Fast Track

Christopher R. O'Donnell[1]
Hongyun Wang[2]
William B. Dunbar[1]

[1]Department of Computer Engineering, Baskin School of Engineering, University of California, Santa Cruz, CA, USA
[2]Department of Applied Math and Statistics, Baskin School of Engineering, University of California, Santa Cruz, CA, USA

Received April 11, 2013
Revised May 28, 2013
Accepted May 28, 2013

Research Article

Error analysis of idealized nanopore sequencing

This numerical study provides an error analysis of an idealized nanopore sequencing method in which ionic current measurements are used to sequence intact single-stranded DNA in the pore, while an enzyme controls DNA motion. Examples of systematic channel errors when more than one nucleotide affects the current amplitude are detailed, which if present will persist regardless of coverage. Absent such errors, random errors associated with tracking through homopolymer regions are shown to necessitate reading known sequences (*Escherichia coli* K-12) at least 140 times to achieve 99.99% accuracy (Q40). By exploiting the ability to reread each strand at each pore in an array, arbitrary positioning on an error rate versus throughput tradeoff curve is possible if systematic errors are absent, with throughput governed by the number of pores in the array and the enzyme turnover rate.

Keywords:
DNA sequencing / Error rates / Nanopore / Nucleic acids / Single-molecule sequencing

DOI 10.1002/elps.201300174

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

High-throughput sequencing technologies can generate genome-scale sequence data with high accuracy, making it possible to identify genomic markers for a growing list of common diseases, including cancers [1]. The leading commercial platforms (Roche, Illumina, Life Technologies) can generate 1–100 s of gigabases per instrument run, with run times on the order of hours to days. For technologies that achieve at most 1% raw error rates, however, read lengths are short, generally tens to hundreds of base pairs. Such short-read sequencing necessitates massive data storage requirements and complex bioinformatics algorithms for genome alignment and assembly, and complicates studies involving linkage analysis. Another drawback of short reads is highly repetitive regions, for example the trinucleotide repeat CGG that cause genomic instability in a condition known as fragile-X syndrome, are almost impossible to map. The short reads also require the devices to have a high degree of parallelization, so that there is sufficient coverage of the sequenced DNA to achieve desired error thresholds. Still another drawback is the need for an amplification step using enzymes that have less than 100% fidelity. In particular, it is common that starting material is amplified to create a library for sequencing, which then undergoes a second amplification reaction to create a clonal colony as in Illumnia's on-chip bridged amplification reaction [2]. Such intensive sample preparation may also require unattainable amounts of starting material. Despite these issues, the short-read and massively parallel devices control the market principally because they provide the highest throughput and sufficiently low error rates.

Single-molecule sequencing (SMS) devices have alleviated the sample preparation requirements of massively parallel devices by eliminating the need for template amplification [3]. The SMS from Helicos Biosciences (HeliScope) preserves the high-throughput feature (~3 Gb/day), but reads remain short (<60 bp) and errors are higher (3–5%), diminishing the value of simpler sample preparation [1]. The SMS from

Correspondence: Associate Professor William B. Dunbar, Department of Computer Engineering, Baskin School of Engineering, University of California, 1156 High Street, MS: SOE3, Santa Cruz, CA 95064, USA
E-mail: dunbar@soe.ucsc.edu
Fax: +1-831-459-4829

Abbreviations: ONT, Oxford Nanopore Technologies; **SMS**, single-molecule sequencing

Colour Online: See the article online to view Figs. 1–3 in colour.

Pacific Biosciences (PacBio RS) boosts read lengths to 10 kb, but throughput is reduced (<0.1 Gb/run) and error rates are considerably higher (15%). Errors can be reduced with this technology by using circular template DNA, but at the price of shorter read lengths [3]. Despite the high error rate, the long-read feature of the PacBio RS technology makes it useful to use in concert with short-read and low-error platforms, specifically for whole-genome sequencing in which the longer reads provide alignment scaffolds for the short read contigs (though DNA mapping technologies [4] are competing for this market).

The ideal sequencing platform would require minimal sample preparation and zero amplification, would be modular and scalable to ensure sufficient throughput for any given application, and would have sufficiently long reads and low errors to permit robust detection of any feature, including rare variants [5] and structural variants such as repetitive regions [1]. No single platform currently possesses all of these assets. Nanopores have been pursued as a candidate SMS platform in university research labs [6], and a subset of the resulting intellectual property has been commercialized, most notably by the company Oxford Nanopore Technologies (ONT) [7]. According to press releases and presentations at sequencing technology meetings, ONT's sequencing platform utilizes chemically modified biological nanopore channels, promising minimal sample preparation and read lengths up to 100 kb [8]. Sample preparation requires no amplification, nor labeling of nucleotides; instead, individual DNA strands are captured by electrophoresis into each nanopore channel from a bulk-phase chamber, and the impeded channel current is used to sense the nucleotides that pass through the limiting constriction of the channel. This paper considers a model method in which intact ssDNA is threaded through a biological pore for sequencing [9, 10], as opposed to an alternative approach in which mononucleotides are sensed in concert with exonuclease-catalyzed ssDNA hydrolysis above the pore [11]. Unfortunately, intact ssDNA passes too fast through the pore when the rates of electrophoresis are unimpeded (~1 Mb/s), when compared to the ionic current measurement bandwidth (~1 kb/s) [6]. To keep ssDNA motion within measurement bandwidths, a leading nanopore sequencing method uses a DNA polymerase enzyme perched on top of the pore to control the rate of each DNA molecule through the pore [9]. In this configuration, the sensitivity of biological pores for identifying the sequence of intact ssDNA has improved, with the occluded current through the MspA pore a function of four nucleotides positioned at the narrowest constriction of the channel [10], and ONT claiming modified pores that are sensitive to three nucleotides at a time.

