.6a,b shows, as a cross-check, the calculation of insertion, deletion, and substitution rates for 2D reads from realignments computed (see Realignment with a trained model) from each guide alignment using the alignment and the trained model. In each case, despite the fact that the starting guide alignments had 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. This indicated that tuned-BLASR was the 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 Table 1.2), 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[82]. The AMAP objective function 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[102], 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 (Fig. 1.12).

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 of 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. 1.17). 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[87], and the models were combined to create an overall HMM, using the evolutionary HMM formalism[87]. 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. 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[103]: 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. Pecan software[81] 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[104] 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. 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)[104]. 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 available (nanopore pipeline at `https://github.com/mitenjain/nanopore`; and marginAlign pipeline at `https://github.com/benedictpaten/marginAlign`.

Accession codes

Primary accessions

European Nucleotide Archive

PRJEB8230 - `http://www.ebi.ac.uk/ena/data/view/PRJEB8230`

ERP009289 - `http://www.ebi.ac.uk/ena/data/view/ERP009289`

Acknowledgements

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U54HG007990. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Contributions

MA conceived experiments and directed research. BP conceived and directed bioinformatics analysis. BP, MJ, IF, and KHM were responsible for bioinformatics analysis and software development. MJ and HEO were responsible for completion of sequencing experiments and data processing. MJ and HEO were responsible for preparing DNA sequencing standards. HEO was responsible for Sanger sequencing of M13 dsDNA. BP and IF were responsible for k-mer and BLAST analysis. BP and MJ were responsible for SNV analysis. BP developed and implemented EM and realignment strategies. KHM conceived and directed BAC experiments and data analysis. All authors contributed to manuscript writing, editing, and completion.

Supplementary Note 1

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

Fig. 1.6c and Supplementary Fig. 1.10 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 sparse, and similarly, but less pronounced, 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 Fig. 1.11). In each case, we saw the same trend. To ascertain if the low substitution error rates were influencing the transition parameters during training (e.g. certain substitutions being traded for higher rates of insertions/deletions, Supplementary Fig. 1.9), 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).

Supplementary Note 2

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 Fig. 1.14). 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 hypothesis that explained 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. For example, 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. This was without the benefit of 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 [77].

Supplementary Note 3

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 Fig. 1.16a-c respectively. For all three read types coverage was largely consistent across the genome, apart from at the very ends of the genome, and did not appear to fluctuate substantially based upon GC content. However, 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 [105] (Supplementary Fig. 1.16d-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 two 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 Table 1.5).

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 five 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 five bases with a slide of one 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. 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. The most enriched 5-mers are G/C rich and did not contain homopolymer repeats (Supplementary Table 1.6).

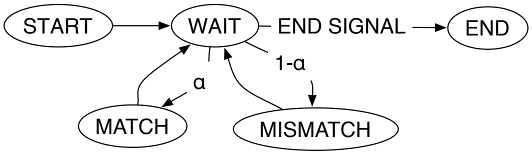
We also compared 5-mers spanning indels in alignments. For this experiment, indels were defined as any 5-mer which has an alignment gap of any size in the four internal positions. We found similar trends in these 5-mers as in the overall counts, with poly-dA and poly-dT 5-mers being under-represented 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 Table 1.6).

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.

Supplementary Note 4

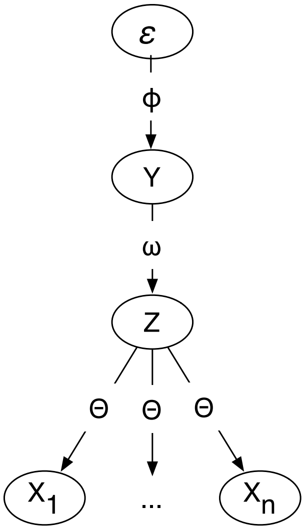
Approach to SNV detection

The relatively high error rates of MinION reads make single nucleotide variant (SNV) discovery potentially challenging (Supplementary Fig. S1.23). Here we describe a method for variant calling that can tolerate this level of error. Let a DNA 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 DNA sequences, Y the given mutated reference DNA sequence, Z the true M13mp18 reference DNA 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 a 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 [87]. 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:



Supplementary Fig. S 1.21: Substitution model.

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):



Supplementary Fig. S 1.22: Model.

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

$$f(X, Y) = \arg\max_{Z'} = P(Z'|Y, \omega)P(Y|\phi)\prod_{i} P(X^i|Z', \theta), \tag{1.1}$$

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 [106]. However, exact dynamic programming algorithms that are exponential in the cardinality of X exist, and a number of principled heuristics have been proposed [107].

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

$$f'(X, Y, \sim) = \arg\max_{Z'} P(Z'|Y, \sim, \omega)P(Y, \phi) \prod_i P(X^i|Z', \sim, \theta), \tag{1.2}$$

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:

$$\arg\max_{Z'_i} P(Z'_i|Y_i, \omega)P(Y_i|\psi) \prod_{X^j_k \sim Y_i} P(X^j_k|Z'_i, \theta), \tag{1.3}$$

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 (1.3) with

$$\arg\max_{Z'_i} P(Z'_i|Y_i,\omega)P(Y_i|\psi)\prod_j \sum_k P(X^j_k|Z'_i,\theta)P(X^j_k \sim Y_i|\phi\circ\omega\circ\theta), \qquad (1.4)$$

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 1.3 marginalising over the probability of all pairwise alignments between each read and the mutated reference sequence.

Instead of calculating 1.4, 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.

In practice, 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. S1.24 and Supplementary Table S1.8 show the results. Note the numbers in the table (and subsequent tables) are the average 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 average over all replicates as a function of the posterior base calling probability threshold. To demonstrate the methods and parameters we chose were reasonable, we compared to a number of parameter and algorithm variations.

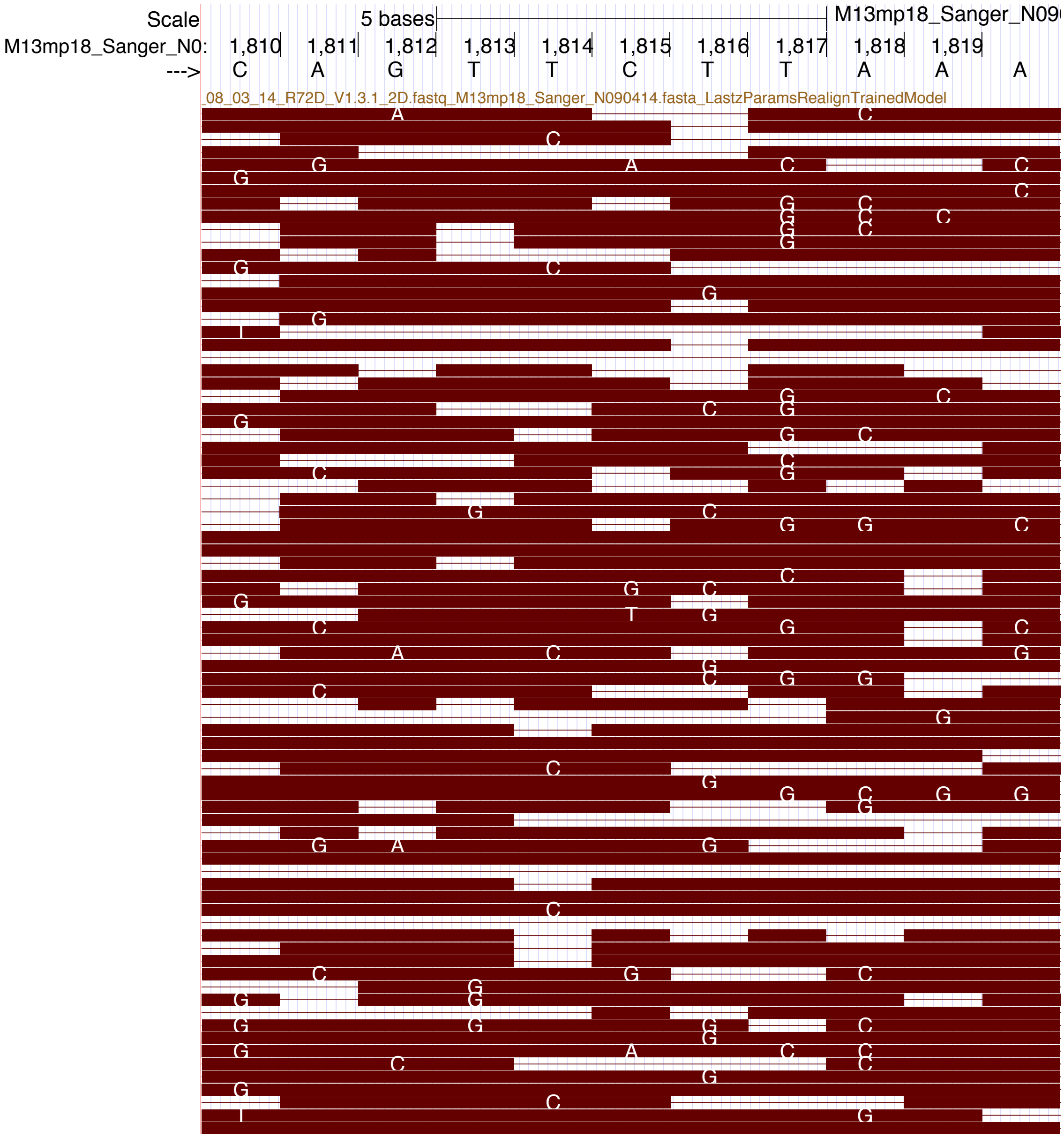
In calculating the posterior match probabilities by setting $\alpha = 0.6$ (a mismatch rate of 40%), we see a decrease in F-score for a 1% mutation frequency (average across all coverages), but a gain for 5% and greater mutation frequencies (Supplementary Fig. S1.25 and Supplementary Table S1.9). This 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 average precision and recall of 98% for a 5% mutation frequency.

For $\alpha = 1.0$ (equivalent to not modeling mismatches), we see significantly lower performance (Supplementary Fig. S1.26 and Supplementary Table S1.10). 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, allow 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).

Next, we calculate the posterior base calling probabilities using a replacement θ instead of θ from our EM trained model. The replacement θ is from a model that has equal probabilities for all substitutions. This strategy is equivalent to picking the base with the highest posterior match probability expectation. In so doing, we see a slight decrease in performance (Supplementary Fig S1.27 and Supplementary Table S1.11). This suggests that 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. S1.28 and Supplementary Table S1.12). This is unsurprising given that the modal posterior match probability is less than 90% (Fig. 1.17(C)).

As might be expected, switching to using template or complement reads instead of 2D reads causes substantially poorer performance (Supplementary Fig. S1.29-1.30 and Supplementary Tables S1.13-1.14). However, this may be somewhat due to using an alignment model trained for 2D reads.

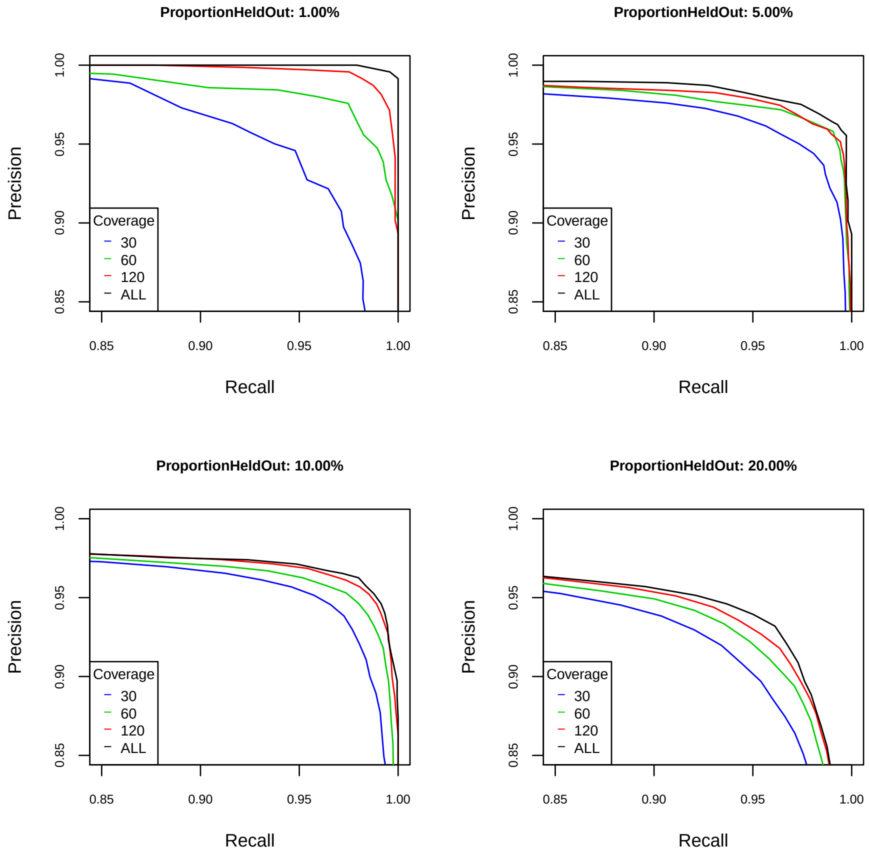


Supplementary Fig. S 1.23: Visualization of an alignment of 2D reads with M13 using trained LAST realignments on the UCSC Genome Browser. The high indel and mismatch rate are clearly evident.

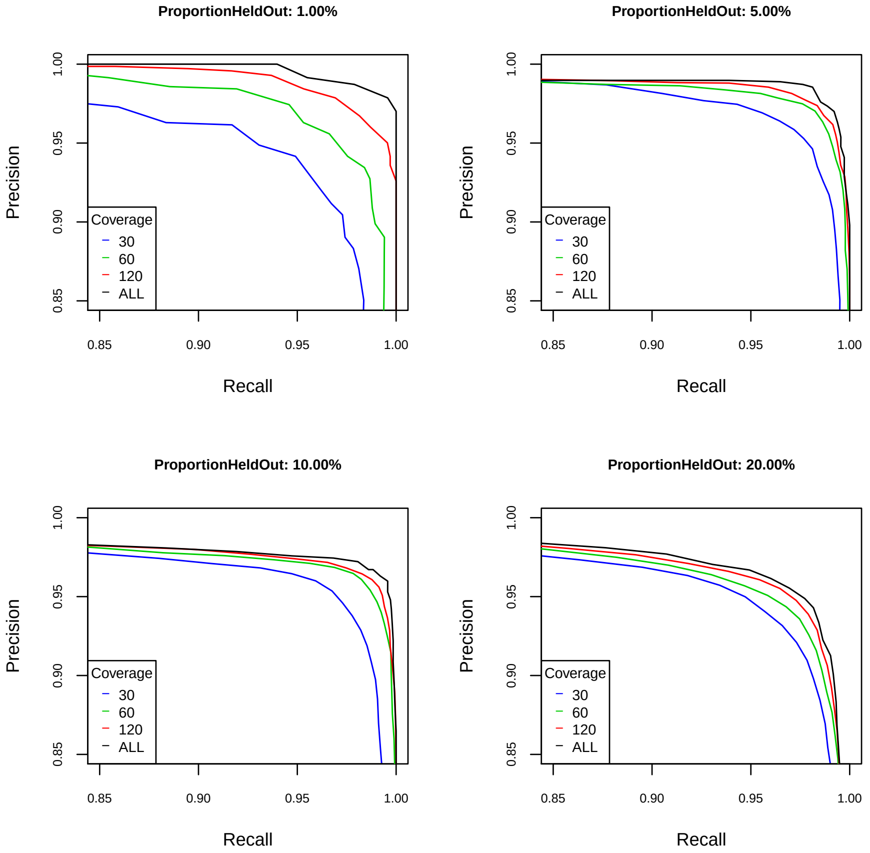
SNV detection using 2D reads

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

Supplementary Table S 1.8: Variant calling on M13 using 2D reads starting with the tuned LAST (run uses the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. We performed variant calling by using posterior match probabilities to integrate over all possible read alignments to the mutated reference. We used 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 were calculated using the EM trained HMM model. Each posterior match probability accounts for substitution differences between the mutated reference and the true reference and assumes 20% divergence. The variant calling results shown are 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 that get called as variants after reads have been aligned. 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. At each coverage value, there are three different samplings of the reads.



Supplementary Fig. S 1.24: Precision/recall curves show variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling used 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0-a 1' flags) mapping algorithm. Variant calling was calculated using posterior match probabilities to integrate over every possible read alignment to the mutated reference. We used the initial guide alignment to band the calculations. Variant calling also used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities were calculated using the EM trained HMM model. Each assumed 20% divergence to account for substitution differences between the mutated reference and the true reference. The variant calling results shown are for a posterior base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of mutated sites called as variants in the aligned reads. Coverage is the total length of reads sampled divided by the length of the reference. ALL symbolizes all the reads for a given experiment. Results shown are across three replicate experiments. There are three different samplings of the reads at each coverage value.

