PDF FORMATTED FOR
SPONSORED BY

Science
AAAS

Real-Time DNA Sequencing from Single Polymerase Molecules

John Eid, *et al.*

Science **323**, 133 (2009);
DOI: 10.1126/science.1162986

The following resources related to this article are available online at www.sciencemag.org (this information is current as of January 4, 2009):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/cgi/content/full/323/5910/133

Supporting Online Material can be found at:
http://www.sciencemag.org/cgi/content/full/1162986/DC1

A list of selected additional articles on the Science Web sites **related to this article** can be found at:
http://www.sciencemag.org/cgi/content/full/323/5910/133#related-content

This article **cites 36 articles**, 17 of which can be accessed for free:
http://www.sciencemag.org/cgi/content/full/323/5910/133#otherarticles

This article appears in the following **subject collections**:
Biochemistry
http://www.sciencemag.org/cgi/collection/biochem

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:
http://www.sciencemag.org/about/permissions.dtl

Downloaded from www.sciencemag.org on January 4, 2009

EX2001
Guardant Health, Inc. v. Univ. of Washington
IPR2022-00817

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2009 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

previously been shown in both birds and butterflies that structural color can enhance pigment color either by an additive or a contrast effect (*8*, *16*, *29*, *30*). This interplay of structure and pigment may therefore also add to the diversity of pollination cues utilized by the flowers of many angiosperm species.

References and Notes

1. A. R. Parker, *J. Opt. A Pure Appl. Opt.* **2**, R15 (2000).
2. A. Sweeney, C. Jiggins, S. Johnsen, *Nature* **423**, 31 (2003).
3. R. L. Rutowski *et al.*, *Biol. J. Linn. Soc. London* **90**, 349 (2007).
4. D. J. Kemp, *Proc. R. Soc. London B Biol. Sci.* **274**, 1043 (2007).
5. N. I. Morehouse, P. Vukusic, R. L. Rutowski, *Proc. R. Soc. London B Biol. Sci.* **274**, 359 (2007).
6. A. R. Parker, Z. Hegedus, *J. Opt. A Pure Appl. Opt.* **5**, S111 (2003).
7. J. E. Kettler, *Am. J. Phys.* **59**, 367 (1991).
8. P. Kevan, W. Backhaus, in *Color Vision: Perspectives from Different Disciplines*, W. Backhaus, R. Kliegl, J. S. Werner, Eds. (Walter de Gruyter, Berlin, 1998), pp. 163–168.
9. K. Noda, B. Glover, P. Linstead, C. Martin, *Nature* **369**, 661 (1994).
10. H. Gorton, T. Vogelmann, *Plant Physiol.* **112**, 879 (1996).
11. C. Hebant, D. W. Lee, *Am. J. Bot.* **71**, 216 (1984).
12. T. C. Vogelmann, *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **44**, 231 (1993).
13. P. B. Green, P. Linstead, *Protoplasma* **158**, 33 (1990).
14. C. Palmer, *Diffraction Grating Handbook* (Newport Corporation, Rochester, NY, ed. 6, 2005).
15. P. Skorupski, T. Döring, L. Chittka, *J. Comp. Physiol. A* **193**, 485 (2007).
16. R. L. Rutowski *et al.*, *Proc. R. Soc. London B Biol. Sci.* **272**, 2329 (2005).
17. P. Kevan, L. Chittka, A. Dyer, *J. Exp. Biol.* **204**, 2571 (2001).
18. K. Daumer, *Z. Vgl. Physiol.* **41**, 49 (1958).
19. F. Gandia-Herrero, F. Garcia-Carmona, J. Escribano, *Nature* **437**, 334 (2005).
20. R. Thorp, D. Briggs, J. Estes, E. Erickson, *Science* **189**, 476 (1975).
21. P. Kevan, *Science* **194**, 341 (1976).
22. L. Chittka, *J. Comp. Physiol. A* **170**, 533 (1992).
23. A. G. Dyer, L. Chittka, *Naturwissenschaften* **91**, 224 (2004).
24. H. M. Whitney, A. G. Dyer, L. Chittka, S. A. Rands, B. J. Glover, *Naturwissenschaften* **95**, 845 (2008).
25. P. L. Jokiel, R. H. York, *Limnol. Oceanogr.* **29**, 192 (1984).
26. Q. O. N. Kay, H. S. Daoud, C. H. Stirton, *Bot. J. Linn. Soc.* **83**, 57 (1981).
27. P. Kevan, M. Lane, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 4750 (1985).
28. B. Heuschen, A. Gumbert, K. Lunau, *Plant Syst. Evol.* **252**, 121 (2005).
29. M. D. Shawkey, G. E. Hill, *Biol. Lett.* **1**, 121 (2005).
30. R. O. Prum, in *Bird Coloration, Mechanisms and Measurements*, G. E. Hill, K. J. McGraw, Eds. (Harvard Univ. Press, Boston, 2006), vol. 1, pp. 295–353.