We consider an idealized nanopore sequencer in which an enzyme controls ssDNA motion through the pore, and the ionic current amplitude is a function of one or more nucleotides. When more than one nucleotide affects the current (e.g. a triple), systematic errors may make it impossible to resolve certain sequences, regardless of depth of coverage; we quantitatively consider examples where this is the case. Notably, such errors would also persist when using nanopores regardless of the method of DNA control (i.e. with enzymes or by any other method). Absent these systematic errors, we consider next random errors introduced by the use of an enzyme to control ssDNA motion through the pore. Specifically, the enzyme is idealized by modeling ssDNA motion as moving in single nucleotide steps with durations from an exponential distribution of known rate (we ignore backtracking that has been experimentally observed [9, 10]). When homopolymer regions move through the pore with no change in current amplitude, the number of nucleotides associated with each detectable amplitude level must be inferred, and this introduces random insertion or deletion errors that can be reduced only by rereading the same sequence multiple times. We derive an analytic expression for the rate of error decay as a function of the number of reads, and examine the resulting error rate trends for known sequences (16.6 kb human mitochondrial DNA [12], 4.6 Mb *Escherichia coli* K-12 [13]). We then simulate nanopore signals to incorporate the effects of added measurement noise and the consequent low-pass filtering required to reduce noise for robust amplitude detection. Using a novel amplitude-level detection and duration binning method for base calling, consensus sequences generated in the noiseless case are shown to match the analytic trends exactly, and increasing noise is shown to increase the error rate. Finally, for devices that scale to what ONT has promised, we examine the tradeoff between throughput and error rate as a function of the number of rereads and the rate of enzyme turnover, with a faster enzyme increasing both throughput and error rate.

2 Materials and methods

2.1 Simulated nanopore signals

All simulations were performed using the Matlab software package. The nanopore sensor was modeled as having single nucleotide sensitivity producing distinct ionic current amplitudes at (3, 2, 1, 0) pA for the nucleobases (A, G, C, T). We varied this amplitude-to-base assignment and observed no measurable difference in the computed error rates for the sequences considered (Supporting Information Fig. 3). The passage of DNA through the nanopore was modeled as unidirectional with the lifetime of each nucleotide in the sensor from an exponential distribution of known rate. Simulated data were produced by first generating an ideal pulse-train signal at 10 MHz for a chosen DNA sequence, where the dwell time for each nucleobase was randomly selected from the exponential distribution with mean 1 ms. White noise was added to the idealized signal, which was then low-pass Bessel filtered at 100 kHz and downsampled to 500 kHz. White noise variance, which we label as "1× noise," was chosen to produce a 2:1 S/N when the Bessel filter was set to 5 kHz bandwidth to emulate conditions comparable to those observed experimentally [9, 10]. Analysis of signals with 2× this noise, and without noise, was also performed. At 1× noise, the mean enzyme rate was also varied (0.1, 0.5, 1, and 10 ms, Supporting Information Fig. 5) to examine its influence on error rate performance.

2.2 Base-calling algorithm, alignment, and consensus

Noise on each simulated signal was reduced by applying a running mean filter followed by a Savitzky–Golay filter of order 2. To identify ionic current levels, a custom step detection algorithm was employed using a gradient threshold to detect transitions between levels and amplitude thresholds to classify levels by nucleobase. The number of nucleotides assigned to each current level was determined using a binning method to sort each level by its duration. The optimal sizes of the bins were chosen to maximize the sum of the probabilities that each current level is assigned the correct number of bases (Supporting Information). Error calculations were performed by comparing the predicted sequence to a known reference sequence. The current levels of the two sequences were globally aligned using the Matlab function "nwalign" with affine gap penalties. The numbers of nucleotides at each aligned level were compared and errors in the predicted sequence were classified as insertions, deletions, or substitutions. Insertions and deletions were counted on a per nucleotide basis, whereas substitutions were counted in terms of the number of current levels with misidentified amplitudes.

Multiread consensus sequences were generated by first performing a progressive multiple alignment of the ionic current levels of the reads using the Matlab function "multialign" with the option "TerminalGapAdjust" set to true. The multiple alignment was used to generate a consensus sequence of current levels using the Matlab function "seqconsensus" with the option "Gaps" set to "all." Nucleotides were then assigned to the consensus sequence current levels using the optimal binning method, where the duration of the consensus levels were determined by computing the mean duration for each level. To ensure that the correct current levels were included in the calculation, each predicted sequence used to generate the consensus was globally aligned with the consensus sequence and only the durations of the aligned current levels were used for computing the mean duration times. Error analysis for the multiread consensus sequences was performed in the same manner as for the single-read predicted sequences.