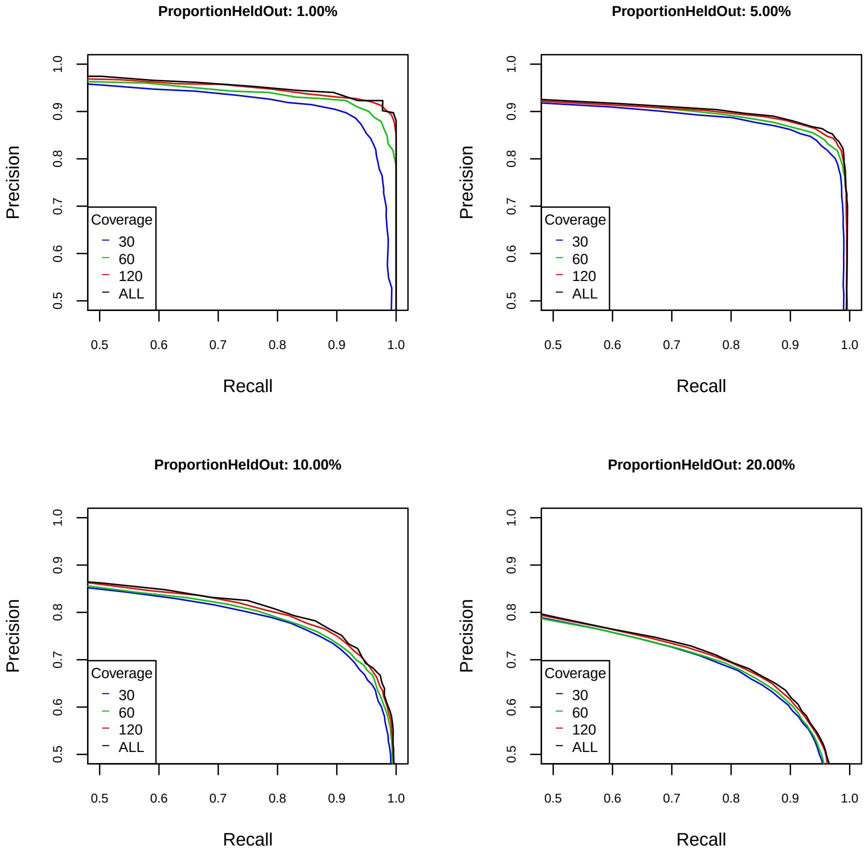


Supplementary Fig. S 1.25: Precision/recall curves show variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling used 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed using posterior match probabilities to integrate over every possible read alignment to the mutated reference. The initial guide alignment was used to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities were calculated using the EM trained HMM model. We assumed 40% divergence to account for substitution differences between the mutated reference and the true reference. The variant calling results shown are 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 that get called as variants when reads are aligned. 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. There are three different samplings of the reads at each coverage value.

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

Supplementary Table S 1.9: Variant calling on M13 using 2D reads starting with the tuned LAST (run uses the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. We performed variant calling by using posterior match probabilities to integrate over all possible read alignments to the mutated reference. We used 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 were calculated using the EM trained HMM model. We assume 40% divergence to account for substitution differences between the mutated reference and the true reference. The variant calling results shown are 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 that get called as variants after reads have been aligned. 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. At each coverage value, there are three different samplings of the reads.



Supplementary Fig. S 1.26: Precision/recall curves show variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling used 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling used posterior match probabilities to integrate over every possible read alignment to the mutated reference. The initial guide alignment was used to band the calculations. Variant calling used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities were calculated using the EM trained HMM model. Here, we do not account for substitution differences between the mutated reference and the true reference. We use the posterior base calling probability threshold that gives the optimal F-score for variant calling. Mutation frequency is the approximate proportion of mutated sites in the reference that get called as variants in aligned reads. 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. There are three different samplings of the reads at each coverage value.

SNV detection using 2D reads

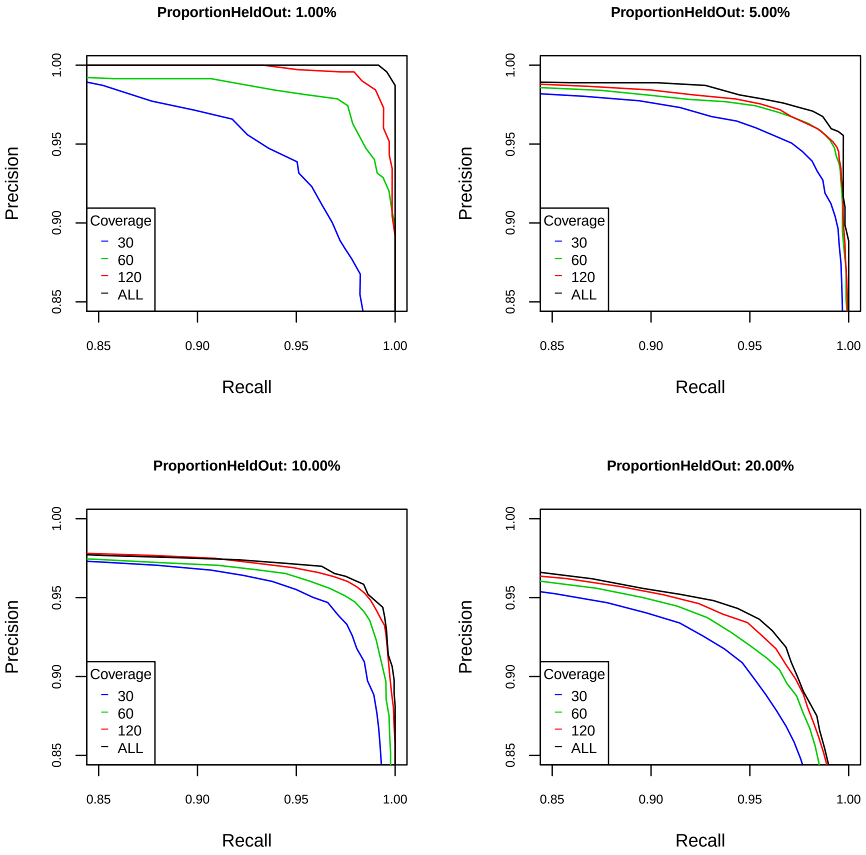
Metric	Mut. Freq.	Coverage			
		30	60	120	ALL
Recall	1	89.60	89.74	91.60	91.88
	5	83.86	84.92	84.86	85.52
	10	74.58	74.87	75.60	77.35
	20	67.12	67.11	67.05	68.10
Precision	1	92.98	96.80	97.90	98.64
	5	94.88	95.21	97.08	96.90
	10	88.10	88.53	89.62	88.24
	20	82.25	82.88	84.35	83.02
F-score	1	91.23	93.12	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

Supplementary Table S 1.10: Variant calling on M13 using 2D reads starting with the tuned LAST (run uses the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. We performed variant calling by using posterior match probabilities to integrate over all possible read alignments to the mutated reference. We used the initial guide alignment to band the calculations. Variant calling also used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities were calculated using the EM trained HMM model. Therein, we disregard substitution differences between the mutated reference and the true reference. We use the posterior base calling probability threshold that gives the optimal F-score for variant calling. Mutation frequency is the approximate proportion of sites mutated in the reference that get called as variants in aligned reads. 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. There are three different samplings of the reads at each coverage value.

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

Supplementary Table S 1.11: Variant calling on M13 uses 2D reads starting with the tuned LAST (run uses the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. We performed variant calling by using posterior match probabilities to integrate over all possible read alignments to the mutated reference. We used the initial guide alignment to band the calculations. This variant calling strategy corresponds to choosing the maximum-frequency/expectation of a non-reference base. Posterior match probabilities were calculated using the EM trained HMM model. Therein, we assume 20% divergence to account for substitutions between the mutated reference and the true reference. We use the posterior base calling probability threshold that gives the optimal F-score for variant calling. Mutation frequency is the approximate proportion of sites mutated in the reference that get called as variants in aligned reads. 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. There are three different samplings of the reads at each coverage value.

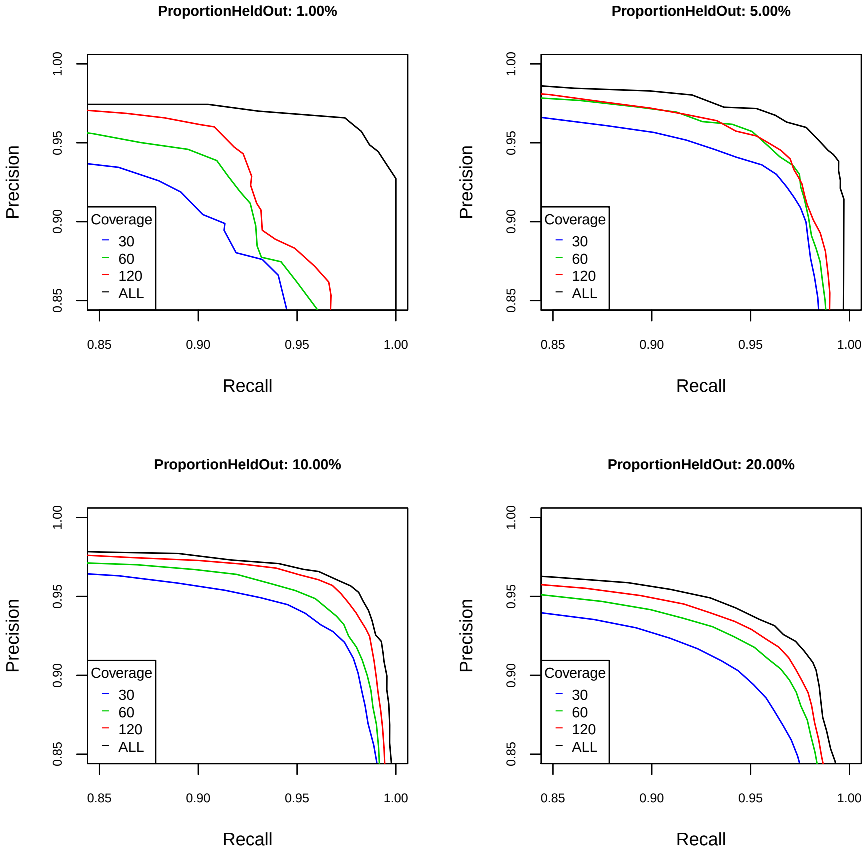


Supplementary Fig. S 1.27: Precision/recall curves show variant calling performance for four different mutation frequencies: 1, 5, 10, and 20 percent. Variant calling used 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. For variant calling, we used posterior match probabilities to integrate over every possible read alignment to the mutated reference. The initial guide alignment was used to band the calculations. This variant strategy corresponds to choosing the maximum-frequency/expectation of a non-reference base. Posterior match probabilities were calculated using the EM trained HMM model. Therein, we assume 20% divergence to account for substitution differences between the mutated reference and the true reference. For variant calling, we used the base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of mutated sites in the reference that get called as variants in aligned reads. 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. There are three different samplings of the reads at each coverage value.

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

Supplementary Table S 1.12: Variant calling on M13 uses 2D reads starting with the tuned LAST (run uses the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was performed conditioned on the fixed input alignment. Variant calling also used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities were calculated using the EM trained HMM model. Therein, we assume 20% divergence to account for substitutions between the mutated reference and the true reference. For variant calling, we used the base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference that get called as variants in aligned reads. 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. There are three different samplings of the reads at each coverage value.

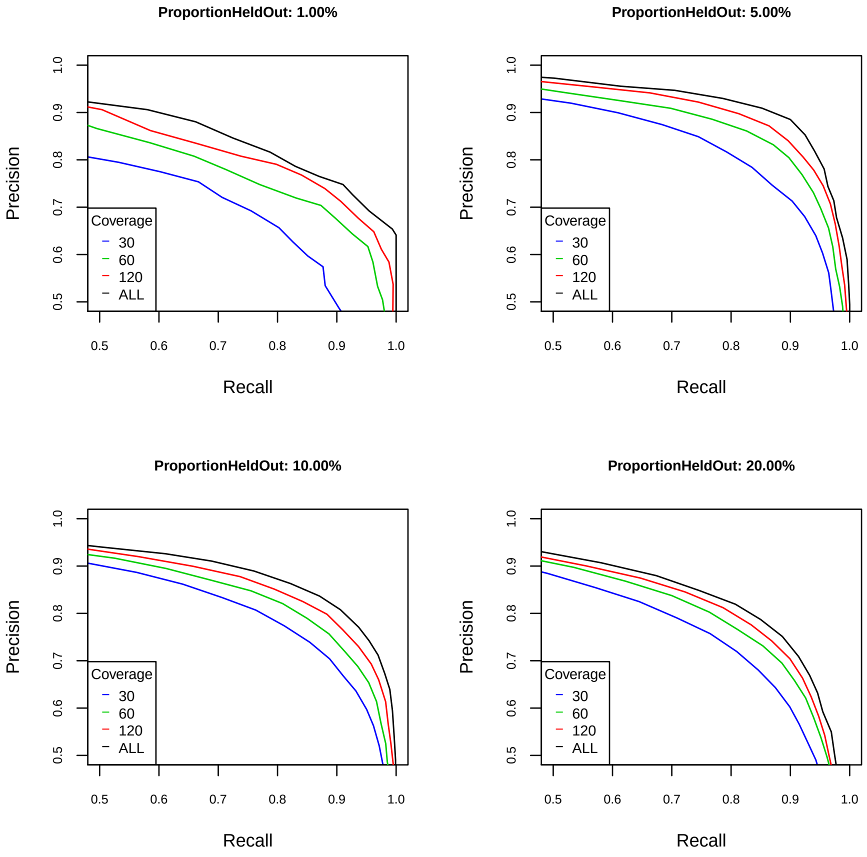


Supplementary Fig. S 1.28: Precision/recall curves show variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling used 2D reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. Variant calling was conditioned on the fixed input alignment. Variant calling also used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities were calculated using the EM trained HMM model. Therein, we assume 20% divergence to account for substitution differences between the mutated reference and the true reference. For variant calling, we used the base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of mutated sites in the reference that get called as variants in aligned reads. 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. There are three different samplings of the reads at each coverage value.

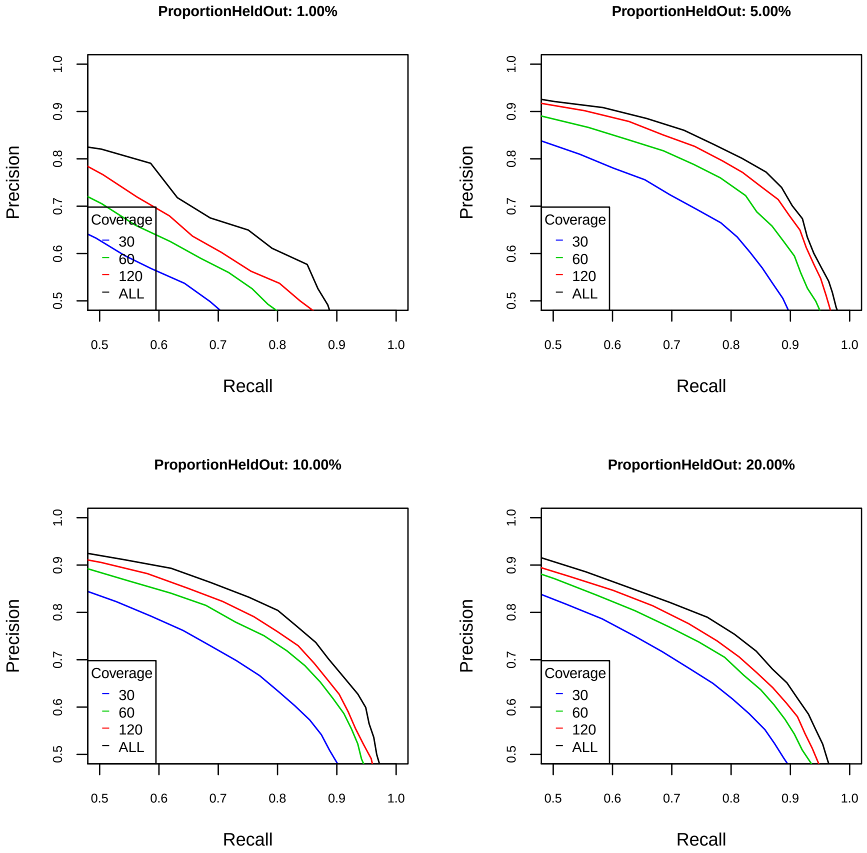
SNV detection using complement reads

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

Supplementary Table S 1.13: Variant calling on M13 uses complement reads starting with the tuned LAST (run uses the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. We performed variant calling by using posterior match probabilities to integrate over all possible read alignments to the mutated reference. We used the initial guide alignment to band the calculations. Variant calling also used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities were calculated using the EM trained HMM model. Therein, we assume 20% divergence to account for substitutions between the mutated reference and the true reference. For variant calling, we used the base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference that get called as variants in aligned reads. 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. There are three different samplings of the reads at each coverage value.



Supplementary Fig. S 1.29: Precision/recall curves show variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling used complement reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. For variant calling, we used posterior match probabilities to integrate over every possible read alignment to the mutated reference. The initial guide alignment was used to band the calculations. Variant calling also used a trained substitution matrix to calculate the maximum likelihood base (see method description). Posterior match probabilities were calculated using the EM trained HMM model. Therein, we assume 20% divergence to account for substitution differences between the mutated reference and the true reference. For variant calling, we used the base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of mutated sites in the reference that get called as variants in aligned reads. 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. There are three different samplings of the reads at each coverage value.