31. We thank S. Rands, P. Rudall, R. Bateman, P. Cicuta, and J. Baumberg for discussions and Syngenta for bees. Funded by Natural Environment Research Council grant NE/C000552/1, Engineering and Physical Sciences Research Council grant EP/D040884/1, the European RTN-6 Network Patterns, the Cambridge University Research Exchange, and German Academic Exchange Service DAAD.

Supporting Online Material
www.sciencemag.org/cgi/content/full/323/5910/130/DC1
Materials and Methods
Figs. S1 to S4
Tables S1 and S2
References

22 September 2008; accepted 6 November 2008
10.1126/science.1166256

Real-Time DNA Sequencing from Single Polymerase Molecules

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,† Stephen Turner†

We present single-molecule, real-time sequencing data obtained from a DNA polymerase performing uninterrupted template-directed synthesis using four distinguishable fluorescently labeled deoxyribonucleoside triphosphates (dNTPs). We detected the temporal order of their enzymatic incorporation into a growing DNA strand with zero-mode waveguide nanostructure arrays, which provide optical observation volume confinement and enable parallel, simultaneous detection of thousands of single-molecule sequencing reactions. Conjugation of fluorophores to the terminal phosphate moiety of the dNTPs allows continuous observation of DNA synthesis over thousands of bases without steric hindrance. The data report directly on polymerase dynamics, revealing distinct polymerization states and pause sites corresponding to DNA secondary structure. Sequence data were aligned with the known reference sequence to assay biophysical parameters of polymerization for each template position. Consensus sequences were generated from the single-molecule reads at 15-fold coverage, showing a median accuracy of 99.3%, with no systematic error beyond fluorophore-dependent error rates.

The Sanger method for DNA sequencing (*1*) uses DNA polymerase to incorporate the 3′-dideoxynucleotide that terminates the synthesis of a DNA copy. This method relies on the low error rate of DNA polymerases, but exploits neither their potential for high catalytic rates nor high processivity (*2*–*4*). Increasing the speed and length of individual sequencing reads beyond the current Sanger technology limit will shorten cycle times, accelerate sequence assembly, reduce cost, enable accurate sequencing analysis of repeat-rich areas of the genome, and reveal large-scale genomic complexity (*5*, *6*). Alternative approaches that increase sequencing performance have been reported [(*7*–*10*), reviewed in (*11*, *12*)]. Several of these methods have been deployed as commercial sequencing systems (*13*–*16*), which have greatly increased overall throughput, enabling many applications that were previously unfeasible. However, because these methods all gate enzymatic activity, using various termination approaches, they have not yielded longer sequence reads (limited to ~400 nucleotides), nor do they exploit the high intrinsic rates of polymerase-catalyzed DNA synthesis.