3 Results and discussion

3.1 The potential for systematic channel errors

In the simplest case that the current amplitude is a function of only one nucleotide in the channel, a necessary and sufficient condition for recovering DNA sequences is that each letter generates a distinct amplitude that can be detected above experimental noise. When more than one nucleotide affects the current amplitude, it is less clear what sequence can be recovered. Consider the case where three nucleotides affect the current. If 64 distinct and detectable amplitude levels are generated for every triple-letter combination of the four nucleotides (i.e. 4^3), then there is no ambiguity in the identified sequence. On the other hand, if there are less than 64 detectable levels, there may or may not be ambiguity. Below, we present a generalized case in which there are an infinite number of sequences that can not be recovered, regardless of how many times the sequence is read.

Assume the current is a function of three nucleotides, and suppose the four triples in the set {CCC, CCA, CAC, ACC} generate the same amplitude. Then, for any $n \geq 1$, there is a set of length-n subsequences $Z_1 \cdots Z_n$ constructed from A and C that cannot be distinguished from each other within the sequence $CCZ_1 \cdots Z_nCC$. As a specific example, within the sequence $\cdots$TCCCACCACCG$\cdots$, the subsequence CACCA cannot be differentiated from CCCCC, ACCAC, or any other five-letter combination of C and A in which As are separated by two or more Cs. A proof of the generalized statement, and comparable statements for cases when the amplitude is a function of two or four nucleotides, are provided in the Supporting Information document. Experimentally building a map from letters to amplitude is required to determine if such channel errors are present for a given nanopore. Practically, two sequences would be considered to have the "same amplitude" if the magnitude of the difference between the two amplitude levels has a S/N of less than 1.5 after applying the low-pass filters designed for signal-to-sequence conversion (S/N 1.5 is a minimum threshold for idealizing the signal by Markov-based methods [14, 15], with S/N 2 or larger required for simpler methods [16], Supporting Information). While additional low-pass filtering can always boost S/N and therefore improve discrimination between amplitude levels, too much filtering will result in excessive deletions of fast events. Thus, the effective filter bandwidth designed for optimal signal-to-sequence conversion will tradeoff S/N for detection time resolution. For the remainder of the paper, we idealize the sequencing problem and assume there are no systematic channel errors. Specifically, the current amplitude is assumed to be a function of one nucleotide at a time (i.e. the channel is single-nucleotide sensitive).

3.2 Errors due to nondeterministic sensing times

By using an enzyme to control the motion of ssDNA through the nanopore [9, 10], the strand is temporarily immobilized for sensing before moving in single-nucleotide steps. A challenge for base calling is that the duration in each immobilized position is nondeterministic. For an ideal enzyme, we can model the duration as following an exponential distribution of known mean dwell time τ. We consider two complications with nondeterministic sensing times. First, without some signal that the enzyme has moved to the next position on the DNA, inferring the length of each detected subsequence is challenging, and in particular one expects errors to grow with the length of homopolymer regions. Second, when experimental noise requires the use of low-pass filtering to permit robust detection of each sequence-specific amplitude level, a fraction of sensing times are too fast for detection and result in an increase in deletions. We consider first the errors that are intrinsic to inferring the length of the sequence that corresponds to each detected amplitude level, and then the errors induced by adding noise to the idealized sensing signal.

The sensing problem is idealized by assuming that the current measurements are noiseless and sensitive to one nucleotide at a time. Specifically, each base (A, G, C, T) generates the current amplitude (3, 2, 1, 0) pA, and each level is detectable regardless of duration (i.e. no durations are too fast since no noise filtering is required). A challenge is that we assume the enzyme does not provide an explicit tracking mechanism within the ionic current signal, which is consistent with the literature [9, 10], and so no new information can be extracted from the signal unless a sequence-specific amplitude shift is detected. This makes it difficult to identify the length of the sequence that corresponds to each detectable level. Mathematically, let τ_i be the duration during which the i-th nucleotide along the DNA is at the sensing position that determines the amplitude level. Each τ_i is an exponentially distributed random variable with mean τ. In our model problem, the transition from the i-th nucleotide to the $(i+1)$-th nucleotide being at the sensing position is detectable only if these two nucleotides are different (and thus yield different amplitude levels). Let s_j be the duration of the j-th segment along the time series of four distinct amplitude levels. If the sequence was entirely nonrepeating, $\tau_i = s_i$ for all $i = 1, \ldots, n_t$, with n_t the length of the sequence. To quantify the challenge of inferring sequence length in general, consider the example of the sequence TCCCAGG moving through the nanopore sensor starting from the right end. Sensing G first, we measure amplitude 2 pA for the duration $s_1 = \tau_1 + \tau_2$. Next, we measure A at amplitude 3 pA for the duration $s_2 = \tau_3$. Next, we measure C at amplitude 1 pA for the duration $s_3 = \tau_4 + \tau_5 + \tau_6$. Finally, sensing T we measure amplitude 0 pA for duration $s_4 = \tau_7$. The length of the sequence at each detected level must be inferred. For a single pass through the sequence, if s_4 gets a large sample value from the exponential distribution, it would appear that more than one T is present; likewise, if s_3 is made up of three faster-than-average durations, it would appear that less than three Cs are present. Clearly, such random errors can be reduced only by repeatedly taking measurements of s_j for each level, and generating a consensus (average) time for that level from which the sequence length estimate is made.