Supplementary Fig. S 1.30: Precision/recall curves show variant calling performance for four different mutation frequencies: 1, 5, 10 and 20 percent. Variant calling used template reads starting with the tuned LAST (run using the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. For variant calling, we used posterior match probabilities to integrate over every possible read alignment to the mutated reference. The initial guide alignment was used to band the calculations. This variant strategy corresponds to choosing the maximum-frequency/expectation of a non-reference base. Posterior match probabilities were calculated using the EM trained HMM model. Therein, we assume 20% divergence to account for substitution differences between the mutated reference and the true reference. For variant calling, we used the base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of mutated sites in the reference that get called as variants in aligned reads. 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. There are three different samplings of the reads at each coverage value.

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

Supplementary Table S 1.14: Variant calling on M13 uses template reads starting with the tuned LAST (run uses the '-s 2 -T 0 -Q 0 -a 1' flags) mapping algorithm. We performed variant calling by using posterior match probabilities to integrate over all possible read alignments to the mutated reference. We used the initial guide alignment to band the calculations. This strategy corresponds to choosing the maximum frequency/expectation of a non-reference base. Posterior match probabilities were calculated using the EM trained HMM model. Therein, we assume 20% divergence to account for substitutions between the mutated reference and the true reference. For variant calling, we used the base calling probability threshold that gives the optimal F-score. Mutation frequency is the approximate proportion of sites mutated in the reference that get called as variants in aligned reads. 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. There are three different samplings of the reads at each coverage value.

Chapter 2

Linear Assembly of a Human Y Centromere using Nanopore Long Reads

Miten Jain[1,*], Hugh E Olsen[1,*], Daniel J Turner[2], David Stoddart[2], Benedict Paten[1], David Haussler[1], Huntington F Willard[3], Mark Akeson[1], and Karen H Miga[1]

[1]UC Santa Cruz Genomics Institute, Santa Cruz, California, USA. [2]Oxford Nanopore Technologies, Oxford, UK. [3]Department of Human Genetics, The University of Chicago, Chicago, Illinois. [*]These authors contributed equally to this work. Correspondence should be addressed to K.H.M. (khmiga@soe.ucsc.edu).

Abstract

The human genome remains incomplete due to the challenge of assembling long tracts of near-identical tandem repeats, or satellite DNAs, that are highly enriched in centromeric regions. To address this, we have implemented a nanopore long-read sequencing strategy to generate high-quality reads capable of spanning hundreds of kilobases of highly repetitive DNAs. Here, we use this advance to produce an initial sequence assembly and characterization of the centromeric region of a human Y chromosome.

Sequence-based studies rely on accurate chromosome assemblies to explore genome biology and function. However, most complex genomes that have been sequenced remain incomplete due to the inability to generate true, haplotype-resolved linear assemblies of centromeric regions, which are known to be enriched in long, often multi-megabase sized tracts of near-identical tandem repeats, or satellite DNAs [108]. Efforts to resolve these regions capitalize on a small number of sparsely arranged sequence variants that offer unique markers to break the repeat monotony and ensure proper overlap-layout-consensus assembly DNAs [109, 110]. Identifying and spanning sequence variants that may be spaced hundreds of kilobases away within a given array requires long and highly accurate sequence reads. Achieving this will involve an advancement in standard single-molecule sequencing, which is error-prone and offers a low throughput of sufficiently long-reads (100 kb+) [9, 111].

Here we present a strategy that generates long reads capable of spanning the complete sequence insert of bacterial artificial chromosomes (BACs) that are hundreds of kilobases in length (∼100-300kb). We demonstrate that these reads are sufficient to resolve the linear ordering of repeats within a single satellite array on the Y chromosome, allowing the first complete sequence characterization of a human centromere.

BACs used in this study were previously determined to span the centromere (DYZ3) locus, and are known to contain centromeric alpha satellite DNAs [112]. Notably, DYZ3 sequences, unlike shorter satellite DNAs [113, 114], have been observed to be stable and cloned without bias [115, 116]. To bypass the challenge of repeat assembly within each BAC, we generated long reads with nanopore sequencing (MinION sequencing device, Oxford Nanopore Technologies), capable of traversing the entire BAC insert. To do so, we optimized a transposase-based method (1D Longboard Strategy) designed to linearize the BAC with a single cut-site. This resulted in a linearization of the BAC followed by addition of the necessary sequencing adaptors (as described in Figure 2.1a and Material and Methods). Plotting read lengths to evaluate nanopore sequencing yield reveal an enrichment for complete read lengths of the BAC DNA (i.e. vector and full-length insert) (Figure 2.1b and Figure 2.2). In total, we generated over >3500 full-length 1D reads that span the entirety of 10 BACs (1 control BAC from Xq24 and 9 BACs that mapped to the DYZ3 locus) with MinION sequencing (Table 2.1).

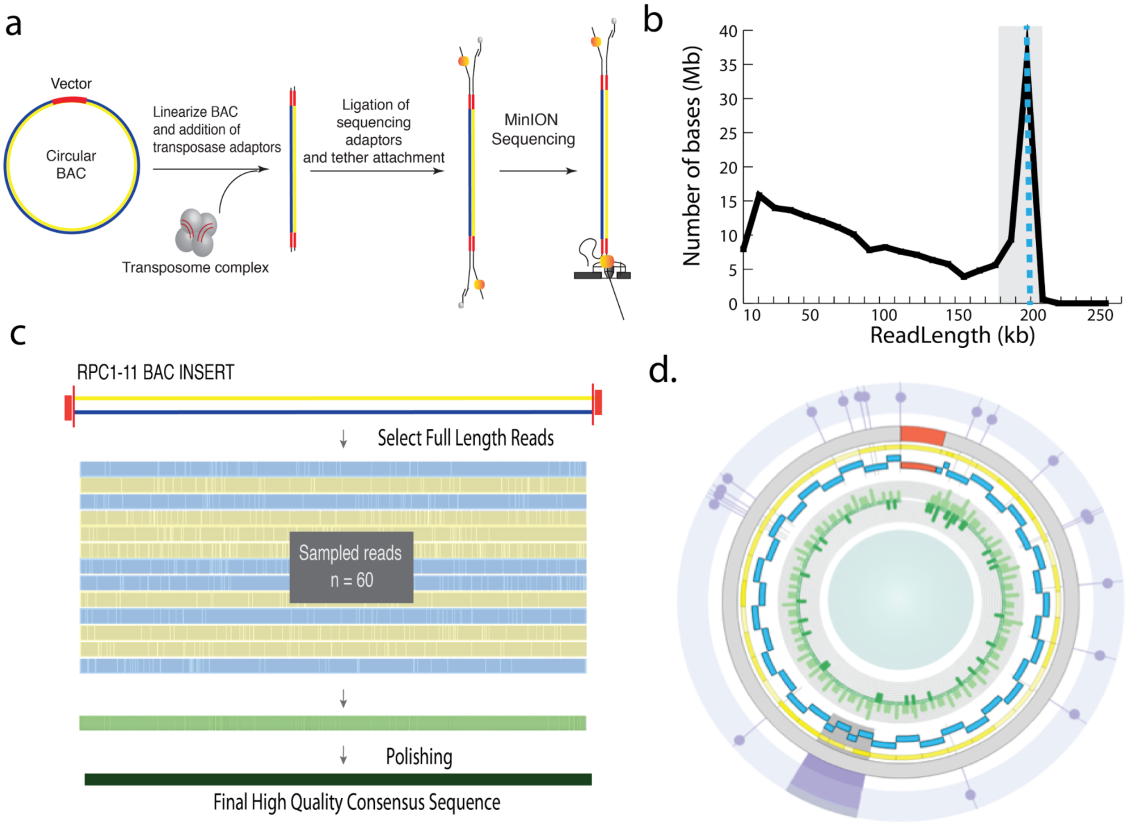


Figure 2.1: BAC based 1D Longboard nanopore sequencing strategy on the MinION. (a) Optimized strategy to cut each circular BAC once with transposase, resulting in a linear and complete DNA fragment of the BAC. After ligation of sequencing adaptors we perform MinION sequencing. (b) Yield plots of BAC DNA (RP11-648J18) provide enrichment, or peaks, supporting BAC lengths. Shading demonstrates the selection of a narrow range of read lengths used in deriving the consensus, the blue dotted line reveals the median value within the selected region providing the closest estimate of insert size. (c) To generate the high quality consensus sequence for each BAC we performed multiple alignment of 60 full length 1D reads (shown as blue and yellow for both orientations) sampled at random with 10 iterations, followed by polishing steps (green) with the entire nanopore long read data and Illumina data. (d) A Circos representation of the polished RP11-718M18 BAC consensus sequence (insert shown in grey: 217 kb, vector in red: 8.8 kb). Blue boxes indicate the positions of each 5.7kb DYZ3 repeat found in a head-to-tail orientation. Purple shading indicates low copy variants, marking tandem DYZ3 repeat structural variants (6 kb).

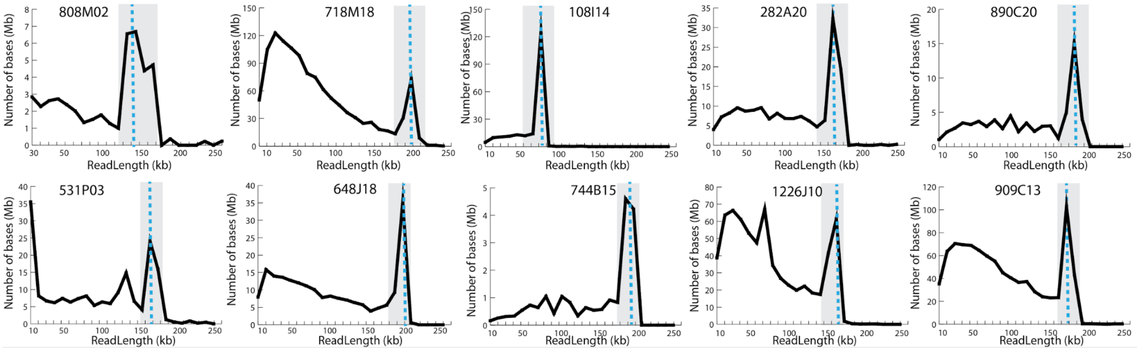


Figure 2.2: MinION yield versus read length. Each subpanel corresponds to the yield in megabases vs read length for a particular BAC with the selected sequence used to generate the consensus sequence highlighted in grey and blue dotted lines providing information for the median value, or expected size of the full-length BAC.

Table 2.1: Throughput for each of the BAC runs.

BAC ID	Type	# reads	Yield (Mb)	Selected full-length Range (kb)	# reads	Yield (Mb)	Consensus length (bp)
108I14	Yp-arm	21143	663	75-85	2172	174	80455
531P03	DZY3 array; unknown order	25666	484	150-165	140	22	160837
808M02	DZY3 array; unknown order	25981	180	140-170	154	24	148529
1226J10	Yq-arm+HSATIII	67218	1392	150-170	858	138	160770
909C13	DZY3 array; unknown order	56610	1695	160-180	727	126	174237
744B15	DZY3 array; unknown order	22017	778	170-190	98	18	180480
890C20	DZY3 array; unknown order	10584	419	170-185	76	14	179559
648J16	DZY3 array; unknown order	20675	561	180-190	59	109	185171
718M18	DZY3 array; unknown order	81594	2704	190-210	596	121	204864
482A22	Control (Xq24)	32386	2426	150-200	1483	254	170145

BAC-based assembly across the DYZ3 locus requires overlap among a few informative sequence variants, thus placing great importance on the accuracy of base-calls. Individual 1D reads (MinION R9.4 chemistry) provide inadequate sequence identity to ensure proper assembly [9, 111]. Using individual reads obtained from a control BAC

(Xq24; RP11-482A22) we observed a median alignment identity of 84.8% and estimated the median rates of insertions, deletions, and mismatches to be 3.6%, 4.6%, and 3.4% respectively. These measurements are consistent with those observed for the nanopore human genome [111]. To improve the overall base quality we derived a consensus from multiple alignments of 1D reads that span the full insert length for each BAC (Material and Methods). We found that we were able to improve the consensus quality with modest coverage increase and sampling (multiple alignments from 60 full-length reads with 10 iterations) (Figure 2.1c). Additional polishing steps were performed using realignment of all full-length nanopore reads for each BAC to improve consensus sequence base quality (99.2% observed for control BAC, RP11-482A22; and an observed range of 99.4 - 99.8% for vector sequences in DYZ3-containing BACs; Material and Methods).

To provide a truth set of repeat sequence variants and to evaluate any inherent nanopore sequence biases, we performed Illumina high-coverage BAC resequencing (coverage range: ∼700-2400; Material and Methods). In reference to the Illumina 5 base pair frequencies for each BAC, we observed homopolymer improvement in our consensus alignment strategy compared to the initial 1D reads (Figure 2.3). DYZ3 satellite repeat copy number, as determined by Illumina read depth were concordant with estimates obtained from nanopore BAC consensus sequences (r = 0.97; Figure 2.4a). Sequence coverage observed for the vector in Illumina datasets were used to confidently identify and validate single-copy sequence variants within each BAC (illustrated for RP11-718M18; Figure 2.4b,c; Material and Methods). Using a k-mer strategy (where k=21 bp), we observed an average positive prediction value of 95.8 between Illumina and our polished sequences, allowing us to identify and mask all sites not supported by Illumina reads as a false positive variant. Finally, standard quality polishing with pilon [117] was applied strictly to unique (that is, non-satellite DNA) sequences on the proximal p and q arm to improve final quality. Alignment of polished consensus sequences from our control BAC from Xq24 (RP11-482A22) and non-satellite DNA in the p-arm adjacent to the centromere (Yp11.2, RP11-531P03), revealed base-quality improvement to > 99% identity. Given the improvements of this strategy, we present high-quality BAC sequences that describe long-range repeat structure, (e.g. 217 kb for RP11-718M18) to guide the ordered assembly of BACs from p-arm to q-arm, spanning the Y centromere (Figure 2.1d).

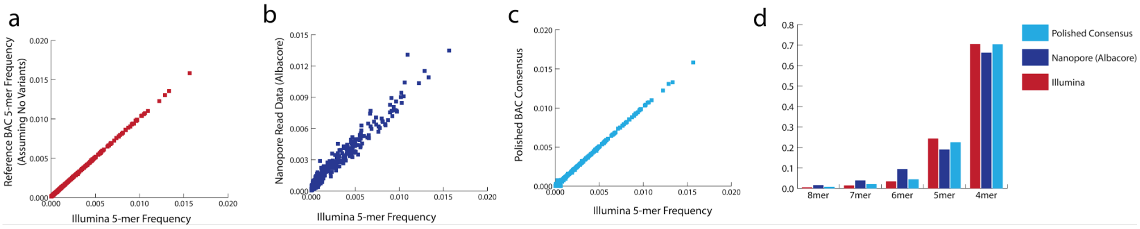


Figure 2.3: Comparisons of 5-mer enrichment and biases. As shown for the RP11-718M18 data, (a) Illumina 5-mer frequencies relative to 5-mer frequencies for a synthetic BAC insert assuming no variants, (b) Illumina 5-mer frequency relative to the reads obtained from the MinION (Albacore base-calling), (c) Illumina 5-mer frequency relative to consensus polished RP11-718M18 nanopore data, (d) investigating the expected proportion of homopolymers (AAAA/TTTT), we observe a correction in calls in our polishing step, which agrees with data from the Illumina read database.

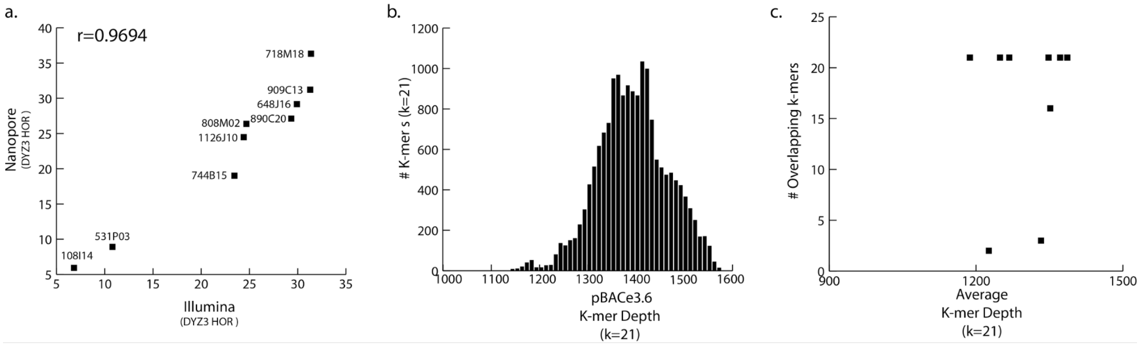
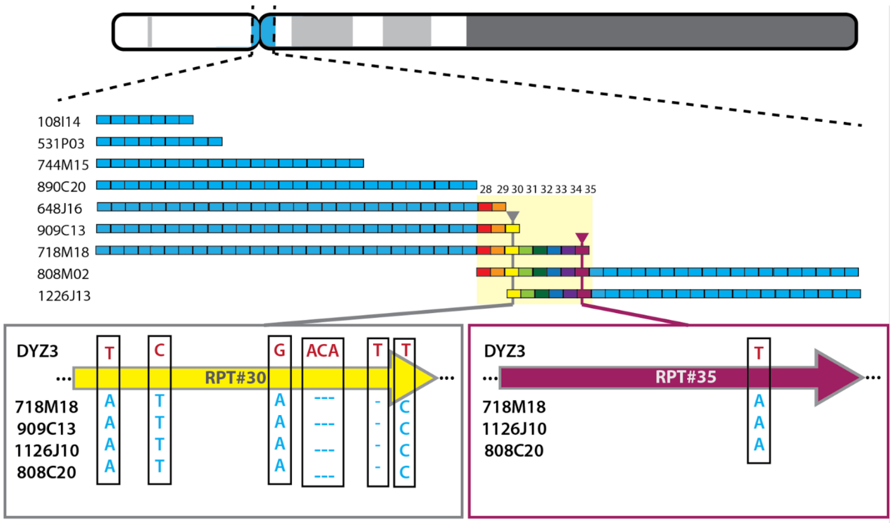


Figure 2.4: Illumina BAC resequencing data are concordant for DYZ3 repeat copy estimates. (a) Using k-mer (where k=21) counts for sites specific to the vector sequence we can determine a range of expected depth or frequency to identify single copy sites for each BAC sequence library, as shown for RP-11 718M18 (b) Overlapping 21-mers with frequency counts within the range of the vector sequence for each BAC library is useful in identifying informative satellite variants (c).