The use of DNA polymerase as a real-time sequencing engine—that is, direct observation of processive DNA polymerization with base-pair resolution—has long been proposed but has been difficult to realize (*7*, *8*, *17*–*22*). To fully harness the intrinsic speed, fidelity, and processivity of these enzymes, several technical challenges must be met simultaneously. First, the speed at which each polymerase synthesizes DNA exhibits stochastic fluctuation, so polymerase molecules would need to be observed individually while they undergo template-directed synthesis. Because of the high nucleotide concentrations required by DNA polymerases (*20*), a reduction in the observation volume beyond what is afforded by conventional methods, such as confocal or total internal reflection microscopy, directly improves single-molecule detection. Second, deoxyribonucleoside triphosphate (dNTP) substrates must carry detection labels that do not inhibit DNA polymerization even when 100% of the native nucleotides are replaced with their labeled counterparts. Third, a surface chemistry is required that retains activity of DNA polymerase molecules and inhibits nonspecific adsorption of labeled dNTPs. Finally, an instrument is required that can faithfully detect and distinguish incorporation of four different labeled dNTPs. Here, we provide proof-of-concept for an approach to highly

Pacific Biosciences, 1505 Adams Drive, Menlo Park, CA 94025, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: jkorlach@pacificbiosciences.com (J.K.); sturner@pacificbiosciences.com (S.T.)

Downloaded from www.sciencemag.org on January 4, 2009

multiplexed single-molecule, real-time DNA sequencing based on the observation of the temporal order of fluorescently labeled nucleotide incorporations during unhindered DNA synthesis by a polymerase molecule.

For the observation of incorporation events, we used a nanophotonic structure, the zero-mode waveguide (ZMW), which can reduce the volume of observation by more than three orders of magnitude relative to confocal fluorescence microscopy (*20*). This level of confinement enables single-fluorophore detection despite the relatively high labeled dNTP concentrations—between 0.1 and 10 μM—required by DNA polymerase for fast, accurate, and processive synthesis. This range produces average molecular occupancies between ~0.01 and 1 molecules for a ZMW 100 nm in diameter (*20*, *23*), compared with ~3 to 300 molecules for total internal reflection microscopy (*24*–*26*). The ZMW fabrication process was recently improved, resulting in a higher yield of devices suitable for single-molecule sequencing (*23*).

Other DNA sequencing approaches have used base-linked fluorescent nucleotides (*7*, *8*, *14*, *17*, *20*, *27*, *28*). These cannot be used in real-time sequencing because they are poorly incorporated in consecutive positions by DNA polymerase. In contrast, when a fluorophore is linked to the terminal phosphate moiety (phospholinked), phosphodiester bond formation catalyzed by the DNA polymerase results in release of the fluorophore from the incorporated nucleotide, thus generating natural, unmodified DNA (*21*, *29*–*31*). Φ29 DNA polymerase was selected for these studies because it is a stable, single-subunit enzyme with high speed, accuracy, and processivity that efficiently uses phospholinked dNTPs (*32*). It is capable of strand-displacement DNA synthesis and has been used in whole-genome amplification, showing minimal sequencing context bias (*33*). We introduced site-specific mutations in the enzyme and devised a linkage chemistry that allows 100% replacement of native nucleotides with four distinct phospholinked dNTPs while retaining near wild-type polymerase kinetics (*32*).

Recently, we reported a surface chemistry that enables selective immobilization of DNA polymerase molecules in the detection zone of ZMW nanostructures with high yield (*34*). Binding of polymerase molecules to the side walls is inhibited through the use of an alumina-specific polyphosphonate passivation layer. Here, an additional biotinylated polyethylene glycol layer was used to orient the polymerase and to prevent direct protein contact with the silica floor of the ZMW (*26*).

Extensions in the state-of-the-art of single-molecule detection were required to enable continuous, high-fidelity detection and discrimination of four spectrally distinct fluorophores simultaneously in large numbers of ZMWs. We reported a high-multiplex confocal fluorescence detection system (*35*) that uses targeted, uniform multilaser illumination of 3000 ZMWs through holographic phase masks. The instrument uses a confocal pinhole array to reject out-of-focus background, and a prism dispersive element for wavelength discrimination that provides flexibility in the choice of fluorescent dyes used while transmitting >99% of the incident light.