We derive a time-binning strategy that estimates the length k of each sequence from the measured duration s_j at each nucleotide-specific amplitude level. Since each $s_j = \sum_{i=1}^{k} \tau_{i+i_0}$ is the sum of k independent samples of an exponentially distributed random variable, each s_j has a Gamma distribution. By rereading the sequence n times, denoting the measured set of durations $\{s_j^1, \ldots, s_j^n\}$, the estimate for sequence length (k_{est}) for each detected level is computed using the variable $x = \left(\frac{1}{n}\sum_{l=1}^{n} s_j^l\right)/\tau$ with the simple equation:

$$k_{\text{est}}(n) = \begin{cases} 1, & \text{if } x \le b_1 \\ 2, & \text{if } b_1 < x \le b_2 \\ 3, & \text{if } b_2 < x \le b_3 \\ \vdots & \end{cases}, \qquad (1)$$

with optimized bin values $(b_1, b_2, b_3, \ldots) = (1.472, 2.483, 3.488, \ldots)$ chosen to minimize the error rate (Supporting Information). If the average s_j is between 2.483τ and 3.488τ for a detected level at 1 pA, for example, the estimated length is $k_{\text{est}}(n) = 3$ producing the sequence estimate CCC. Since the random variable $(x \cdot n)$ has a gamma distribution with shape parameter $(k \cdot n)$ and scale parameter 1, the error rate per nucleotide for a k-repeat based on measurements from n reads is $\text{Err}(k, n) = \frac{1}{k}\sum_j |j - k| \Pr(k_{\text{est}}(n) = j)$. This error rate has an analytical expression that can be computed in Matlab using the incomplete gamma function (Supporting Information). From this expression, the per-nucleotide error rate $g(n)$ is computed for any given sequence as a function of the number of reads n, and is given by the equation:

$$g(n) = \frac{1}{n_t}\sum_{k=1}^{m} q_k \cdot \text{Err}(k, n), \qquad (2)$$

where n_t is the length of the sequence, q_k is the total number of nucleotides belonging to length-k repeats in the sequence, and m is the longest repeat length present in the given sequence.

We computed error rates using Eq. (2) for four different sequences, including the 16.6 kb human mitochondrial DNA sequence [12], and for the 4.6 Mb *E. coli* K-12 sequence [13] (Fig. 1). The other two sequences are 50 nucleotides in length and are used to show the influence of homopolymer length. Not surprisingly, the nonrepeating sequence has the fastest rate of error decay and, as expected, longer stretches of homopolymer regions require a greater number of reads

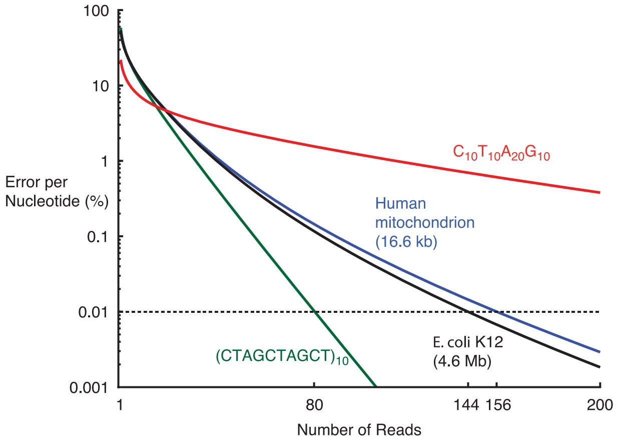

Figure 1. Analytic error rates for enzyme-controlled ssDNA nanopore sequencing. The idealization assumes a noiseless single-nucleotide sensor, but with no mechanism for tracking single-nucleotide displacements through homopolymer regions. Durations of each nucleotide in the sensor are from a single exponential distribution of known rate, consistent with an ideal enzyme controlling DNA motion through the sensor. The analytic error rates are computed for human mitochondrion [12], and for *E. coli* K-12 [13], for a 50 mer nonrepeating sequence (green) and a 50 mer with length 10 and 20 mer homopolymer regions (red). Error reduction is accomplished only by rereading the same sequence and averaging the duration at each resolvable sequence-specific amplitude level.

to reduce the error. With no mechanism for tracking the motion progress through homopolymer regions of ssDNA, rereading the same strand is required to reduce errors to acceptable levels. The figure suggests that to achieve the Q40 standard (99.99% confidence) requires reading known sequences ~150 times. Achieving multiple reads could be accomplished by single-pass reading of many copies in parallel in a multichannel array, or by rereading the same strand at each pore [9]. When considering nanopore sensors that are a function of more than one nucleotide, the analytic error rate performance improves, but only if there are no systematic channel errors (Supporting Information Fig. 6). Notably, the single-read error improves from 40.5% per nucleotide for a single-nucleotide sensor to 1.24% per nucleotide for a four-nucleotide sensor, for the 16.6 kb human mitochondrial DNA. The improvement is a byproduct of being able to detect the length of homopolymers that are the same length or shorter than the sensor footprint. The improvement is less dramatic, however, when higher accuracy is needed (the four-nucleotide nanopore sensor requires 130 reads for Q40 accuracy, Supporting Information Fig. 6).