We predicted the ordering of nine BACs using 38 sequence variants (with emphasis on 7 variants that ensured proper overlap between BACs spanning the p-arm and q-arm, as shown in Figure 2.5) resulting in 354,250 bp of assembled centromeric alpha satellite DNA (Figure 2.6). The majority of the centromeric sequence is defined by a 302 kb array of a 5.7 kb DYZ3 higher-order repeat (HOR) [118, 119] that, at the genomic level, is functionally associated with kinetochore assembly and centromere identity [116, 120]. The predicted length of the RP11 Y centromere is consistent with DYZ3 array estimates for 96 individuals from the same Y haplogroup (R1b) (Figure 2.7; mean: 315 kb; median: 350 kb) [121, 122]. This finding is in agreement with pulse-field gel electrophoresis (PFGE) DYZ3 size estimates presented in previous physical mapping of the Y centromeric region [119, 123, 124]. Using a Y-haplogroup matched cell line [125], we find concordant PFGE array size estimates across 6 restriction digests with our RP11 centromere Y length prediction (Figure 2.8). Pairwise comparison between 52 HOR repeats in the DYZ3 array reveal limited sequence divergence between copies (average: 99.7%), as expected for highly homogenized HORs. Further, in agreement with previous assessment of sequence variation within the DYZ3 array [118, 119], we detect a 6 kb HOR structural variant, and provide evidence for 9 copies that are, in all but one occasion, found in tandem [119]. We defined nine DYZ3 repeat haplotypes using linkage between variant bases that are frequent in the array, revealing three local blocks that are enriched for distinct haplotype groups, consistent with previous demonstrations of short-range homogenization of satellite DNA sequence variants [119, 124].



```
718M18      GAATGCTTCTGTGTAGCTTTAATATGAAGACATTTAGTTTT
909C13      GAATGCTTCTGTGTAGCTTTAATATGAAGACATTTAGTTTT
1126J10     GAATGCTTCTGTGTAGCTTTAATATGAAGACATTTAGTTTT
808C20      GAATGCTTCTGTGTAGCTTTAATATGAAGACATTTAGTTTT

718M18      TCAAATGGAAGGTTCAAAACTGTGACATGAATGCCCACATC
909C13      TCAAATGGAAGGTTCAAAACTGTGACATGAATGCCCACATC
1126J10     TCAAATGGAAGGTTCAAAACTGTGACATGAATGCCCACATC
808C20      TCAAATGGAAGGTTCAAAACTGTGACATGAATGCCCACATC

718M18      TGAGTGCACAAATCACAAAGAAGTTTCTCAAAATGCTTCTG
909C13      TGAGTGCACAAATCACAAAGAAGTTTCTCAAAATGCTTCTG
1126J10     TGAGTGCACAAATCACAAAGAAGTTTCTCAAAATGCTTCTG
808C20      TGAGTGCACAAATCACAAAGAAGTTTCTCAAAATGCTTCTG

718M18      TCCACTTTCAGATTCT---ACAAGAGAGGTTCAAAACTA
909C13      TCCACTTTCAGATTCT---ACAAGAGAGGTTCAAAACTA
1126J10     TCCACTTTCAGATTCT---ACAAGAGAGGTTCAAAACTA
808C20      TCCACTTTCAGATTCT---ACAAGAGAGGTTCAAAACTA

718M18      AAACACATCACAAATAAG-TTCC
909C13      AAACACATCACAAATAAG-TTCC
1126J10     AAACACATCACAAATAAG-TTCC
808C20      AAACACATCACAAATAAG-TTCC

718M18      TCGGAATTCTTCTGTGTAGTATTTATGTGAAGATATTTCCT
1126J10     TCGGAATTCTTCTGTGTAGTATTTATGTGAAGATATTTCCT
808C20      TCGGAATTCTTCTGTGTAGTATTTATGTGAAGATATTTCCT
```

Figure 2.5: Evidence for satellite variants in overlap region between repeats 28-35. Informative variants useful in ensuring proper overlap are shown for repeat 30 (yellow) and repeat 35 (light purple). Support for seven variant positions are shown for all BACs (blue) with the reference base indicated in red. Relevant alignments for each variant are provided with shared variant bases/positions indicated in red.

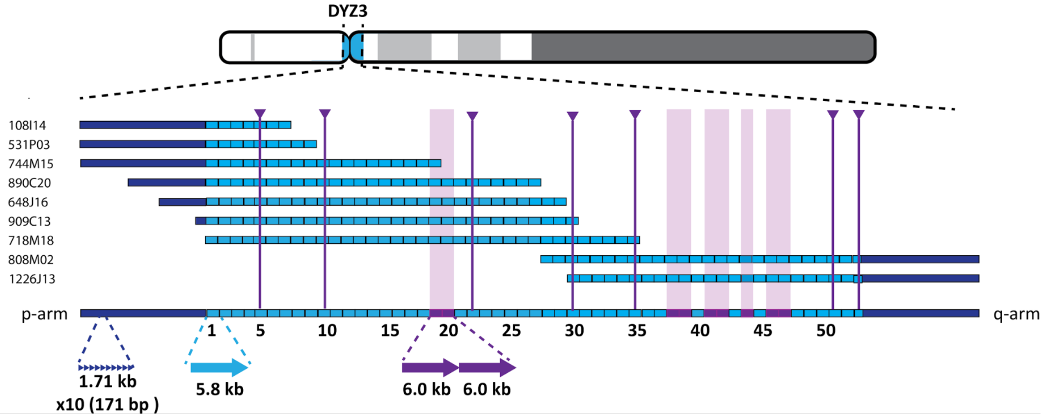


Figure 2.6: Linear assembly of the RP11 Y centromere. Ordering of nine DYZ3-containing BACs spanning from proximal p-arm to proximal q-arm provided evidence for a 354,250 bp region enriched in alpha satellite DNA. Vertical purple lines and shaded boxes indicate DYZ3 satellite repeats that contain at least one single copy variant used to ensure proper BAC overlap-layout-consensus assembly. Highly divergent monomeric alpha satellite (∼171 bp, dark blue), indicative of the edges of the otherwise highly homogeneous array, is observed at > 99% sequence identity with sequences in the GRCh38 assembly. The centromere locus is defined by the DYZ3 conical 5.7 kb higher-order repeat (HOR) (light blue), that is observed in a head to tail orientation from p-arm to q-arm, for a total of 301 kb. Nine HOR variants (6 kb, shown in purple) have been identified, with all but one identified in tandem. DYZ3 HORs were classified into nine haplotypes using four frequent satellite DNA variants in the array (haplotype (H)1 red, H2 orange, H3 yellow, H4 green, H5 blue, H6 dark orange; H7 purple, H8 dark purple, H9 grey). We identified three predominant blocks: H1 proximal to the p-arm (I), H4 in the middle of the array (II), and H5 adjacent to the q-arm (III).

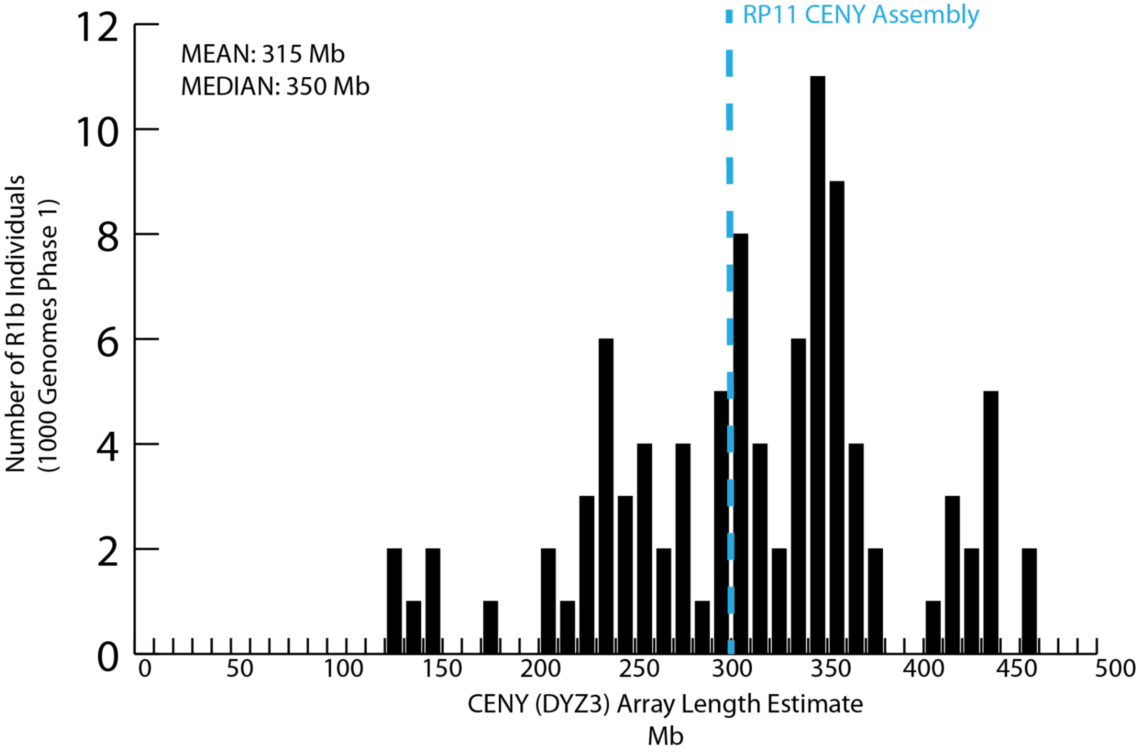
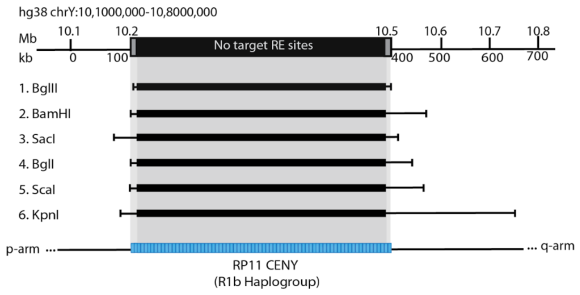


Figure 2.7: Distribution of array lengths estimates for R1b Y-haplogroups from the Phase 1 1000 genome project. The assembled DYZ3 array length for the RP11 donor genome is shown as a dashed blue line.



				R1b Y-Haplogroup	
		hg38_start	hg38_end	RP11_Array Estimate	HuRef PFGE
1	**BglII**	10255458	10617035	269,065	290,000
2	**BamHI**	10237287	10686008	356,209	375,000
3	**SacI**	10214376	10624533	317,645	330,000
4	**BglI**	10243115	10654962	319,335	300,000
5	**ScaI**	10242349	10682450	347,589	339,000
6	**KpnI**	10220391	10855565	542,662	540,000

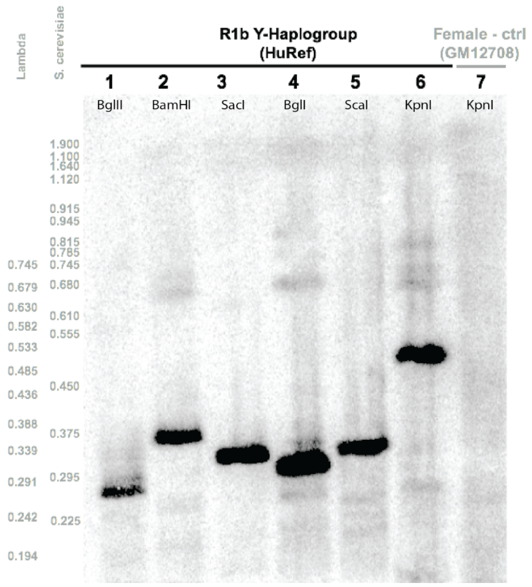


Figure 2.8: DYZ3 array length estimates by pulse field gel electrophoresis (PFGE) Southern using digests with a Y-haplogroup R1b matched individual (HuRef cell line). DNA digest is shown in top panel for 6 enzymes used with corresponding CHEF gel. Lane 7 is used as a negative control (GM12708 female cell line). Size estimates were made using chromosomes from *S. cerevisiae* strain YNN295 and lambda DNA as markers (marker sizes in kilobase pairs at left). Size estimates assuming the RP11 DYZ3 assembly are presented in the table relative to the relative positions of restriction sites in the human reference assembly flanking the centromeric region (GRCh38).

In conclusion, we have implemented a long read strategy to generate high quality, finished sequences to advance sequence characterization of repetitive DNAs. In doing so, we report the long-range repeat organization and structure of a human centromere on chromosome Y. We expect that this work will contribute to ongoing efforts to complete complex genomes.

Material and Methods

A. 1D Longboard MinION Protocol

BAC DNA Preparation and Validation. Bacterial artificial chromosomes (BACs) clones used in this study were obtained from BACPAC RPC1-11 library, Children's Hospital Oakland Research Institute in Oakland, California, USA (http://www.chori.org/bacpac/). BACs that span the human Y centromere: RP11-108I14, RP11-1226J10, RP11-808M2, RP11-531P03, RP11-909C13, RP11-890C20, RP11-744B15, RP11-648J16, RP11-718M18, and RP11-482A22, were determined based on previous hybridization with DYZ3-specific STSs probes sY715 and sY78 [112]. BAC DNA was prepared using the QIAGEN Large-Construct Kit (Cat No./ID: 12462). To ensure removal of the E.coli genome, it was important to include an exonuclease incubation step at 37C for 1 hour, as provided within the QIAGEN Large-Construct Kit. BAC DNAs were hydrated in TE buffer. BAC Insert length estimates were determined by pulsed-field gel electrophoresis (PFGE) (data not shown).