The architecture of our method is shown in Fig. 1A. DNA sequence is determined by detecting fluorescence from binding of correctly base-paired (cognate) phospholinked dNTPs in the active site of the polymerase (Fig. 1B). A fluorescence pulse is produced by the polymerase retaining the cognate nucleotide with its color-coded fluorophore in the detection region of the ZMW. It lasts for a period governed principally by the rate of catalysis, and ends upon cleavage of the dye-linker-pyrophosphate group, which quickly diffuses from the ZMW detection region. The duration of the fluorophore retention is much longer than the time scales associated with diffusion (2 to 10 μs) or noncognate sampling (<1 ms), which manifest as a low and constant background signal. Translocation prepares the polymerase active site for binding of the subsequent cognate phospholinked dNTP, which marks the beginning of the next pulse. Thus, the interpulse duration is a combination of the translocation and subsequent nucleotide binding times. The sequence of fluorescence pulses recorded in the plot of intensity versus time is referred to as a read.

To illustrate the principle of our approach to DNA sequencing, we used a synthetic, linear, single-stranded DNA template with a two-base artificial sequence pattern (Fig. 2A). Alternating template sections that omitted either cytosine or guanine were interrogated with their complementary phospholinked dNTPs, A555-dCTP and A647-dGTP. Reactions were initiated by addition of catalytically essential metal ions while collecting movies of fluorescence emissions simultaneously from an array containing 3000 ZMWs (movie S1). Time-resolved fluorescence spectra are presented for an example ZMW in Fig. 2B. For each movie frame, these spectra were reduced by dye-weighted summation to two values, representing the emission rate from each of the two phospholinked dNTPs as a function of time (Fig. 2C).

Single-molecule events corresponding to phospholinked dNTP incorporations manifested as fluorescent pulses whose variable duration reflected the enzyme kinetics and exhibited stochastic fluctuations in intensity (because of counting statistics and dye photophysics). The reads contained pulses with the expected pattern: alternating blocks of like-colored pulses corresponding to the alternating blocks in the template. Furthermore, we observed the hallmarks of single-molecule fluorescent events: single-frame rise and fall times at the start and end of the pulse, respectively (<<10 ms), which facilitate

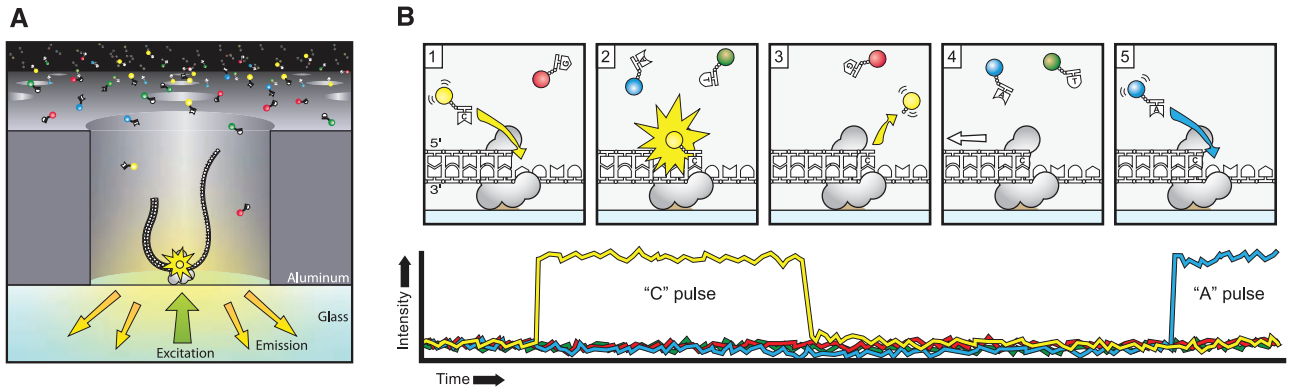


Fig. 1. Principle of single-molecule, real-time DNA sequencing. (**A**) Experimental geometry. A single molecule of DNA template-bound Φ29 DNA polymerase is immobilized at the bottom of a ZMW, which is illuminated from below by laser light. The ZMW nanostructure provides excitation confinement in the zeptoliter (10^{-21} liter) regime, enabling detection of individual phospholinked nucleotide substrates against the bulk solution background as they are incorporated into the DNA strand by the polymerase. (**B**) Schematic event sequence of the phospholinked dNTP incorporation cycle, with a corresponding expected time trace of detected fluorescence intensity from the ZMW. (1) A phospholinked nucleotide forms a cognate association with the template in the polymerase active site, (2) causing an elevation of the fluorescence output on the corresponding color channel. (3) Phosphodiester bond formation liberates the dye-linker-pyrophosphate product, which diffuses out of the ZMW, thus ending the fluorescence pulse. (4) The polymerase translocates to the next position, and (5) the next cognate nucleotide binds the active site beginning the subsequent pulse.