3.3 Influence of measurement noise and enzyme rate on base-calling errors

To consider next the effect of measurement noise on base-calling performance, we simulate ionic current signals. The mapping of bases (A, G, C, T) to the amplitude (3, 2, 1, 0) pA was again used, with the sequence of the first 50 nucleotides of the Mitochondrial DNA sequence [12] used to generate each signal. For each signal, durations for each base were randomly drawn from an exponential distribution with mean $\tau = 1$ ms. A gradient-based algorithm was developed for level detection, and the time-binning strategy in Eq. (1) was used to assign the number of bases for each detected level. The reference sequence was used to compute the errors for each estimated sequence and for multiread consensus sequences (Supporting Information). To emulate experimental noise, a sufficient amount of white noise is added to the unfiltered ideal signal to produce ~0.5 pA root-mean-square after low-pass Bessel filtering at 5 kHz bandwidth. This noise we label as "1× noise" and results in S/N of 2 between adjacent amplitude levels at 5 kHz bandwidth, which is sufficient for detection by standard methods [16] and by our gradient-based algorithm. The simulation makes use of a model of the nanopore instrument that has been experimentally validated [17], specifically by including the low-pass Bessel filter used in the current sensing amplifier. The Bessel filter is set at 100 kHz bandwidth and additional filtering is performed for robust level detection, which is similar to what is done experimentally [10]. The case of 2× noise has two times the variance of the added white noise before filtering, and is also considered. We used the same base-calling logic and filter settings for 1× and 2× noise, though the filter settings were optimized for robust level detection at the 1× noise condition. The tradeoff in noise filtering and level-detection fidelity is central to sequencing error performance; therefore, settings would be optimized for the given S/N and time resolution constraints imposed by the instrument and enzyme rate in actual experiments. Example current traces show the difference between no noise, 1× and 2× noise (see Fig. 2A). The error rates for multiread consensus sequences are generated as a function of the number of reads, and compared to the analytic curve for the 50 nucleotide sequence (see Fig. 2B).

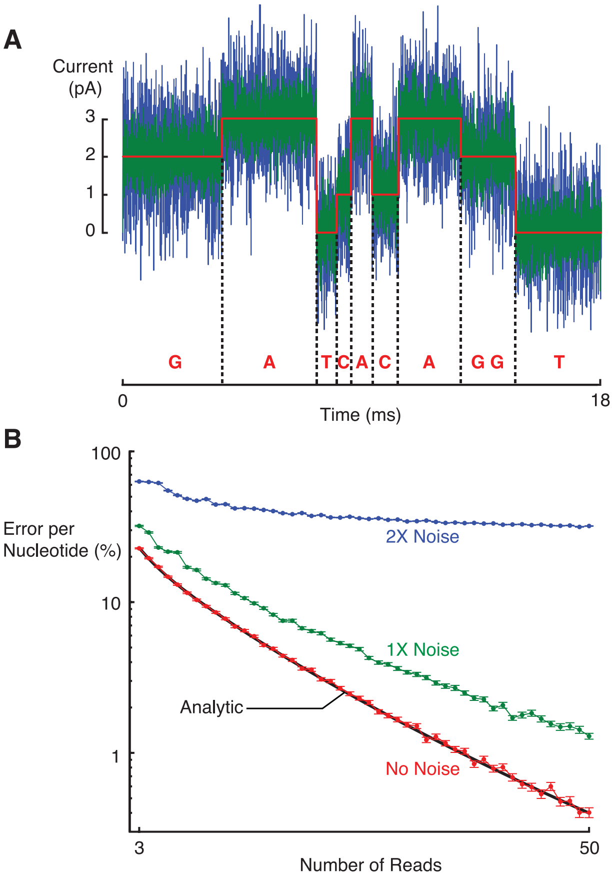

Figure 2. Base-calling logic applied to simulated nanopore signals shows error rate performance that matches the analytic error rate in the absence of noise, and increasing error rates with measurement noise. (A) Example signal traces for the first ten nucleotides in the sequence, with no noise (red), 1× noise (green), and 2× noise (blue). The randomness of level durations shows the need for multiple reads to identify sequence lengths with confidence. (B) Mean error rates as a function of number of reads for the first 50 nucleotides of the human mitochondrial DNA sequence [12]. Data points are the mean error per nucleotide from 900 independent multiread consensus sequences, with each consensus computed using the reported number of reads and with each read being drawn from a set of 10 000 simulated signals. Error bars are the standard error, computed as the SD of the error divided by $\sqrt{900}$.

In the absence of noise, the error-rate performance of our computed consensus sequences matches the analytic trend exactly, validating our simulation and base-calling algorithm. The errors for both analytic and simulated (noiseless) trends are broken down as 82% insertions and 18% deletions on average, with no substitutions (see Supporting Information Fig. 1). The largest source of error is insertions because 68% of the nucleotides in this specific sequence are nonrepeating, and only insertion errors are possible in the noiseless case since every level is detectable. As the number of reads is increased, insertions become the dominant remaining error source. When 1× noise is added, the filtering required to reduce noise causes faster levels to go undetected, creating an increase in the fraction of deletions for each estimated sequence. Specifically, the fastest dwell that our level-detection method can robustly detect is 170 μs (see Supporting Information Fig. 2), and for a mean enzyme duration of 1 ms, $1 - 1/\exp(0.17/1) = 0.16$ or 16% of dwells are too fast for robust detection. The presence of noise can also cause fast levels to transiently appear at the wrong amplitude, resulting in substitutions. An increasing fraction of deletions and substitutions are observed in the error breakdown for consensus sequences at 1× noise, particularly for a low number of reads (for three read consensus, 72% insertions, 26% deletions, 2% substitutions, see Supporting Information Fig. 1). As the number of reads is increased, insertions become the dominant remaining error source, consistent with the noiseless case (see Supporting Information Fig. 1). At 1× noise, we varied the base-to-amplitude mapping to test if our original mapping choice was biasing the error rate performance with noise. The results show no significant difference in the error rate trends (see Supporting Information Fig. 3). When noise is further increased to 2×, a substantial new source of error is that spurious level-changes induced by the noise are detected, causing substitutions, deletions, and incorrect calculation of durations. Insertions remain a large source of error (40–70%), and deletions become the greatest source of error for consensus sequences using more than 25 reads (see Supporting Information Fig. 1). While the base-calling performance at 2× noise is unacceptably bad, it should again be qualified that the filtering and base-calling logic was not re-optimized for the 2× noise case but kept the same as for the 1× noise case. Practically, both filtering and logic will be optimized according to the noise and level-detection performance of a given nanopore platform.