Transposase-mediated 1D long reads. MinIONs can process long fragments, as has been previously documented [9, 111]. While these long reads demonstrate the processivity of nanopore sequencing, they are also few in numbers. To systematically enrich for the number of long reads per MinION sequencing run, we developed a strategy that uses ONT Rapid Sequencing Kit (RAD002). We performed a titration between the transposase from this kit (RAD002) and circular BAC DNA. This was done to achieve conditions that would optimize the probability of circular BAC fragments being cut by the transposase only once. To this end, we diluted the 'live' transposase from the RAD002 kit with the 'dead' transposase provided by ONT. For pulsed-field gel electrophoresis (PFGE) based tests, we used 1 μl of 'live' transposase and 1.5 μl of 'dead' transposase per 200 ng of DNA in a 10 μl reaction volume. This reaction mix was then incubated at 30^{o}C for 1 minute and 75^{o}C for 1 minute, followed by PFGE. Our PFGE tests used a 1% high-melting agarose gels and were run with standard 180^{o} FIGE conditions for 3.5 hours. An example PFGE gel is shown below:

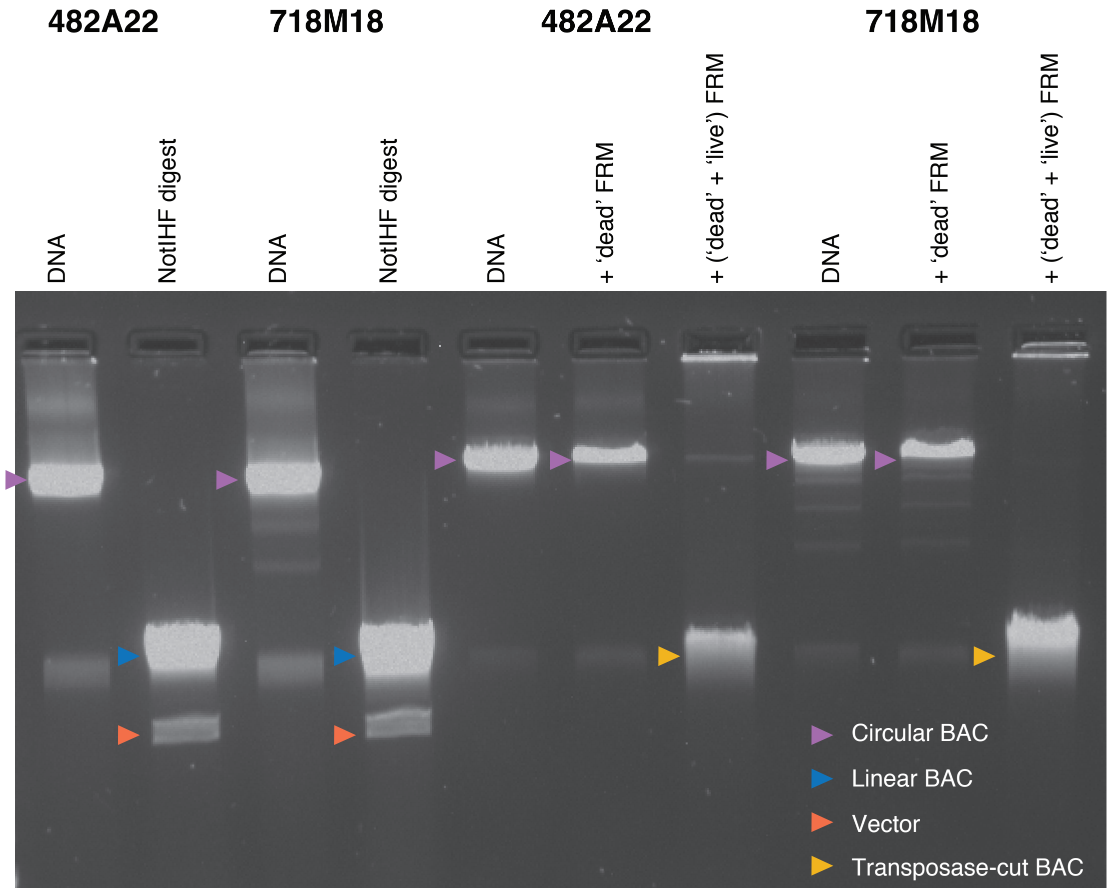


Figure 2.9: Representative gel image from a pulsed-field gel electrophoresis assay to test NotIHF digest of BACs and to assess titration of DNA:ONT transposase ('dead' FRM + 'live' FRM). Data shown for two BACs: RP11-482A22 (∼175 kb Control BAC from Xq24) and RP11-718M18 (∼217 kb DYZ3-containing BAC). Circularized BACs are indicated in purple. High fidelity NotI (NotI-HF; NEB R3189S) was used to identify insert sequence (blue) and vector sequence (orange). Addition of transposase ('dead' FRM + 'live' FRM) indicates that the the majority of linearized DNAs (light orange, transposase-cut BAC) are full-length or only cut once.

For MinION sequencing library preparation, we used 1.5 μl of 'live' transposase and 1 μl of 'dead' transposase (supplied by ONT) per 1 μg of DNA in a 10 μl reaction volume. Briefly, this reaction mix was then incubated at 30°C for 1 minute and 75°C for

1 minute. We then added 1 μl of the sequencing adapter and 1 μl of Blunt/TA Ligase Master Mix (New England Biolabs) and incubated the reaction for 5 minutes. This was the adapted BAC DNA library for the MinION. R9.4 SpotON flow cells were primed using ONT recommended protocol. We prepared 1 ml of priming buffer with 500 μl running buffer (RBF) and 500 μl water. Flow cells were primed with 800 μl priming buffer via the side loading port. We waited for 5 minutes to ensure initial buffering before loaded the remaining 200 μl of priming buffer via the side loading port but with the SpotON open. We next added 35 μl RBF and 28 μl water to the 12 μl library for a total volume of 75 μl. We loaded this library on the flow cell via the SpotON port and proceeded to start a 48 hour MinION run.

When a nanopore run is underway, the amplifiers controlling individual pores can alter voltage to get rid of unadapted molecules which will otherwise block the pore. With R9.4 chemistry, ONT introduced global flicking that reversed the potential every 10 minutes by default to clear all nanopores of all molecules. At 450 bps speed, a 200 kb BAC would take around 7.5 minutes to process. To ensure sufficient time for capturing BAC molecules on the MinION, we changed the global flicking time period to 30 minutes. This is no longer the case with an update to ONT's MinKNOW software, and on the later BAC sequencing runs we did not change any parameters.

B. Multiple alignment strategy and polishing steps to improve sequence quality.

We selected BAC full-length reads as determined by observed enrichment in our yield plots. Full-length reads used in this study were determined to contain at least 3 kb of vector sequence, as determined by BLASR [126] (*-sdpTupleSize 8 -bestn 1 -nproc 8 -m 0*) alignment with the pBACe3.6 vector (GenBank Accession: U80929.2). Reads were converted to the forward strand. Reads were reoriented relative to a fixed 3 kb vector sequence by aligning the transition from vector to insert. In cases where the vector sequence was not identified at the end of the read, the sequence preceding the vector was added to the end of the original sequence. We sampled 60 reoriented reads at random and performed a multiple sequence alignment (MSA) using kalign [127]. We computed the consensus from the MSA using a custom python script (github, needed), where the most prevalent base at each position was called. Gaps were only considered in the consensus if the second most frequent nucleotide at that position was present in less than 10 reads. We performed random sampling followed by MSA iteratively 10x, resulting in a panel of 10 consensus sequences. We next performed an MSA on the collection of consensus sequences to generate a final consensus sequence, as discussed. Consensus sequence polishing was performed by aligning full-length 1D nanopore reads for each BAC to the consensus (BLASR [126], *-sdpTupleSize 8 -bestn 1 -nproc 8 -m 0*). We used pysamstats [128] to identify read support for each base call. We masked consensus bases that had less than 50% support in the total read alignments. Finally, we performed two rounds of pilon [117] polishing using Illumina sequence data (described below) for the non-alpha higher-order regions on the p and q arm that had unique alignments.

We performed Illumina re-sequencing (Miseq V3 600bp; 2 x 100 bp) for all nine DYZ3-containing BACs to validate identified repeat variants in the nanopore consensus sequence and ultimately guide BAC-based assembly of the array. Overall read depth for each BAC was determined by mapping Illumina reads from each BAC library to the pBACe3.6 vector. DYZ3 copy number estimates were determined by the frequency of Illumina reads that mapped to a reference 5785 bp DYZ3 repeat (presented in tandem to remove edge-effects mapping artifacts), multiplied by the total BAC consensus length. DYZ3 copy number in each consensus sequence derived from nanopore reads was determined using HMMER3 [129] (v3.1b2) with a profile constructed from the DYZ3 reference repeat. To characterize individual consensus bases, Illumina read data was reformatted into a k-mer library (where k=21 bp, with 1 bp slide) in forward and reverse orientation. K-mers that matched the pBACe3.6 sequence exactly were labeled as ‘vector’. As the vector sequence is expected to be present once in each BAC the distribution of counts for 21-mers that had an exact match with pBACe3.6 provided a range of k-mer frequency and/or k-mer depth expected for single-copy DNA (as shown in Supplementary Figure 3b for RP11-718M18). DYZ3 repeat variants (satVARs) were determined as 21-mers not identified to have an exact match with either the vector or DYZ3 reference repeat. Single copy satVARs were observed once in the consensus sequence and had a k-mer depth profile in the range of the corresponding BAC vector k-mer distribution. Additionally, satVARs used in overlap-layout-consensus assembly must be supported by 2 or more Illumina 21-mers that support a single-copy site.

C. Prediction and validation of DYZ3 array.

BAC ordering was determined using 34 overlapping informative satVARs (including the nine DYZ3 6 kb structural variants) in addition to alignments directly to either assembled sequence on the p-arm or q-arm of the human reference assembly (GRCh38). Full length DYZ3 HORs (ordered 1-52) were evaluated by MSA (using kalign [127]) between overlapping BACs, with emphasis on repeats 28-35 that define the overlap between BACs anchored to the p-arm or q-arm. RPC1-11 BAC library has been previously referenced as derived from a known carrier of haplogroup R1b [130, 131]. We compared our predicted DYZ3 array length with 93 R1b Y-haplogroup matched individuals by intersecting previously published DYZ3 array length estimates for 1000 genome phase 1 data [121, 122] with donor-matched Y-haplogroup information [132]. To investigate concordancy of our array prediction with previous physical maps of the Y-centromere, we identified the positions of referenced restriction sites that directly flank the DYZ3 array in the human chromosome Y assembly (GRCh38) [119, 124, 123]. It is unknown if previously published individuals are from the same population cohort as the RPC1-11 donor genome. Therefore, we performed similar PFGE DYZ3 array PFGE length estimates using the HuRef B-lymphoblast cell line (now available from Coriell Institute as GM25430), which was previously characterized to be in the R1-b Y-haplogroup [125].

PFGE alpha satellite Southern. High-molecular-weight HuRef genomic DNA was resuspended in agarose plugs using 5e6 cells per 100 μl of 0.75% Clean-Cut Agarose (CHEF Genomic DNA Plug Kits Cat #: 170-3591 BIORAD). A female lymphoblastoid cell line (GM12708) was included as a negative control. Agarose plug digests were performed overnight (8-12hrs) with 30-50U of each enzyme with matched NEB buffer. PFGE Southern experiments used 1/4 - 1/2 agarose plug per lane (with an estimate of 5-10μg) in an 1% SeaKem LE Agarose gel and 0.5 X TBE. CHEF Mapper conditions were optimized to resolve 0.1-2.0 Mb DNAs: voltage 6V/cm, runtime: 26:40 hrs, in angle: 120°, initial switch time: 6.75 s, final switch time: 1m33.69s, with a linear ramping factor. We used the Lambda (NEB; N0340S) and *S. cerevisiae* (NEB; N0345S) as markers. Methods of transfer to nylon filters, prehybridization, and chromosome specific hybridization with 32P-labeled satellite probes have been described [133]. Briefly, DNA was transferred to nylon membrane (Zeta Probe GT nylon membrane; CAT# 162-0196) for ~24hrs. DYZ3 probe (50 ng DNA labelled ~2 cpm/mL; amplicon product using previously published STS DYZ3 Y-A and Y-B primers [134]) was hybridized for 16 hrs at 42°C. In addition to standard wash conditions [133], we performed two additional stringent wash (buffer: 0.1% SDS and 0.1x SSC) steps for 10 min at 72°C to remove non-specific binding. An image was recovered after a 20hr exposure.

Contributions

KM and HW conceived the project KM, MJ, HO, and MA designed the experiments. MJ and HO were involved with BAC sample preparation. MJ and HO performed MinION sequencing and base-calling. MJ and KM analyzed the BAC sequencing data and validation analyses. KM and HW performed the pulse-field gel electrophoresis array length estimates. KM, MJ and HO contributed to analysis and figure generation. MA, BP, DP, and DH provided technical advice; All authors contributed to the writing, editing, and completion of the manuscript.

Chapter 3

Relevant MinION work

Below, I present some of the MinION work that I have contributed to. To keep this writeup succinct, only the abstracts for the research are shown along with the author contributions. I will place website links for all the papers and informatics tools presented in this thesis in the Appendix.

MinION Analysis and Reference Consortium: Phase 1 data release and analysis

Camilla LC Ip[1,*], Matthew Loose[2,*], John R Tyson[3,*], Mariateresa de Cesare[1,*], Bonnie L Brown[4,*], Miten Jain[5,*], Richard M Leggett[6,*], David A Eccles[7], Vadim Zalunin[8], John M Urban[9], Paolo Piazza[1], Rory J Bowden[1], Benedict Paten[5], Solomon Mwaigwisya[10], Elizabeth M Batty[1], Jared T Simpson[11], Terrance P Snutch[3], Ewan Birney[8,*], David Buck[1,*], Sara Goodwin[12,*], Hans J Jansen[13,*], Justin O'Grady[10,*], Hugh E Olsen[5,*], MinION Analysis and Reference Consortium

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]School of Life Sciences, Queens Medical Centre, University of Nottingham, Nottingham, UK. [3]Michael Smith Laboratories and Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, Canada. [4]Virginia Commonwealth University, Richmond, VA, USA. [5]University of California, Santa Cruz, Santa Cruz, CA, USA. [6]The Genome Analysis Centre, Norwich Research Park, Norwich, UK. [7]Malaghan Institute of Medical Research, Wellington, New Zealand. [8]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. [9]Division of Biology and Medicine, Brown University, Providence, RI, USA. [10]Norwich Medical School, University of East Anglia, Norwich, UK. [11]Informatics and Biocomputing, Ontario Institute for Cancer Research, ON, Canada. [12]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [13]ZF-screens B.V., Leiden, Netherlands. [*]Equal contributors. Correspondence should

be addressed to Camilla LC Ip (camilla.ip@well.ox.ac.uk), Matthew Loose (loose@nottingham.ac.uk), John R Tyson (jtyson@msl.ubc.ca), Mariateresa de Cesare (decesare@well.ox.ac.uk), Ewan Birney (birney@ebi.ac.uk), David Buck (dbuck@well.ox.ac.uk), Sara Goodwin (sgoodwin@cshl.edu), Hans J Jansen (jansen@zfscreens.com), Justin O'Grady (Jusitn.OGrady@uea.ac.uk), or Hugh E Olsen (heolsen@soe.ucsc.edu).

Abstract

The advent of a miniaturized DNA sequencing device with a high-throughput contextual sequencing capability embodies the next generation of large scale sequencing tools. The MinION^{TM} Access Programme (MAP) was initiated by Oxford Nanopore Technologies^{TM} in April 2014, giving public access to their USB-attached miniature sequencing device. The MinION Analysis and Reference Consortium (MARC) was formed by a subset of MAP participants, with the aim of evaluating and providing standard protocols and reference data to the community. Envisaged as a multi-phased project, this study provides the global community with the Phase 1 data from MARC, where the reproducibility of the performance of the MinION was evaluated at multiple sites. Five laboratories on two continents generated data using a control strain of *Escherichia coli* K-12 by preparing and sequencing samples according to a revised ONT protocol. Here, we provide the details of the protocol used, along with a preliminary analysis of the characteristics of typical runs including the consistency, rate, volume, and quality of data produced. Further analysis of the Phase 1 data are presented here, and additional experiments in Phase 2 of *E. coli* from MARC are already underway to

identify ways to improve and enhance MinION performance.

Contributions

EB coordinated the study. EB, DB, JT, JOG, and BB designed the study. MdC, PP, DB, SG, JOG, RML, SM, HJ, HEO, and MJ designed the MARC protocol and performed the experiments. VZ, MJ, BP, CI, ML, and RML collated the data for the group and ran bioinformatics pipelines over the data. CI, ML, RML, MJ, BP, EB, RB, LB, and JT analysed the data. CI, RB, DB, EB, ML, RML, MJ, BP, HEO, PP, MdC, MS, JU, JOG, SG, JT, TPS, BB, and DE drafted the manuscript. All authors participated in discussions relating to the generation and analysis of the data.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Synopsis

The MARC consortium comprises of laboratories from all over the globe. In this work, we performed biological and technical replicate experiments using the MinION in five different laboratories. We then performed marginAlign [9] analyses (alignments, EM, statistics) on these data. The data were achieved using R7.3 chemistry-based experiments and were termed as Phase 1 release from MARC.

I helped with data generation and performed MinION sequencing experiments using genomic DNA from *E. coli.* I also helped with data uploads and downloads, and sequence data extraction. In addition, I performed sequence alignments using marginAlign [9] (with bwa [99] chaining and bwa [99] EM) to estimate the error model from the data and to compute alignment statistics. I also helped with writing the manuscript.

Whole genome sequencing and assembly of *Caenorhabditis elegans* genomes using the MinION sequencing device

JR Tyson[1,*], NJ O'Neil[2,*], M Jain[3,*], HE Olsen[3], P Hieter[4], TP Snutch[1]

[1]Michael Smith Laboratories and Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, Canada. [2]Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada V6T 1Z4. [3]UC Santa Cruz Genomics Institute and Department of Biomolecular Engineering, University of California, Santa Cruz, California, USA. [4]Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada V6T 1Z4; Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada V6T 1Z3. Correspondence should be addressed to T.P.S. (snutch@mail.ubc.ca).

Abstract

Advances in 3rd generation sequencing have opened new possibilities for 'bench-top' whole genome sequencing. The MinION is a portable device that uses nanopore technology and can sequence long DNA molecules. MinION long sequence reads are well suited for *de novo* assembly of novel complex genomes as they facilitate the construction of highly contiguous physical genome maps obviating the need for labor-intensive physical genome mapping. MinION derived contigs can be polished using highly accurate Illumina derived sequence data to generate an accurate highly contiguous genome sequence. To assess the feasibility of this hybrid sequencing approach to de novo assembly of large complex genomes, we sequenced the genome of two *Caenorhabditis*

elegans strains, a wild type strain and a strain containing two complex rearrangements. MinION sequence data was used to assemble a highly contiguous wild type *C. elegans* genome containing 55 contigs (N50 contig length = 3.0 Mb) that covered >99% of the 100,286,401 base reference genome. MinION sequence data and *de novo* genome assembly also identified complex rearrangements in the mutant strain. This demonstrates that large complex genomes can be assembled from MinION data and that the long reads of MinION sequencing can be used to elucidate complex genomic rearrangements.

Contributions

JRT and NO'N performed the experiments and analysis. MJ and HEO helped with assembly and sequence-level analysis. TPS directed the project. All authors contributed to writing the manuscript.