Downloaded from www.sciencemag.org on January 4, 2009

pulse detection and base calling even when pulses are close together (see, e.g., Fig. 2C, bottom right). Pulse activity ceased after completion of the 123 total nucleotide incorporations needed to reach the end of the linear template.

Read parameters for 10 representative molecules as well as an entire data set from a ZMW array are shown in table S1. The average DNA synthesis rate for 740 single-molecule reads was 4.7 ± 1.7 bases/s. Median and standard error values for the fluorescence intensity were 7353 ± 2970 (A555-dCTP) and 8408 ± 3381 (A647-dGTP) detected photons per second, yielding signal-to-noise ratios of 24 ± 10 and 25 ± 10, respectively. The observed pulse brightness showed uniformity within individual traces consistent with a stationary single-molecule process (modal coefficient of variation was 25%). When essential reaction components were withheld (e.g., DNA polymerase, primer/template, one or more phospholinked dNTPs or metal ions), <0.01 pulses ZMW^{-1} s^{-1} were observed, versus >1 pulses ZMW^{-1} s^{-1} for DNA polymerization conditions, additionally confirming that fluorescence pulse activity corresponded to DNA polymerase catalytic activity.

Immobilized enzymes in ZMWs maintained high activity. In the current implementation, polymerase molecules are randomly distributed among the ZMWs, leading to a Poisson distribution of occupancy (*34*). At optimal loading, the distribution is 36.8% empty ZMWs, 36.8% with just one polymerase, and 26.4% with two or more. In this experiment, ~35% of the ZMWs produced traces indicative of single DNA polymerase occupancy, of which 82% produced full-length reads. The yield limitations inherent to random polymerase immobilization could be exceeded using molecular self-assembly to place a single polymerase in each ZMW.

The real-time nature of the collected data allows the kinetics of the enzyme to be directly observed (fig. S1 and table S2). Single-polymerase kinetic parameters from these two-color block traces responded as expected when reaction conditions were modified. Reduction of the pH from 7.1 to 6.5 increased median pulse widths, consistent with a decrease of the rate of phosphodiester bond formation (*36*). Increasing the phospholinked dNTP concentration from 100 to 250 nM increased the median DNA synthesis rate, consistent with rate-limited dNTP binding at low concentration.

To investigate the potential of this method for long-read DNA sequencing, we performed a similar two-base signature sequence pattern experiment using a single-stranded 72-base circular DNA template (Fig. 3A). The template was designed such that cytosines were present on only half of the circle, and guanines on the other half. Φ29 DNA polymerase is highly processive (>70,000 bases) without cofactors in bulk reactions (*2*). It will carry out multiple laps of DNA strand-displacement synthesis around the circular template and has been shown to retain this activity in ZMWs (*34*).

A representative read (Fig. 3B) showed the expected continuous signature sequence pattern of alternating periods of A555-dCTP and A647-dGTP pulses. Pulse characteristics were similar to those described for Fig. 2 and remained uniform throughout a read. DNA polymerization activity lasted for thousands of seconds, allowing observation of several kilobases of DNA synthesis (top axis, Fig. 3B). Occasional pauses in DNA polymerization activity are visible as gaps in the trace. The total synthesized DNA length as a function of time (Fig. 3C) shows periods of different persistent polymerization rates during these long reads. Two characteristic polymerization rates of ~2 bases/s and ~4 bases/s were determined, suggesting the existence of different long-lived polymerase modes that occasionally and suddenly interconvert. No spatial correlation in the polymerase speed was observed across a ZMW array. Pulse characteristics underlying these two states were statistically identical, with the exception of a decreased interpulse duration for the faster state (fig. S2). Similar behavior was also observed using different combinations of the fluorophores and bases and for templates with different sequences (fig. S3), which implies that these states are specific neither to the phospholinked dNTPs used nor to the sequence context. Several prolonged observations were made in these experiments, showing continuous polymerase activity for more than 1 hour and >4000 bases