With an enzyme that follows a single exponential distribution, the mean dwell time τ will not be precisely known in practice. Using an estimate $\hat{\tau}$ of the true mean τ in the base-calling logic will incur errors. If $\hat{\tau} < \tau$, the length of each sequence will be overestimated causing insertion errors. Likewise, $\hat{\tau} > \tau$ will cause underestimation of sequence lengths and result in deletion errors. Additionally, the larger or smaller $\hat{\tau}/\tau$ becomes, the greater the errors. As an example, if $x = 1.3$ is computed with known τ and for a level corresponding to the single-nucleotide C, the estimated and correct length is $k_{\text{est}}(n) = 1$ from Eq. (1). However, if $\hat{\tau} = 0.85\tau$ is used to compute x, then it becomes $x = 1.53$ and Eq. (1) produces an insertion error with estimated length $k_{\text{est}}(n) = 2$. To assess the effect of incorrectly estimating τ on error rate performance, we considered two extreme cases at 1× noise: overestimating the mean dwell by double ($\hat{\tau} = 2\tau$), and underestimating the mean dwell by half ($\hat{\tau} = 0.5\tau$). The incorrect estimates for the mean were used in the calculation of $x = \left(\sum_{l=1}^{n} s_j^l\right) / (n\hat{\tau})$, which is used to compute the length estimate (k_{est}) of each sequence at each detected level in Eq. (1). As expected, overestimating the mean dwell creates deletion errors, with an error rate of 16% that persists even up to 30 reads (see Supporting Information Fig. 4). Error rate performance is considerably worse when underestimating the mean dwell time by half, with a persistent error over 100% that is comprised almost exclusively of insertion errors (see Supporting Information Fig. 4).

Assuming the mean enzyme dwell time τ is known, we considered also the effect of different τ values on the error rate performance, again using the 1× noise condition. The filtering required for robust amplitude-level detection at 1× noise results in an increasing fraction of levels that go undetected as τ is decreased, and error rates are considerably worse as τ decreases below 1 ms. On the other hand, our base-calling method applied to 1× noisy signals is observed to perform as well as is theoretically possible (i.e. matching the analytic trends derived for noiseless signals) when $\tau > 10$ ms (see Supporting Information Fig. 5). The phi29 enzyme as a replication-driven ratchet has mean dwell $\tau = 36$ ms, computed as the reported 25 ms median dwell [9] divided by ln(2), but this does not suggest that the theoretically optimal error rate is achievable. Specifically, our idealization ignores backtracking that is experimentally observed with phi29, and the noise and channel sensing characteristics do not match our idealization. Nonetheless, $\tau = 36$ ms means that 99.5% of levels are resolvable if the setup can robustly detect 170 μs (see Supporting Information Fig. 2). Although a slower enzyme will reduce the number of deletion errors caused by filtering out fast events, it also means lower throughput. The viability of a commercial nanopore sequencer will be determined by both the throughput and the error rates that are achievable, and these are a function of the scale of the multichannel array that can be incorporated (fluidics, circuitry) into a platform of a given size [8].

The final topic of this paper examines the tradeoff between throughput and error rate in an idealized multichannel nanopore sequencing platform, while varying both the number of channels n_c and the mean enzyme turnover rate$1/\tau$. The idealized throughput $G(n)$ is a function of the number of reads n per DNA molecule in each pore, and is computed as:

$$G(n) = 3.6(\text{Mb/h})\frac{n_c}{n \cdot \tau(\text{ms}^{-1})}, \tag{3}$$

where 3.6 is a conversion factor so that τ in units (ms) results in G in units (Mb/h). The calculation ignores the open channel time between DNA captures, and assumes a common read number n is implemented for every pore in the array. The analytic error rate $g(n)$ from Eq. (2) is plotted versus the throughput $G(n)$ for the 16.6 kb human mitochondrial DNA sequence [12], varying n from 3 to 200 reads and with $\tau = 10$ ms (see line in Fig. 3). For reference, the figure shows the throughput required to sequence a diploid human genome (6 billion base pairs) in 24 h (6 Gb/day). The analytic error