Synopsis

In this work, we used MinION data to understand genome complexities in the worm genome. We were able to improve high-quality, contiguous assemblies using MinION reads. This assembly could then be polished to almost 100% completion using Illumina short-read data.

I helped with running the genome assemblies using SPAdes [135] (for Illumina data) and Canu [136] (for nanopore data). For Illumina data, I helped with quality-value based filtering of data for assembly. I assessed assembly quality using QUAST [137] and performed sequence alignments between the various assemblies and the sequence data (both nanopore and Illumina) using marginAlign [9] (bwa [99]). Using these alignments and QUAST [137] output, I calculated alignment and summary statistics. I also helped with writing the manuscript.

Mapping DNA methylation with high-throughput nanopore sequencing

Arthur C Rand[1,2], Miten Jain[1,2], Jordan M Eizenga[1,2], Audrey Musselman-Brown[1], Hugh E Olsen[1], Mark Akeson[1] & Benedict Paten[1]

[1]Department of Biomolecular Engineering and Genomics Institute, University of California, Santa Cruz, Santa Cruz, California, USA. [2]These authors contributed equally to this work. Correspondence should be addressed to B.P. (benedict@soe.ucsc.edu).

Abstract

DNA chemical modifications regulate genomic function. We present a framework for mapping cytosine and adenosine methylation with the Oxford Nanopore Technologies MinION using this nanopore sequencer's ionic current signal. We map three cytosine variants and two adenine variants. The results show that our model is sensitive enough to detect changes in genomic DNA methylation levels as a function of growth phase in *Escherichia coli*.

Contributions

BP conceived of the experiments. BP and MA directed the research. ACR implemented the models and performed analysis. MJ and HEO performed the sequencing experiments and performed sequence data analysis. JME implemented the HDP model and Gibbs sampler. AM-B performed initial experiments. All authors contributed to writing the manuscript.

Synopsis

In this work, we demonstrated that MinION sequencing can detect cytosine and adenine DNA methylation using synthetic DNA and genomic DNA. We used synthetic DNA constructs to discriminate among 3 C-5 variants of Cytosine (C, 5-mC, and 5-hmC) at 80% median accuracy [138].

Using this approach for genomic *E. coli*, we mapped the methylation status (C vs. 5-mC) for 96% of cytosines in the C$\underline{\text{C}}$[A/T]GG context (underlined C being probed for methylation). We also mapped the methylation status for 86% of adenines (A vs. 6-methyladenine (6-mA)) in the G$\underline{\text{A}}$TC context in pUC19 plasmid DNA. The methylation status was mapped using modest coverage (20X for cytosines and 40X for adenines).

I helped with designing and executing MinION sequencing experiments that used synthetic DNA substrates and *E. coli* (genomic and whole-genome amplified). The genomic DNA runs included various growth phases of *E. coli* (0.4 OD, 0.8 OD, and stationary phase). I performed sequence-level data analysis using marginAlign [9] to estimate error-rates from these data, as well as to compute alignment-level statistics.

Nanopore sequencing and assembly of a human genome with ultra-long reads

M Jain[1,§], S Koren[2,§], J Quick[3,§], AC Rand[1,§], TA Sassani[4,5,§], JR Tyson[7,§], AD Beggs[8], AT Dilthey[2], IT Fiddes[1], S Malla[9], H Marriott[9], KH Miga[1], T Nieto[8], J O'Grady[10], HE Olsen[1], BS Pederson[4,5], A Rhie[2], H Richardon[10], AR Quinlan[4,5,6], TP Snutch[7], L Tee[8], B Paten[1], AM Phillippy[2], JT Simpson[11,12], NJ Loman[3,*], M Loose[9,*]

[1]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA. [2]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA. [3]Institute of Microbiology and Infection, University of Birmingham, Birmingham B15 2TT, UK. [4]Department of Human Genetics, University of Utah, Salt Lake City, UT, USA. [5]USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT, USA. [6]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA. [7]Michael Smith Laboratories and Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, Canada. [8]Surgical Research Laboratory, Institute of Cancer & Genomic Science, University of Birmingham, UK. [9]DeepSeq, School of Life Sciences, University of Nottingham, UK. [10]Norwich Medical School, University of East Anglia, Norwich, UK [11]Ontario Institute for Cancer Research, Toronto M5G 0A3, Canada. [12]Department of Computer Science, University of Toronto, Toronto M5S 3G4, Canada. [§]These authors contributed equally to this work. [*]Authors for correspondence n.j.loman@bham.ac.uk, matt.loose@nottingham.ac.uk

Abstract

Nanopore sequencing is a promising technique for genome sequencing due to its portability, ability to sequence long reads from single molecules, and to simultaneously assay DNA methylation. However, until recently, nanopore sequencing has been mainly applied to small genomes, due to the limited output attainable. We present nanopore sequencing and assembly of the GM12878 Utah/Ceph human reference genome generated using the Oxford Nanopore MinION and R9.4 version chemistry. We generated 91.2 Gb of sequence data (30x theoretical coverage) from 39 flowcells. *De novo* assembly yielded a highly complete and contiguous assembly (NG50 3Mb). We observed considerable variability in homopolymeric tract resolution between different basecallers. The data permitted sensitive detection of both large structural variants and epigenetic modifications. Further, we developed a new approach exploiting the long read capability of this system and found that adding an additional 5-coverage of 'ultra-long' reads (read N50 of 99.7kb) more than doubled the assembly contiguity. Modeling the repeat structure of the human genome predicts extraordinarily contiguous assemblies may be possible using nanopore reads alone. Portable *de novo* sequencing of human genomes may be important for rapid point-of-care diagnosis of rare genetic diseases and cancer, and monitoring of cancer progression. The complete dataset including raw signal is available as an Amazon Web Services Open Dataset at: `https://github.com/nanopore-wgs-consortium/NA12878`.

Contributions

NJL, ML, JTS, and JRT conceived the study. JQ developed the long read protocol. ADB, MJ, ML, HM, SM, TN, JO'G, JQ, HR, JRT, and LT prepared materials and/or performed sequencing. ATD, ITF, MJ, SK, NJL, ML, KHM, HEO, BP, BSP, AMP, ARQ, ACR, AR, TAS, JTS, and JRT performed bioinformatics analysis and wrote or modified software. ITF, MJ, SK, NJL, ML, KHM, JO'G, HEO, BP, AMP, JQ, ARQ, ACR, TAS, JTS, TPS, and JRT wrote and edited the manuscript. All authors approved the manuscript and provided strategic oversight for the work.

Synopsis

Over the past three years, MinION nanopore sequencing improved in both sequence quality and throughput. The Nanopore Human Genome consortium sequenced the GM12878 human genome using 44 flow cells. Six laboratories (hailing from the UK, Canada, and the USA) generated the sequence data. The sequencing was performed by eight laboratories that are spread across the UK, Canada, and USA. The data analysis was then performed by several laboratories in tandem.

I helped generate sequence data at UCSC. I used marginAlign EM [9] on these data to estimate the error model. Additionally, I performed kmer analysis to understand under and over-represented 5mers in the data. I also compared the sequence quality from the different basecallers that were used as part of the study. These included sequence calls from Metrichor (ONT's cloud basecaller), Nanonet (ONT's open-source basecaller), and Scrappie (ONT's homopolymer basecaller). I also helped with writing the manuscript.

Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing

Andrew M Smith[1], Miten Jain[1], Logan Mulroney[1], Daniel R Garalde[2] and Mark Akeson[1]

[1]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, USA 95064. [2]Oxford Nanopore Technologies, Oxford, UK. Correspondence should be addressed to M.A. (makeson@soe.ucsc.edu)

Abstract

The ribosome small subunit is expressed in all living cells. It performs numerous essential functions during translation, including formation of the initiation complex and proofreading of base-pairs between mRNA codons and tRNA anticodons. The core constituent of the small ribosomal subunit is a 1.5 kb RNA strand in prokaryotes (16S rRNA) and a homologous 1.8 kb RNA strand in eukaryotes (18S rRNA). Traditional sequencing-by-synthesis (SBS) of rRNA genes or rRNA cDNA copies has achieved wide use as a 'molecular chronometer' for phylogenetic studies [139], and as a tool for identifying infectious organisms in the clinic [140]. However, epigenetic modifications on rRNA are erased by SBS methods. Here we describe direct MinION nanopore sequencing of individual, full-length 16S rRNA absent reverse transcription or amplification. As little as 5 picograms (∼10 attomole) of *E. coli* 16S rRNA was detected in 4.5 micrograms of total human RNA. Nanopore ionic current traces that deviated from canonical patterns

revealed conserved 16S rRNA base modifications, and a 7-methylguanosine modification that confers aminoglycoside resistance to some pathological *E. coli* strains. This direct RNA sequencing technology has promise for rapid identification of microbes in the environment and in patient samples.

Contributions

AMS designed and performed RNA bench experiments, conceived and designed MinION experiments, helped perform MinION experiments and bioinformatics, and co-wrote the paper. MJ helped conceive and design MinION experiments, helped perform MinION experiments and bioinformatics, and co-wrote the paper. LM helped design RmtB experiments and engineered an *E. coli* strain that carried RmtB. DRG co-wrote the manuscript and helped conceive and design MinION experiments. MA co-wrote the manuscript, helped conceive and design experiments, and oversaw the project.

Synopsis

In this work, we developed and implemented direct RNA sequencing to 16S ribosomal RNA. We developed an enrichment-based approach for selective sequencing of rRNA in a human total RNA background. We also demonstrated simultaneous detection of two nucleotide modifications on 16S rRNA from *E. coli.* We also performed microbial classification using 16S rRNA data from four different microorganisms.

I helped with design and execution of MinION sequencing experiments for data generation and processing. Thereafter, I performed signal-level data analysis using marginAlign [9] and nanoraw [141]. I helped with estimating error model in the data using marginAlign EM [9], and computing alignment statistics. I also contributed towards writing the manuscript.

MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry

Miten Jain[1,*], John R Tyson[2,*], Matthew Loose[3,*], Camilla LC Ip[4,5,*], David A Eccles[6], Justin O'Grady[7], Sunir Malla[3], Richard M Leggett[8], Ola Wallerman[9], Hans J Jansen[10], Vadim Zalunin[11], Ewan Birney[11,*], Bonnie L Brown[12,*], Terrance P Snutch[2,*], Hugh E Olsen[1,*], MinION Analysis and Reference Consortium

[1]University of California at Santa Cruz, Santa Cruz, CA, USA. [2]Michael Smith Laboratories and Djavad Mowfaghian Centre for Brain Health, University of British Columbia, Vancouver, Canada. [3]School of Life Sciences, University of Nottingham, Nottingham, UK. [4]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [5]Peter Medawar Building for Pathogen Research, University of Oxford, Oxford, UK. [6]Malaghan Institute of Medical Research, Wellington, New Zealand. [7]Norwich Medical School, University of East Anglia, Norwich, UK. [8]Earlham Institute, Norwich Research Park, Norwich, UK. [9]Science for Life Laboratory, IGP, Uppsala University, Uppsala, Sweden. [10]ZF-screens B.V., Leiden, Netherlands. [11]European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. [12]Virginia Commonwealth University, Richmond, VA, USA. [*]Equal contributors. Correspondence should be addressed to Miten Jain (miten@soe.ucsc.edu), John R Tyson (jtyson@msl.ubc.ca), Matthew Loose (loose@nottingham.ac.uk), Camilla LC Ip (camilla.ip@well.ox.ac.uk), Ewan Birney (birney@ebi.ac.uk), Bonnie L Brown (blbrown@vcu.edu), Terrance P Snutch (snutch@msl.ubc.ca), or Hugh E Olsen (he-

olsen@soe.ucsc.edu).

Abstract

Background: long read sequencing is rapidly evolving and reshaping the suite of opportunities for genomic analysis. For the MinION in particular, as both the platform and chemistry develop, the user community requires reference data to set performance expectations and maximally exploit third-generation sequencing. We performed an analysis of MinION data derived from whole genome sequencing of *Escherichia coli* K-12 using the R9.0 chemistry, comparing the results with the older R7.3 chemistry.

Methods: We computed the error-rate estimates for insertions, deletions, and mismatches in MinION reads.

Results: Run-time characteristics of the flow cell and run scripts for R9.0 were similar to those observed for R7.3 chemistry, but with an 8-fold increase in bases per second (from 30 bps in R7.3 and SQK-MAP005 library preparation, to 250 bps in R9.0) processed by individual nanopores, and less drop-off in yield over time. The 2-dimensional ("2D") N50 read length was unchanged from the prior chemistry. Using the proportion of alignable reads as a measure of base-call accuracy, 99.9% of "pass" template reads from 1-dimensional ("1D") experiments were mappable and 97% from 2D experiments. The median identity of reads was 89% for 1D and 94% for 2D experiments. The total error rate (miscall + insertion + deletion) decreased for 2D "pass" reads from 9.1% in R7.3 to 7.5% in R9.0 and for template "pass" reads from 26.7% in R7.3 to 14.5% in R9.0.

Conclusions: These Phase 2 MinION experiments serve as a baseline by providing estimates for read quality, throughput, and mappability. The datasets further enable the development of bioinformatic tools tailored to the new R9.0 chemistry and the design of novel biological applications for this technology.

Abbreviations: K: thousand, Kb: kilobase (one thousand base pairs), M: million, Mb: megabase (one million base pairs), Gb: gigabase (one billion base pairs).

Contributions

MJ and JT coordinated the study. The MARC group collectively designed the study. ML, SM, and JT performed the experiments. VZ, RL, ML, MJ, RL, and CI ran data pre-processing steps. MJ, CI, and JT analysed the data. MJ and BB drafted the manuscript. All authors participated in discussions relating to the generation and analysis of the data and edited and approved the final manuscript for submission.

Synopsis

In this work, we performed biological and technical replicate experiments using the MinION in two different laboratories. We then performed marginAlign [9] analyses (alignments, EM, statistics) on these data. The data were achieved using R9.0 chemistry-based experiments and were termed as Phase 2 release from MARC.

I helped with running the consortium, and performing data analyses. I used marginAlign [9] (with and without EM) for estimating the error-model in the data, and to compute the alignment statistics. I also helped with writing the manuscript.

Conclusion

The work I present demonstrates the utility and recent advances in nanopore sequencing. One of the most striking improvements has been in read length. DNA read lengths of 10 kb became routine with the release of the MinION in 2014. Now, 200 kb+ read lengths are routinely achievable, with the longest measured at 882 kb. This advance in read lengths allowed us to assemble the first human centromere. It is likely these unprecedented read lengths will reveal new insights into the complex, unresolved regions of the human genome.

The MinION DNA sequencing accuracy went from 66% in June 2014 to ~95% in early-2017. This was achieved by ONT with successive improvements in chemistry, from R6 in June 2014 to R9.5 in March 2017 (with R7, R7.3, R9.0, and R9.4 chemistries in between). The main changes in chemistry was with the use of a new nanopore, CsgG, in R9.0 chemistry and ones following it. This new pore, combined with software improvements substantially improved both accuracy and throughput. The new 1d2 (1d-squared) chemistry now achieves >95% median sequencing accuracy, and yields of >15 Gb per MinION flow cell are routine.

The improvements with R9.4 chemistry made it feasible for a consortium to sequence a human genome using the MinION. This effort required 39 flow cells to achieve ~30X coverage of the genome. If redone with the subsequent improvements, this sequencing could be done today with ~10 flow cells.

Nanopore direct RNA sequencing became available in 2016, and has opened a new frontier for single-molecule analysis of native RNA. Full-length reads of native RNA molecules permit further annotation of various RNAs, along with direct detection of nucleotide modifications simultaneously.

ONT democratized sequencing with MAP in 2014. This allowed individual laboratories to perform their own sequencing as well as analyses. From this spawned a host of software tools for analysis of nanopore data. Some of these tools also are able to combine information from other sequencing platforms, such as Illumina short-reads, with nanopore data. These tools are becoming a standard practice in the community for nanopore analysis now. Some examples of these tools include marginAlign and GraphMap for sequence alignment, Canu for genome assembly, and marginCaller and nanopolish detection of single nucleotide variants.

It is foreseeable that MinION's will become a common laboratory equipment, akin to a PCR machine or a gel electrophoresis equipment. Nanopore sequencing can play an essential role in resolving complex genomic regions, analyzing nucleotide modifications in native DNA and RNA, and understanding DNA and RNA structure.