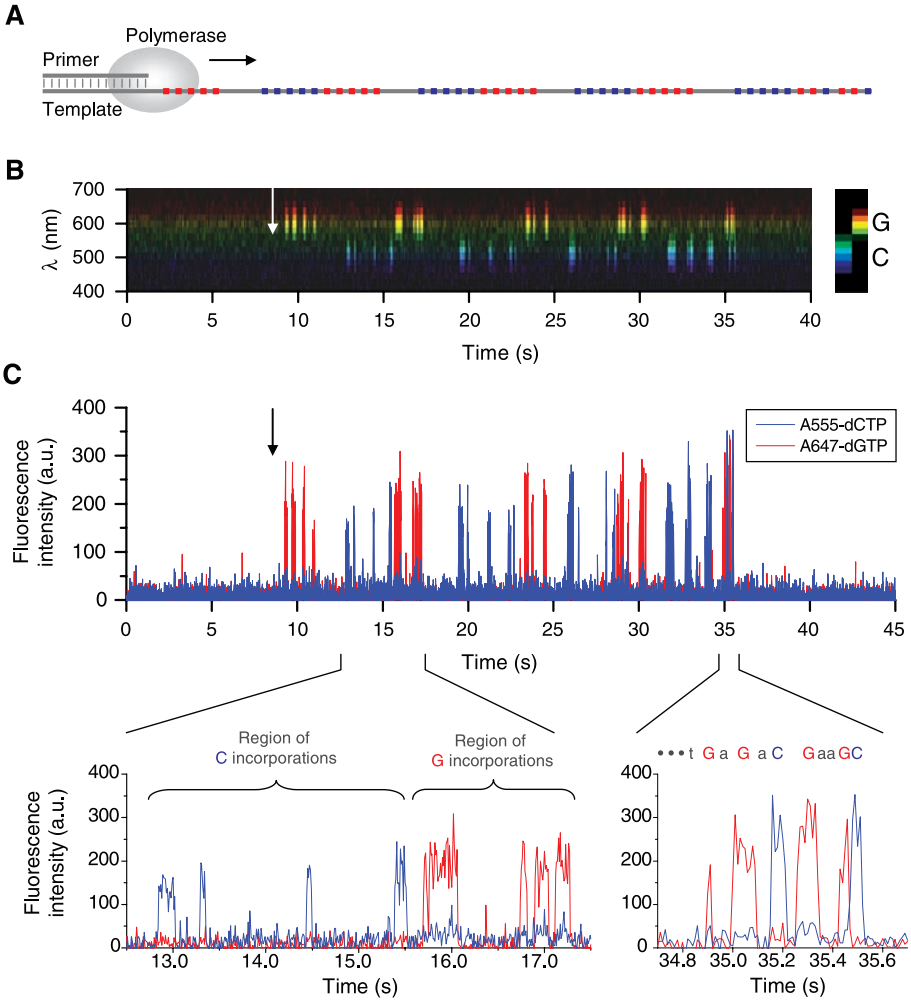


Fig. 2. Real-time detection of single-molecule DNA polymerase activity. (**A**) DNA template design for two-base sequence pattern detection. The sequence of a linear, single-stranded DNA template was designed to yield incorporation of alternating blocks of two phospholinked nucleotides (A555-dCTP and A647-dGTP), interspersed with the other two, unmodified dNTPs. (**B**) Time-resolved fluorescence intensity spectrum from a ZMW. Data from a 15 × 5 pixel area from each movie frame were spatially collapsed to a 15-pixel spectrum, which is shown as a function of time. The expected fluorescence emission profiles for the two labeled nucleotides are shown at the right. The arrow denotes addition of the catalytic metal ion that initiated the polymerization reaction. The complete data set from which the time trace was extracted, containing 3000 ZMWs measured simultaneously, is shown in movie S1. (**C**) Corresponding fluorescence time trace after spectral processing. Two regions are magnified and annotated with the expected nucleotide incorporation sequence. Pulse heights show level setting with a coefficient of variation of 27% and a maximal excursion of 61%.