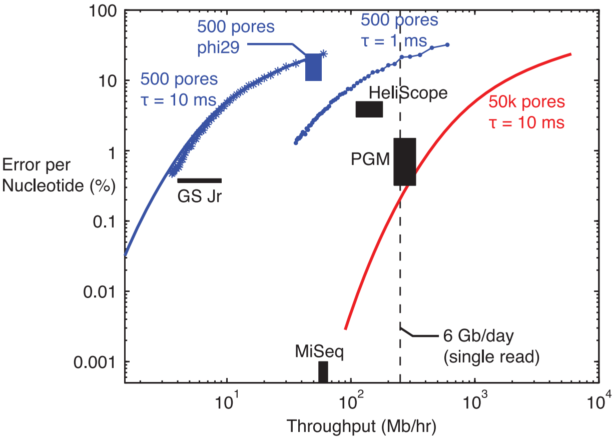

Figure 3. Examining the tradeoff between error rate and throughput while comparing commercial sequencing platforms with prospective nanopore sequencing platforms, varying the mean enzyme rate, number of channels and number of reads per DNA. Commercial devices (black) include benchtop platforms (Roche/454 GS Jr, Life/Ion Torrent PGM, Illumina MiSeq) and the SMS from Helicose (HeliScope) with base call error rate and throughput ranges reported in [1,3,18]. Prospective nanopore devices (color) are the noiseless idealized case (lines), the 1× noise case (data), and based on the reported experimental performance for the phi29 enzyme [9] (box). Throughput is defined by Eq. (3) and assumes all 500 (blue) or 50 000 (red) pores operate independently and continuously at mean enzyme rate $1/\tau$, ignoring the time between captured molecules. The vertical dashed line represents the throughout for single-read sequencing of 6 Gb (approximate size of human genome in somatic cells) in 24 h. Error rates for simulations are based on the derived analytic trends for the noiseless case (16.6 kb, 3–200 reads, see Fig. 1), or the derived base-calling method applied to 1× noisy signals (50 nt, 3–50 reads, see Fig. 2B) for the Human Mitochondrial DNA sequence [12]. With phi29 as a replication-driven ratchet, the error rate of 10–24.5% is for a single read per DNA. Throughput is based on mean dwell $\tau = 36$ ms, computed as the reported 25 ms median dwell divided by ln(2), and adding ±6 ms range in calculation of throughput for display purposes.

rate curves show the best possible error rate versus throughput tradeoff for $n_c = 500$ and 50 000 channels, by assuming there are no systematic channel errors that mask the sequence and there is no noise on the signal. Improvement on these trends is possible only if, in addition to these assumptions, a robust mechanism that signals tracking through homopolymer regions is available in the measured current (no such mechanism has been reported in the literature). If noise is present but τ is sufficiently large (10 ms in our simulations), base-calling logic could in principle perform comparably to the optimal tradeoff (ignoring backtracking by the enzyme). As τ decreases tenfold from 10 ms to 1 ms, throughput increases tenfold while error rate increases by only twofold on average for n from 3 to 50 reads (see data points in Fig. 3). An enzyme with τ below 1 ms performs considerably worse in terms of error rate (Supporting Information Fig. 5). This suggests that a given platform will have an optimal enzyme rate that can maximize throughput while providing acceptable error rates. For a given enzyme, positioning on the tradeoff curve can exploit the ability to reread each strand in each pore [9], which would be preferable to single reads of many copies of the same strand in all pores, particularly for rare variant detection. Also shown is the single-read performance of phi29, adapted from the reported error rate [9] and assuming 500 single-read pores for the throughput calculation (see blue box in Fig. 3). To contrast these prospective platforms with commercial sequencing platforms, the error rate versus throughput reported in the literature [1, 3, 18] is shown for three benchtop devices and the Heliscope (see black boxes in Fig. 3).

4 Concluding remarks

In conclusion, our error analysis shows the need to reread the same molecule at each pore, or to read identical copies of the molecule serially or in parallel pores, when ionic current nanopore sequencing is used in conjunction with enzymes to control DNA motion. Systemic errors caused by the channel's inability to sense and differentiate specific sequences may or may not be present for a given pore. If present, such errors define an error rate threshold below which the platform cannot go, regardless of the number of reads. Random indel errors, on the other hand, can be reduced by increasing the number of reads, and we provided the first analytic expression that defines the best possible rate of error reduction as a function of the number of reads. The error rate versus throughput tradeoff results show that platforms on the scale of what ONT has promised (512 channel hand-held MinION, or 25 2000-channel GridIONs for 50 k total channels) are on the spectrum of commercial sequencing device performance, and will be directly competitive if systematic error thresholds are below that of the competitions.

Reduction in instrument complexity is an advantage for prospective nanopore devices that may trump any disadvantages associated with higher systematic error threshold or indel error frequency, though this will only become clear when such devices become available to users. Specifically, the prospective device presumably eliminates the need to build or amplify sequencing libraries, reduces the complexity of fluidics required during the sequencing operation (unlabeled nucleotides), and could make resequencing permissible with no fluid exchange. Even with the same raw error rate (5–15%) and read lengths (250 bp–10 kbp) as Pacific Biosciences RS platform [1], a considerably less complex device can be much cheaper and portable. There is presently no "cheap, quick, and dirty" sequencing technology; however, a hand-held nanopore sequencer may be such a technology. Even with modestly higher error rates, long read-length, and portable sequencing platforms would undoubtedly find applications, e.g. for fast resequencing or targeted sequencing of pathogen strains [19], provided the user interface is as simple as other devices used routinely in clinical settings.