Appendix

Table 3.1: Manuscript titles and weblinks

Title	**Weblink**
The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community	`https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1103-0`
Improved data analysis for the MinION nanopore sequencer	`http://www.nature.com/nmeth/journal/v12/n4/abs/nmeth.3290.html`
Linear Assembly of a Human Y Centromere using Nanopore Long-Reads	
Mapping DNA methylation with high-throughput nanopore sequencing	`http://www.nature.com/nmeth/journal/v14/n4/abs/nmeth.4189.html`
Nanopore sequencing and assembly of a human genome with ultra-long reads	`http://biorxiv.org/content/early/2017/04/20/128835`

Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing	`http://biorxiv.org/content/early/2017/04/29/132274`
Whole genome sequencing and assembly of a *Caenorhabditis elegans* genome with complex genomic rearrangements using the MinION sequencing device	`http://biorxiv.org/content/early/2017/01/08/099143`
MinION Analysis and Reference Consortium: Phase 1 data release and analysis	`https://f1000research.com/articles/4-1075/v1`
MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry	`https://f1000research.com/articles/6-760/v1`

Table 3.2: Software pipelines and weblinks

Pipeline	**Link**
marginAlign	`https://github.com/benedictpaten/marginAlign`
nanopore	`https://github.com/mitenjain/nanopore`

Table 3.3: Data repositories and weblinks

Data	Link
M13mp18	PRJEB8230 - `http://www.ebi.ac.uk/ena/data/view/PRJEB8230` ERP009289 - `http://www.ebi.ac.uk/ena/data/view/ERP009289`
NA12878	`https://github.com/nanopore-wgs-consortium/NA12878`

Bibliography

[1] Daniel Branton, David W Deamer, Marziali Andre, Bayley Hagan, Steven A Benner, Butler Thomas, Massimiliano Di Ventra, Garaj Slaven, Hibbs Andrew, Huang Xiaohua, Stevan B Jovanovich, Predrag S Krstic, Lindsay Stuart, Xinsheng Sean Ling, Carlos H Mastrangelo, Meller Amit, John S Oliver, Yuriy V Pershin, J Michael Ramsey, Riehn Robert, Gautam V Soni, Tabard-Cossa Vincent, Wanunu Meni, Wiggin Matthew, and Jeffery A Schloss. The potential and challenges of nanopore sequencing. *Nat. Biotechnol.*, 26(10):1146–1153, 2008. 2

[2] David Deamer, Mark Akeson, and Daniel Branton. Three decades of nanopore sequencing. *Nat. Biotechnol.*, 34(5):518–524, 6 May 2016. 2, 3, 23

[3] J J Kasianowicz, E Brandin, D Branton, and D W Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.*, 93(24):13770–13773, 26 November 1996. 2

[4] Gerald M Cherf, Kate R Lieberman, Rashid Hytham, Christopher E Lam, Karplus

Kevin, and Akeson Mark. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nat. Biotechnol.*, 30(4):344–348, 2012. 2

[5] Mariam Ayub and Hagan Bayley. Individual RNA base recognition in immobilized oligonucleotides using a protein nanopore. *Nano Lett.*, 12(11):5637–5643, 14 November 2012. 2

[6] Elizabeth A Manrao, Ian M Derrington, Andrew H Laszlo, Kyle W Langford, Matthew K Hopper, Gillgren Nathaniel, Pavlenok Mikhail, Niederweis Michael, and Jens H Gundlach. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat. Biotechnol.*, 30(4):349–353, 2012. 2

[7] Philip M Ashton, Satheesh Nair, Tim Dallman, Salvatore Rubino, Wolfgang Rabsch, Solomon Mwaigwisya, John Wain, and Justin O'Grady. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.*, 33(3):296–300, March 2015. 2, 20

[8] Mohan T Bolisetty, Gopinath Rajadinakaran, and Brenton R Graveley. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biol.*, 16:204, 30 September 2015. 2, 11

[9] Miten Jain, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson. Improved data analysis for the MinION nanopore sequencer. *Nat.*

Methods, 12(4):351–356, April 2015. 2, 9, 10, 12, 13, 15, 16, 18, 107, 109, 118, 130, 131, 134, 136, 139, 143, 146

[10] J Quick, A Quinlan, and N Loman. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. Technical report, 26 September 2014. 2

[11] Andy Kilianski, Jamie L Haas, Elizabeth J Corriveau, Alvin T Liem, Kristen L Willis, Dana R Kadavy, C Nicole Rosenzweig, and Samuel S Minot. Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *Gigascience*, 4:12, 26 March 2015. 2, 20

[12] Ivan Sović, Mile Šikić, Andreas Wilm, Shannon Nicole Fenlon, Swaine Chen, and Niranjan Nagarajan. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat. Commun.*, 7:11307, 15 April 2016. 2, 13, 17, 18

[13] Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael Schatz, and W Richard McCombie. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Technical report, 6 January 2015. 2, 13

[14] Alexander L Greninger, Samia N Naccache, Scot Federman, Guixia Yu, Placide Mbala, Vanessa Bres, Doug Stryke, Jerome Bouquet, Sneha Somasekar, Jeffrey M Linnen, Roger Dodd, Prime Mulembakani, Bradley S Schneider, Jean-Jacques Muyembe-Tamfum, Susan L Stramer, and Charles Y Chiu. Rapid metagenomic

identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med.*, 7(1):99, 29 September 2015. 2, 6, 20

[15] Adam D Hargreaves and John F Mulley. Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. *PeerJ*, 3:e1441, 24 November 2015. 2

[16] Minh Duc Cao, Devika Ganesamoorthy, Alysha G Elliott, Huihui Zhang, Matthew A Cooper, and Lachlan J M Coin. Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinION$^{\text{TM}}$ sequencing. *Gigascience*, 5(1):32, 26 July 2016. 2

[17] Kim Judge, Simon R Harris, Sandra Reuter, Julian Parkhill, and Sharon J Peacock. Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. *J. Antimicrob. Chemother.*, 70(10):2775–2778, October 2015. 2, 21

[18] E Karlsson, A Lärkeryd, A Sjödin, M Forsman, and P Stenberg. Scaffolding of a bacterial genome using MinION nanopore sequencing. *Sci. Rep.*, 5:11996, 7 July 2015. 2

[19] Mehdi Kchouk, Kchouk Mehdi, and Elloumi Mourad. Error correction and De-Novo genome assembly for the MinION sequencing reads mixing illumina short reads. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2015. 2

[20] Richard M Leggett, Darren Heavens, Mario Caccamo, Matthew D Clark, and Robert P Davey. NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. *Bioinformatics*, 32(1):142–144, 1 January 2016. 2, 13

[21] Nicholas J Loman and Mark J Pallen. Twenty years of bacterial genome sequencing. *Nat. Rev. Microbiol.*, 13(12):787–794, December 2015. 2

[22] Nicholas J Loman and Aaron R Quinlan. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*, 30(23):3399–3401, 1 December 2014. 2, 12

[23] Mohammed-Amin Madoui, Stefan Engelen, Corinne Cruaud, Caroline Belser, Laurie Bertrand, Adriana Alberti, Arnaud Lemainque, Patrick Wincker, and Jean-Marc Aury. Genome assembly using nanopore-guided long and error-free DNA reads. *BMC Genomics*, 16:327, 20 April 2015. 2, 13

[24] Alexander S Mikheyev and Mandy M Y Tin. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.*, 14(6):1097–1102, November 2014. 2, 12

[25] George Miles, Jessica Hoisington-Lopez, and Eric Duncavage. Nanopore sequencing of a DNA library prepared from Formalin-Fixed Paraffin-Embedded tissue. In *Laboratory Investigation*, volume 95, pages 520A–521A, 2015. 2

[26] Ruth R Miller, Vincent Montoya, Jennifer L Gardy, David M Patrick, and Patrick

Tang. Metagenomics for pathogen detection in public health. *Genome Med.*, 5(9):81, 20 September 2013. 2

[27] M J Pallen. Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. *Parasitology*, 141(14):1856–1862, December 2014. 2, 20

[28] Joshua Quick, Philip Ashton, Szymon Calus, Carole Chatt, Savita Gossain, Jeremy Hawker, Satheesh Nair, Keith Neal, Kathy Nye, Tansy Peters, Elizabeth De Pinna, Esther Robinson, Keith Struthers, Mark Webber, Andrew Catto, Timothy J Dallman, Peter Hawkey, and Nicholas J Loman. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of salmonella. *Genome Biol.*, 16:114, 30 May 2015. 2, 20

[29] J Quick and N J Loman. Bacterial whole-genome read data from the Oxford Nanopore Technologies MinION^{TM} nanopore sequencer. *GigaScience Database*, 2014. 2

[30] Joshua Quick, Nicholas J Loman, Sophie Duraffour, Jared T Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, Nobila Ouédraogo, Babak Afrough, Amadou Bah, Jonathan H J Baum, Beate Becker-Ziaja, Jan Peter Boettcher, Mar Cabeza-Cabrerizo, Álvaro Camino-Sánchez, Lisa L Carter, Juliane Doerrbecker, Theresa Enkirch, Isabel García-Dorival, Nicole Hetzelt, Julia Hinzmann, Tobias Holm, Liana Eleni Kafetzopoulou, Michel Koropogui, Abigael Kosgey, Eeva Kuisma, Christopher H Logue, Antonio Mazzarelli, Sarah Meisel, Marc Mertens, Janine Michel, Didier Ngabo,

Katja Nitzsche, Elisa Pallasch, Livia Victoria Patrono, Jasmine Portmann, Johanna Gabriella Repits, Natasha Y Rickett, Andreas Sachse, Katrin Singethan, Inês Vitoriano, Rahel L Yemanaberhan, Elsa G Zekeng, Trina Racine, Alexander Bello, Amadou Alpha Sall, Ousmane Faye, Oumar Faye, N'faly Magassouba, Cecelia V Williams, Victoria Amburgey, Linda Winona, Emily Davis, Jon Gerlach, Frank Washington, Vanessa Monteil, Marine Jourdain, Marion Bererd, Alimou Camara, Hermann Somlare, Abdoulaye Camara, Marianne Gerard, Guillaume Bado, Bernard Baillet, Déborah Delaune, Koumpingnin Yacouba Nebie, Abdoulaye Diarra, Yacouba Savane, Raymond Bernard Pallawo, Giovanna Jaramillo Gutierrez, Natacha Milhano, Isabelle Roger, Christopher J Williams, Facinet Yattara, Kuiama Lewandowski, James Taylor, Phillip Rachwal, Daniel J Turner, Georgios Pollakis, Julian A Hiscox, David A Matthews, Matthew K O'Shea, Andrew Mcd Johnston, Duncan Wilson, Emma Hutley, Erasmus Smit, Antonino Di Caro, Roman Wölfel, Kilian Stoecker, Erna Fleischmann, Martin Gabriel, Simon A Weller, Lamine Koivogui, Boubacar Diallo, Sakoba Keïta, Andrew Rambaut, Pierre Formenty, Stephan Günther, and Miles W Carroll. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–232, 11 February 2016. 2, 15, 19, 20, 21

[31] Joshua Quick, Aaron R Quinlan, and Nicholas J Loman. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *Gigascience*, 3(1):1–6, 2014. 2

[32] Anna C Ramgren, Hannah S Newhall, and Karen E James. DNA barcoding and metabarcoding with the Oxford Nanopore MinION. In *Genome*, volume 58, pages 268–268, 2015. 2

[33] Judith Risse, Marian Thomson, Sheila Patrick, Garry Blakely, Georgios Koutsovoulos, Mark Blaxter, and Mick Watson. A single chromosome assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore sequencing data. *Gigascience*, 4:60, 4 December 2015. 2

[34] Jing Wang, Nicole E Moore, Yi-Mo Deng, David A Eccles, and Richard J Hall. MinION nanopore sequencing of an influenza genome. *Front. Microbiol.*, 6:766, 18 August 2015. 2, 20

[35] Jeremy R Wang and Corbin D Jones. Fast alignment filtering of nanopore sequencing reads using locality-sensitive hashing. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2015. 2

[36] Alan C Ward and Wonyong Kim. MinION: New, long read, portable nucleic acid sequencing device. *J. Bacteriol. Virol.*, 45(4):285, 2015. 2

[37] Mick Watson, Marian Thomson, Judith Risse, Richard Talbot, Javier Santoyo-Lopez, Karim Gharbi, and Mark Blaxter. poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics*, 31(1):114–115, 1 January 2015. 2, 12

[38] Shan Wei and Zev Williams. Rapid Short-Read sequencing and aneuploidy detec-

tion using MinION nanopore technology. *Genetics*, 202(1):37–44, January 2016. 2, 22

[39] Jacob Schreiber, Zachary L Wescoe, Robin Abu-Shumays, John T Vivian, Baldandorj Baatar, Kevin Karplus, and Mark Akeson. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.*, 110(47):18910–18915, 19 November 2013. 5, 24

[40] Zachary L Wescoe, Jacob Schreiber, and Mark Akeson. Nanopores discriminate among five C5-cytosine variants in DNA. *J. Am. Chem. Soc.*, 136(47):16582–16587, 26 November 2014. 5, 6, 24

[41] Arthur C Rand, Miten Jain, Jordan Eizenga, Audrey Musselman-Brown, Hugh E Olsen, Mark Akeson, and Benedict Paten. Cytosine variant calling with high-throughput nanopore sequencing. 1 January 2016. 6, 24

[42] Jared T Simpson, Rachael Workman, Philip C Zuzarte, Matei David, Lewis Jonathan Dursi, and Winston Timp. Detecting DNA methylation using the Oxford Nanopore Technologies MinION sequencer. Technical report, 4 April 2016. 6, 24

[43] Matthew Loose, Sunir Malla, and Michael Stout. Real-time selective sequencing using nanopore technology. *Nat. Methods*, 13(9):751–754, September 2016. 6, 8, 14

[44] Camilla L C Ip, Matthew Loose, John R Tyson, Mariateresa de Cesare, Bonnie L Brown, Miten Jain, Richard M Leggett, David A Eccles, Vadim Zalunin, John M Urban, Paolo Piazza, Rory J Bowden, Benedict Paten, Solomon Mwaigwisya, Elizabeth M Batty, Jared T Simpson, Terrance P Snutch, Ewan Birney, David Buck, Sara Goodwin, Hans J Jansen, Justin O'Grady, Hugh E Olsen, and MinION Analysis and Reference Consortium. MinION analysis and reference consortium: Phase 1 data release and analysis. *F1000Res.*, 4:1075, 15 October 2015. 9, 12, 15, 23

[45] Yao-Tseng Chen, Christian Iseli, Charis A Venditti, Lloyd J Old, Andrew J G Simpson, and C Victor Jongeneel. Identification of a new cancer/testis gene family, CT47, among expressed multicopy genes on the human X chromosome. *Genes Chromosomes Cancer*, 45(4):392–400, April 2006. 10

[46] Alexis L Norris, Rachael E Workman, Yunfan Fan, James R Eshleman, and Winston Timp. Nanopore sequencing detects structural variants in cancer. *Cancer Biol. Ther.*, 17(3):246–253, 3 March 2016. 11

[47] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 16 March 2013. 12, 17

[48] Martin C Frith, Michiaki Hamada, and Paul Horton. Parameters for accurate genome alignment. *BMC Bioinformatics*, 11:80, 9 February 2010. 12, 17

[49] Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial

genome assembled de novo using only nanopore sequencing data. *Nat. Methods*, 12(8):733–735, August 2015. 13, 17, 18, 24

[50] Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.*, 33(6):623–630, June 2015. 13

[51] Tamas Szalay and Jene A Golovchenko. De novo sequencing and variant calling with nanopores using PoreSeq. *Nat. Biotechnol.*, 33(10):1087–1091, October 2015. 13, 19

[52] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, 17(1):132, 20 June 2016. 14

[53] minoTour. minoTour/minoTour. `https://github.com/minoTour/minoTour`. Accessed: 2016-6-26. 14

[54] Matei David, Lewis Jonathan Dursi, Delia Yao, Paul C Boutros, and Jared T Simpson. Nanocall: An open source basecaller for Oxford Nanopore sequencing data. 1 January 2016. 14

[55] Vladimír Boža, Broňa Brejová, and Tomáš Vinař. DeepNano: Deep recurrent

neural networks for base calling in MinION nanopore reads. 30 March 2016. 14, 15

[56] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex Dewinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2 January 2009. 17, 20

[57] Christopher Lee, Catherine Grasso, and Mark F Sharlow. Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464, March 2002. 17

[58] Chenhao Li, Kern Rei Chng, Jia Hui Esther Boey, Hui Qi Amanda Ng, Andreas Wilm, and Niranjan Nagarajan. INC-Seq: Accurate single molecule reads using nanopore sequencing. *GigaScience*, 4:60, 2 August 2016. 20

[59] Thomas Hoenen, Allison Groseth, Kyle Rosenke, Robert J Fischer, Andreas Hoenen, Seth D Judson, Cynthia Martellaro, Darryl Falzarano, Andrea Marzi,

R Burke Squires, Kurt R Wollenberg, Emmie de Wit, Joseph Prescott, David Safronetz, Neeltje van Doremalen, Trenton Bushmaker, Friederike Feldmann, Kristin McNally, Fatorma K Bolay, Barry Fields, Tara Sealy, Mark Rayfield, Stuart T Nichol, Kathryn C Zoon, Moses Massaquoi, Vincent J Munster, and Heinz Feldmann. Nanopore sequencing as a rapidly deployable Ebola outbreak tool. *Emerg. Infect. Dis.*, 22(2):331–334, February 2016. 21

[60] Ebola situation report - 11 november 2015 — ebola. `http://apps.who.int/ebola/current-situation/ebola-situation-report-11-november-2015`. Accessed: 2016-6-21. 21

[61] Sophie Zaaijer, Columbia University Ubiquitous Genomics 2015 class, and Yaniv Erlich. Using mobile sequencers in an academic classroom. *Elife*, 5, 7 April 2016. 21

[62] Citizen sequencers: Taking oxford nanopore's MinION to the classroom and beyond - Bio-IT world. `http://www.bio-itworld.com/2015/12/9/citizen-sequencers-taking-oxford-nanopores-minion-classroom-beyond.html`. Accessed: 2016-6-29. 22