Downloaded from www.sciencemag.org on January 4, 2009

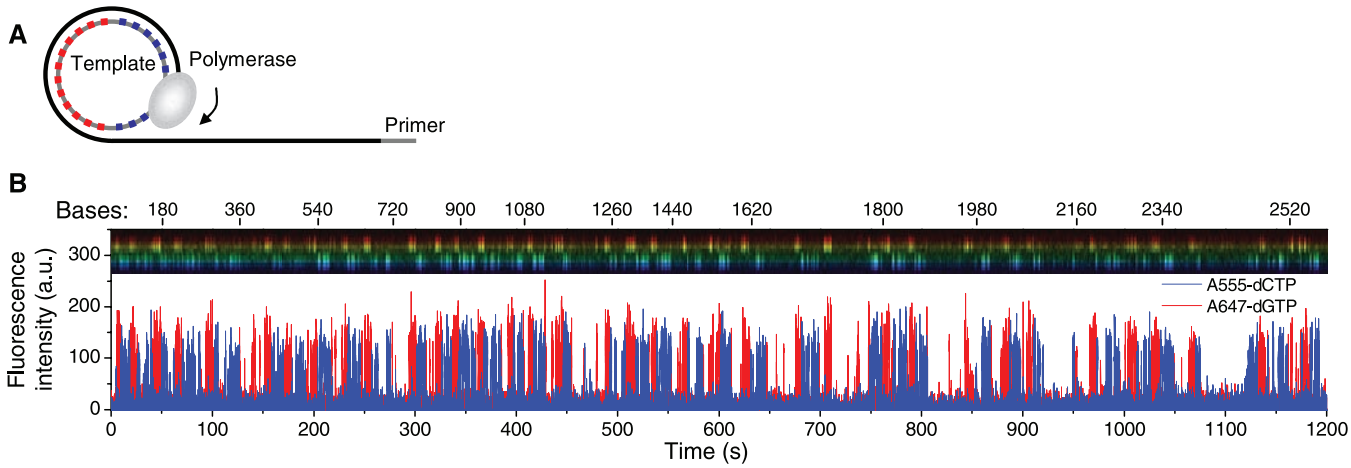
synthesized. The mean number of pulses per block was uniform over at least 1000 bases of incorporation (Fig. 3D); hence, this sequencing approach maintains accuracy irrespective of read length.

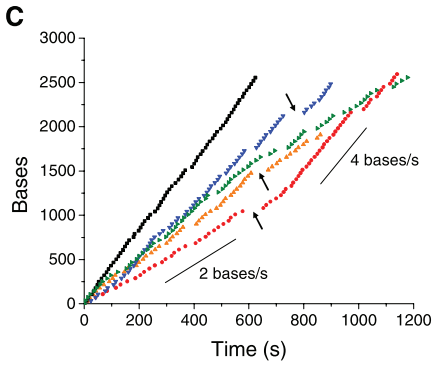
The measurements described above were extended to four-color DNA sequencing. All four native nucleotides were fully replaced with the following set of phospholinked dNTPs: A555-dATP, A568-dTTP, A647-dGTP, and A660-dCTP (fig. S4). Two lasers were used for the excitation of the four fluorophores. Fluorescence pulses were identified by a threshold detection algorithm (*37*) based on dye-weighted summation as above. The base identities of pulses were automatically assigned by least-squares fitting of the four phospholinked dNTP reference spectra to the measured spectra (fig. S4) (*26*). The read extracted from the measured pulses matched well with the underlying sequence of the nascent DNA strand, spanning the entire length of the 150-base linear template (Fig. 4, A and B). Of the 158 total bases in the alignment, 131 were correctly identified by the automated base caller. The 27 errors consisted of 12 deletions, eight insertions, and seven mismatches.