We conclude our paper with a brief discussion on assigning error probabilities to sequences, as this is a forward-looking issue that will benefit from basic research as nanopore sequencing technologies come to market. Assigning a statistical measure of confidence to sequencing data is important for determining the suitability of sequencing results for a given application, and also for providing a quantitative basis for comparing data generated from different technologies [20]. The de facto metric for comparing the probability of error for a sequence across platforms is the position-specific quality score (Q-score). Quality scores originated with the base calling program phred, which uses an algorithm and a four-parameter set associated with the error characteristics of the Sanger method to compute the score [21]. The accuracy of the quality scores has been key to the utility of Sanger sequencing data [22]. The quality scores currently reported for next-generation high-throughput sequencing techniques are on the same numerical scale as phred quality scores, but are not as accurate [18, 20, 22]. Quality scores are less accurate in part because the parameters derived for the Sanger sequencing method do not isomorphically (in a mathematical sense) capture error characteristics of the other sequencing methods. To identify an accurate metric of base quality for a nanopore sequencing method, appropriate parameters built on the base-call error characteristics of nanopore signals needs to be identified. More broadly, until a universal standard is developed for defining accuracy for next-generation sequencing, the value of combining sequence data from different technologies will not reach its full potential.

This work was supported by National Institutes of Health grant 1R21HG006561-01 (W. B. D.). We thank T. Morin for reviewing a draft of this manuscript.

The authors have declared the following potential conflict of interest: W. B. Dunbar has an equity interest in TPG, Inc. The terms of this arrangement have been reviewed and approved by the University of California, Santa Cruz in accordance with its conflict of interest policies.

5 References

[1] Xuan, J., Yu, Y., Qing, T., Guo, L., Shi, L., *Cancer Lett.* 2013, DOI: 10.1016/j.canlet.2012.11.025.

[2] Myllykangas, S., Buenrostro, J., Ji, H. P., in: Rodríguez-Ezpeleta, N., Hackenberg, M., Aransay, A. M. (Eds.), *Bioinformatics for High Throughput Sequencing*, Springer, New York 2012, pp. 11–25.

[3] Thompson, J., Milos, P., *Genome Biol.* 2011, *12*, 217.

[4] Lam, E. T., Hastie, A., Lin, C., Ehrlich, D., Das, S. K., Austin, M. D., Desh- pande, P., Cao, H., Nagarajan, N., Xiao, M., Kwok, P.-Y., *Nat. Biotechnol.* 2012, *30*, 771–776.

[5] Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., *Proc. Natl. Acad. Sci. U.S.A.* 2011, *108*, 9530–9535.

[6] Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Ventra, M. D., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S. B., Krstic, P. S., Lindsay, S., Ling, X. S., Mastrangelo, C. H., Meller, A., Oliver, J. S., Pershin, Y. V., Ramsey, J. M., Riehn, R., Soni, G. V., Tabard-Cossa, V., Wanunu, M., Wiggin, M., Schloss, J. A., *Nat. Biotechnol.* 2008, *26*, 1146–1153.

[7] Pollack, A., *The New York Times*, Feb 17, 2012, pp. B2.

[8] Maitra, R. D., Kim, J., Dunbar, W. B., *Electrophoresis* 2012, *33*, 3418–3428.

[9] Cherf, G. M., Lieberman, K. R., Rashid, H., Lam, C. E., Karplus, K., Akeson, M., *Nat. Biotechnol.* 2012, *30*, 344–348.

[10] Manrao, E. A., Derrington, I. M., Laszlo, A. H., Langford, K. W., Hopper, M. K., Gillgren, N., Pavlenok, M., Niederweis, M., Gundlach, J. H., *Nat. Biotechnol.* 2012, *30*, 349–353.

[11] Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., Bayley, H., *Nat. Nanotechnol.* 2009, *4*, 265–270.

[12] Sanchez-Cespedes, M., Parrella, P., Nomoto, S., Cohen, D., Xiao, Y., Esteller, M., Jeronimo, C., Jordan, R. C. K., Nicol, T., Koch, W. M., Schoenberg, M., Mazzarelli, P., Fazio, V. M., Sidransky, D., *Cancer Res.* 2001, *61*, 7015–7019.

[13] Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J. Mau, B., Shao, Y., *Science* 1997, *277*, 1453–1462.

[14] Venkataramanan, L., Sigworth, F. J., *Biophys. J.* 2002, *82*, 1930–1942.

[15] Qin, F., Auerbach, A., Sachs, F., *Biophys. J.* 2000, *79*, 1928–1944.

[16] Sakmann, B., Neher, E. (Eds.), *Single-Channel Recording*, Plenum Press, New York 1995.

[17] Garalde, D. R., O'Donnell, C. R., Maitra, R. D., Wiberg, D. M., Wang, G., Dunbar, W. B., *IEEE Trans. Control Syst. Technol.* 2013, DOI: 10.1109/TCST.2012.2224349.

[18] Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., Pallen, M. J., *Nat. Biotechnol.* 2012, *30*, 434–439.

[19] Loman, N. J., Constantinidou, C., Chan, J. Z., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E. R., Pallen, M. J., *Nat. Rev. Microbiol.* 2012, *10*, 599–606.

[20] Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W., Russ, C., Lander, E., Nusbaum, C., Jaffe, D., *Genome Res.* 2008, *18*, 763–770.

[21] Ewing, B., Green, P., *Genome Res.* 1998, *8*, 186–194.

[22] Holt, R. A., Jones, S. J. M., *Genome Res.* 2008, *18*, 839–846.