[63] Shengpei Chen, Sheng Li, Weiwei Xie, Xuchao Li, Chunlei Zhang, Haojun Jiang, Jing Zheng, Xiaoyu Pan, Hancheng Zheng, Jia Sophie Liu, Yongqiang Deng, Fang Chen, and Hui Jiang. Performance comparison between rapid sequencing platforms for ultra-low coverage sequencing strategy. *PLoS One*, 9(3):e92192, 20 March 2014. 22

[64] Now they're sequencing DNA in outer space. *MIT Technology Review*, 10 June 2016. Accessed: 2017-5-22. 22

[65] Sequencing DNA in the palm of your hand. 29 September 2015. Accessed: 2017-5-22. 22

[66] Sarah L Castro-Wallace, Charles Y Chiu, Kristen K John, Sarah E Stahl, Kathleen H Rubins, Alexa B R McIntyre, Jason P Dworkin, Mark L Lupisella, David J Smith, Douglas J Botkin, Timothy A Stephenson, Sissel Juul, Daniel J Turner, Fernando Izquierdo, Scot Federman, Doug Stryke, Sneha Somasekar, Noah Alexander, Guixia Yu, Christopher Mason, and Aaron S Burton. Nanopore DNA sequencing and genome assembly on the international space station. 1 January 2016. 22

[67] Andrew M Smith, Robin Abu-Shumays, Mark Akeson, and David L Bernick. Capture, unfolding, and detection of individual tRNA molecules using a nanopore device. *Front Bioeng Biotechnol*, 3:91, 24 June 2015. 24

[68] Robert Y Henley, Brian Alan Ashcroft, Ian Farrell, Barry S Cooperman, Stuart M Lindsay, and Meni Wanunu. Electrophoretic deformation of individual transfer RNA molecules reveals their identity. *Nano Lett.*, 16(1):138–144, 13 January 2016. 24

[69] Progress at UC Santa Cruz: Long DNA fragments, tRNA and modified bases — Mark Akeson, UC Santa Cruz. 24

[70] R P Horgon and L C Kenny. SAC review: Omic technologies: genomics, transcriptomics. *Proteomics and metabolomics. TOG*, 13:189–195, 2011. 25

[71] Jeff Nivala, Douglas B Marks, and Mark Akeson. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat. Biotechnol.*, 31(3):247–250, March 2013. 25

[72] Leroy E Hood, Gilbert S Omenn, Robert L Moritz, Ruedi Aebersold, Keith R Yamamoto, Michael Amos, Jennie Hunter-Cevera, Laurie Locascio, and Workshop Participants. New and improved proteomics technologies for understanding complex biological systems: addressing a grand challenge in the life sciences. *Proteomics*, 12(18):2773–2783, September 2012. 25

[73] MJ Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC bioinformatics*, 13:238, 2012. 34, 70

[74] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 00(00):3, March 2013. 34, 70

[75] Martin C Frith, Raymond Wan, and Paul Horton. Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic acids research*, 38(7):e100, April 2010. 34, 35, 70

[76] Robert S Harris. *Improved pairwise alignment of genomic DNA*. PhD thesis, The Pennsylvania State University, 2007. 34, 70

[77] Joshua Quick, AR Quinlan, and NJ Loman. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *GigaScience*, pages 1–6, 2014. 35, 52, 80

[78] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. GenBank. *Nucleic acids research*, 41(Database issue):D36–42, January 2013. 35, 71

[79] Stephen F Altschup, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic Local Alignment Search Tool. *J. Mol. Biol.*, 215:403–410, 1990. 35, 71

[80] Chuong B Do and Serafim Batzoglou. What is the expectation maximization algorithm? *Nature biotechnology*, 26(8):897–9, August 2008. 42

[81] Benedict Paten, Javier Herrero, Kathryn Beal, Stephen Fitzgerald, and Ewan Birney. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome research*, 18(11):1814–28, November 2008. 42, 72, 76

[82] Ariel S Schwartz and Lior Pachter. Multiple alignment by sequence annealing. *Bioinformatics (Oxford, England)*, 23(2):e24–9, January 2007. 49, 73

[83] Philip M Ashton, Satheesh Nair, Tim Dallman, Salvatore Rubino, Wolfgang Rabsch, Solomon Mwaigwisya, John Wain, and Justin O'Grady. MinION nanopore

sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology*, (December), December 2014. 53, 67

[84] John W Davey, Paul a Hohenlohe, Paul D Etter, Jason Q Boone, Julian M Catchen, and Mark L Blaxter. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature reviews. Genetics*, 12(7):499–510, July 2011. 57

[85] Sarah J Bourlat, Angel Borja, Jack Gilbert, Martin I Taylor, Neil Davies, Stephen B Weisberg, John F Griffith, Teresa Lettieri, Dawn Field, John Benzie, Frank Oliver Glöckner, Naiara Rodríguez-Ezpeleta, Daniel P Faith, Tim P Bean, and Matthias Obst. Genomics in marine monitoring: new opportunities for assessing marine health status. *Marine pollution bulletin*, 74(1):19–31, September 2013. 57

[86] David Stucki and Sebastien Gagneux. Single nucleotide polymorphisms in Mycobacterium tuberculosis and the need for a curated database. *Tuberculosis (Edinburgh, Scotland)*, 93(1):30–9, January 2013. 57

[87] I. Holmes and W. J. Bruno. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, 17(9):803–820, September 2001. 58, 74, 75, 83

[88] YT Chen, Christian Iseli, and CA Venditti. Identification of a new cancer/testis gene family, CT47, among expressed multicopy genes on the human X chromosome. *Genes, Chromosomes & Cancer*, 400(December 2005):392–400, 2006. 61

[89] Todd J Treangen and Steven L Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics*, 13(1):36–46, January 2012. 61

[90] Deanna C Tremblay, Graham Alexander, Shawn Moseley, and Brian P Chadwick. Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. *BMC genomics*, 11(1):632, January 2010. 61

[91] Manisha Brahmachary, Audrey Guilmatre, Javier Quilez, Dan Hasson, Christelle Borel, Peter Warburton, and Andrew J Sharp. Digital genotyping of macrosatellites and multicopy genes reveals novel biological functions associated with copy number variation of large tandem repeats. *PLoS genetics*, 10(6):e1004418, June 2014. 61

[92] Alexander S. Mikheyev and Mandy M.Y. Tin. A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources*, 14(6):1097–1102, September 2014. 67

[93] Jacob Schreiber, Zachary L Wescoe, Robin Abu-Shumays, John T Vivian, Baldandorj Baatar, Kevin Karplus, and Mark Akeson. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, 110(47):18910–5, November 2013. 67

[94] Andrew H Laszlo, Ian M Derrington, Henry Brinkerhoff, Kyle W Langford, Ian C

Nova, Jenny Mae Samson, Joshua J Bartlett, Mikhail Pavlenok, and Jens H Gundlach. Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proceedings of the National Academy of Sciences of the United States of America*, 110(47):18904–9, November 2013. 67

[95] Zachary L Wescoe, Jacob Schreiber, and Mark Akeson. Nanopores discriminate among five C5-Cytosine variants in DNA. *Journal of the American Chemical Society*, 136(47):16582–7, November 2014. 67

[96] Gerald M Cherf, Kate R Lieberman, Hytham Rashid, Christopher E Lam, Kevin Karplus, and Mark Akeson. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Åprecision. *Nature biotechnology*, 30(4):344–8, April 2012. 67

[97] Kate R Lieberman, Joseph M Dahl, Ai H Mai, Mark Akeson, and Hongyun Wang. Dynamics of the translocation step measured in individual DNA polymerase complexes. *Journal of the American Chemical Society*, 134(45):18816–23, November 2012. 67

[98] AEP Schibel, Na An, Qian Jin, Fleming AM, Burrows CJ, and White HS. Nanopore detection of 8-oxo-7,8-dihydro-2'-deoxyguanosine in immobilized single-stranded DNA via adduct formation to the DNA damage site. *Journal of American Chemical Society*, 132(51):17992–17995, 2010. 67

[99] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 70, 131, 134

[100] Martin C Frith, Michiaki Hamada, and Paul Horton. Parameters for accurate genome alignment. *BMC bioinformatics*, 11:80, January 2010. 70

[101] R Durbin, S R Eddy, A Krogh, and G Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* The Press Syndicate of The University of Cambridge, 1998. 71, 72

[102] Benedict Paten, Dent Earl, Ngan Nguyen, Mark Diekhans, Daniel Zerbino, and David Haussler. Cactus: Algorithms for genome multiple sequence alignment. *Genome research*, 21(9):1512–28, September 2011. 72, 73

[103] SR Eddy. Profile hidden Markov models. *Bioinformatics*, pages 755–763, 1998. 75

[104] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–2, March 2010. 76

[105] FM You, N Huo, KR Deal, and YQ Gu. 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. 81

[106] Isaac Elias. Settling the intractability of multiple alignment. *Journal of Computational Biology*, 13(7):1323–1339, September 2006. 85

[107] Oscar Westesson, Gerton Lunter, Benedict Paten, and Ian Holmes. Phylogenetic automata, pruning, and multiple alignment. March 2011. 85

[108] Jorge J Yunis and Walid G Yasmineh. Heterochromatin, satellite DNA, and cell function. *Science*, 174(4015):1200–1209, 17 December 1971. 106

[109] M G Schueler, A W Higgins, M K Rudd, K Gustashaw, and H F Willard. Genomic and genetic definition of a functional human centromere. *Science*, 294(5540):109–115, 5 October 2001. 106

[110] Daniel E Khost, Danna G Eickbush, and Amanda M Larracuente. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in drosophila melanogaster. *Genome Res.*, 27(5):709–721, May 2017. 106

[111] Miten Jain, Sergey Koren, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Karen H Miga, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas James Loman, and Matthew Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. 20 April 2017. 107, 109, 110, 118

[112] C A Tilford, T Kuroda-Kawaguchi, H Skaletsky, S Rozen, L G Brown, M Rosenberg, J D McPherson, K Wylie, M Sekhon, T A Kucaba, R H Waterston, and D C Page. A physical map of the human Y chromosome. *Nature*, 409(6822):943–945, 15 February 2001. 107, 118

[113] D Brutlag, M Carlson, K Fry, and T S Hsieh. DNA sequence organization in

drosophila heterochromatin. *Cold Spring Harb. Symp. Quant. Biol.*, 42 Pt 2:1137–1146, 1978. 107

[114] Roger A Hoskins, Christopher D Smith, Joseph W Carlson, A Bernardo Carvalho, Aaron Halpern, Joshua S Kaminker, Cameron Kennedy, Chris J Mungall, Beth A Sullivan, Granger G Sutton, Jiro C Yasuhara, Barbara T Wakimoto, Eugene W Myers, Susan E Celniker, Gerald M Rubin, and Gary H Karpen. Heterochromatic sequences in a drosophila whole-genome shotgun assembly. *Genome Biol.*, 3(12):RESEARCH0085, 31 December 2002. 107

[115] D L Neil, A Villasante, R B Fisher, D Vetrie, B Cox, and C Tyler-Smith. Structural instability of human tandemly repeated DNA sequences cloned in yeast artificial chromosome vectors. *Nucleic Acids Res.*, 18(6):1421–1428, 25 March 1990. 107

[116] Karen E Hayden, Erin D Strome, Stephanie L Merrett, Hye-Ran Lee, M Katharine Rudd, and Huntington F Willard. Sequences associated with centromere competency in the human genome. *Mol. Cell. Biol.*, 33(4):763–772, February 2013. 107, 112

[117] Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, and Ashlee M Earl. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9(11):e112963, 19 November 2014. 111, 123

[118] J Wolfe, S M Darling, R P Erickson, I W Craig, V J Buckle, P W Rigby, H F Willard, and P N Goodfellow. Isolation and characterization of an alphoid centromeric repeat family from the human Y chromosome. *J. Mol. Biol.*, 182(4):477–485, 20 April 1985. 112, 113

[119] C Tyler-Smith and W R Brown. Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J. Mol. Biol.*, 195(3):457–470, 5 June 1987. 112, 113, 124

[120] C Tyler-Smith, R J Oakey, Z Larin, R B Fisher, M Crocker, N A Affara, M A Ferguson-Smith, M Muenke, O Zuffardi, and M A Jobling. Localization of DNA sequences required for human centromere function through an analysis of rearranged Y chromosomes. *Nat. Genet.*, 5(4):368–375, December 1993. 112

[121] 1000 Genomes Project Consortium, Goncalo R Abecasis, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil A McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 1 November 2012. 112, 124

[122] Karen H Miga, Yulia Newton, Miten Jain, Nicolas Altemose, Huntington F Willard, and W James Kent. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.*, 24(4):697–707, April 2014. 112, 124

[123] Rachel Wevrick and Huntington F Willard. Long-range organization of tandem

arrays of alpha satellite DNA at the centromeres of human chromosomes: high-frequency array-length polymorphism and meiotic stability. *Proceedings of the National Academy of Sciences*, 86(23):9394–9398, 1989. 112, 124

[124] R Oakey and C Tyler-Smith. Y chromosome DNA haplotyping suggests that most european and asian men are descended from one of two males. *Genomics*, 7(3):325–330, July 1990. 112, 113, 124

[125] Samuel Levy, Granger Sutton, Pauline C Ng, Lars Feuk, Aaron L Halpern, Brian P Walenz, Nelson Axelrod, Jiaqi Huang, Ewen F Kirkness, Gennady Denisov, Yuan Lin, Jeffrey R MacDonald, Andy Wing Chun Pang, Mary Shago, Timothy B Stockwell, Alexia Tsiamouri, Vineet Bafna, Vikas Bansal, Saul A Kravitz, Dana A Busam, Karen Y Beeson, Tina C McIntosh, Karin A Remington, Josep F Abril, John Gill, Jon Borman, Yu-Hui Rogers, Marvin E Frazier, Stephen W Scherer, Robert L Strausberg, and J Craig Venter. The diploid genome sequence of an individual human. *PLoS Biol.*, 5(10):e254, 4 September 2007. 112, 124

[126] Mark J Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, 13:238, 19 September 2012. 122

[127] Timo Lassmann and Erik L L Sonnhammer. Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, 6(1):298, 2005. 122, 124

[128] Alistair Miles. pysamstats. 122

[129] S R Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998. 123

[130] Helen Skaletsky, Tomoko Kuroda-Kawaguchi, Patrick J Minx, Holland S Cordum, Ladeana Hillier, Laura G Brown, Sjoerd Repping, Tatyana Pyntikova, Johar Ali, Tamberlyn Bieri, Asif Chinwalla, Andrew Delehaunty, Kim Delehaunty, Hui Du, Ginger Fewell, Lucinda Fulton, Robert Fulton, Tina Graves, Shun-Fang Hou, Philip Latrielle, Shawn Leonard, Elaine Mardis, Rachel Maupin, John McPherson, Tracie Miner, William Nash, Christine Nguyen, Philip Ozersky, Kymberlie Pepin, Susan Rock, Tracy Rohlfing, Kelsi Scott, Brian Schultz, Cindy Strong, Aye Tin-Wollam, Shiaw-Pyng Yang, Robert H Waterston, Richard K Wilson, Steve Rozen, and David C Page. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423(6942):825–837, 19 June 2003. 124

[131] Fernando L Mendez, G David Poznik, Sergi Castellano, and Carlos D Bustamante. The divergence of neandertal and modern human Y chromosomes. *Am. J. Hum. Genet.*, 98(4):728–734, 7 April 2016. 124

[132] Mark A Jobling and Chris Tyler-Smith. Human y-chromosome variation in the genome-sequencing era. *Nat. Rev. Genet.*, 30 May 2017. 124

[133] J S Waye and H F Willard. Structure, organization, and sequence of alpha satellite DNA from human chromosome 17: evidence for evolution by unequal crossing-

over and an ancestral pentamer repeat shared with the human X chromosome. *Mol. Cell. Biol.*, 6(9):3156–3165, 1986. 125

[134] Peter E Warburton, Gillian M Greig, Thomas Haaf, and Huntington F Willard. PCR amplification of chromosome-specific alpha satellite DNA: Definition of centromeric STS markers and polymorphic analysis. *Genomics*, 11(2):324–333, October 1991. 125

[135] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, Alexey V Pyshkin, Alexander V Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A Alekseyev, and Pavel A Pevzner. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, 19(5):455–477, May 2012. 133

[136] Sergey Koren, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, 27(5):722–736, May 2017. 133

[137] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 15 April 2013. 133, 134

[138] Arthur C. Rand, Miten Jain, Jordan M. Eizenga, Audrey Musselman-Brown,

Hugh E. Olsen, Mark Akeson, and Benedict Paten. Mapping dna methylation with high-throughput nanopore sequencing. *Nat Meth*, 14(4):411–413, Apr 2017. Brief Communication. 136

[139] Carl R Woese and George E Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. 141

[140] Impact of 16s rrna gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. 141

[141] Marcus H Stoiber, Joshua Quick, Rob Egan, Ji Eun Lee, Susan E Celniker, Robert Neely, Nicholas Loman, Len Pennacchio, and James B Brown. De novo identification of DNA modifications enabled by Genome-Guided nanopore signal processing. 10 April 2017. 143