Sequencing performance analysis was extended to a set of 449 reads that showed pulse trains consistent with single polymerase occupancy (table S3). In these data, errors are dominated by deletions, which stem from incorporation events or intervals between them that are too short to be reliably detected. Unlabeled nucleotide contamination (dark nucleotides) can be a source of deletion errors in single-molecule sequencing systems. Here, this was not the case because the initial phospholinked dNTP composition was >99.5% pure (fig. S5) and, unlike with base-linked nucleotides, the polymerase showed no preference for unlabeled versus labeled substrates (*32*). Additionally, a comparison of our observed deletion error rate with a deletion rate predicted solely from pulse width distributions shows that dark nucleotides need not be invoked as a source of error. For example, fig. S6 shows the pulse width distribution for A555-dATP and the projected probability of pulse detection for that nucleotide as a function of pulse width. From these data, the deletion rate is estimated to be 7.8%, consistent with the observed 7.4% deletion rate for this nucleotide. This error type can be addressed by engineering the enzyme to reduce the fraction of short incorporation events, increasing fluorophore brightness, and improving efficiency of light collection.

The majority of insertion errors were caused by dissociation of a cognate nucleotide from the active site before phosphodiester bond formation can occur, resulting in the erroneous duplication of a pulse. This error type can be addressed by modifying the enzyme to decrease the free energy of the enzyme-substrate bound state, thus decreasing the dissociation rate before catalysis. Mismatches in the reads were mainly caused by spectral misassignments of the A647 and A660 dyes (accounting for ~60% of the mismatch error), which show the least spectral separation amongst the four dyes (table S3). The remainder of the mismatches involved misassignments between the A555 and A568 dyes (other factors were below the sensitivity of the assay). Finding compatible dye sets with larger spectral separations, as well as increasing the brightness of the dyes and collection efficiency of the instrument, will reduce the frequency of these errors.

To survey possible sequence context dependencies of these error types, we quantitated the two most important kinetic parameters—pulse width and interpulse duration—as a function of sequence position over the 150-base template. To extract these parameters for each template location, we associated individual pulses from the 449 reads with their sequence positions using a Smith-Waterman alignment algorithm (*38*). Pulse widths and interpulse durations are displayed as a function of sequence position in Fig. 4, C and D, respectively. The average pulse widths depend weakly on dNTP identity and show statistically significant but only moderate variation across template position. The average interpulse durations were typically between 200 and 700 ms, except for a few instances with much higher





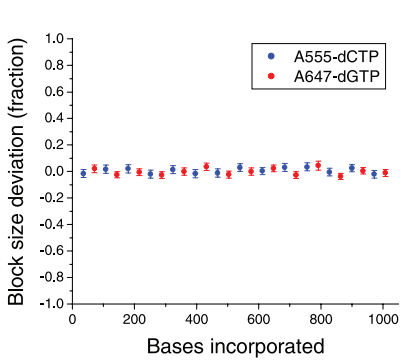


Fig. 3. Long read length activity of DNA polymerase. (**A**) DNA template design. The sequence of a circular, single-stranded template was designed to yield continuous incorporation via strand-displacement DNA synthesis of alternating blocks of two phospholinked nucleotides (A555-dCTP and A647-dGTP), interspersed with the other two unmodified dNTPs. (**B**) Time-resolved spectrum of fluorescence emission as in Fig. 2B with fluorescence time trace from a single ZMW. The corresponding total length of synthesized DNA is indicated by the top axis. (**C**) DNA polymerization rate profiles for several molecules. Examples of pause sites are indicated by arrows. The two lines indicate two persistent polymerization rates. (**D**) Error as a function of length of read for 14 rolling circle cycles (1008 total base incorporations; n = 186 reads). The fractional deviation from the average number of pulses per block (12 A555-dCTP and 12 A647-dGTP observed phospholinked dNTP pulses per cycle, respectively), mean ± SE, is plotted as a function of template position. The 95% confidence interval for the slope is –0.027 to +0.036 blocks per 1008 bases of incorporation.

Downloaded from www.sciencemag.org on January 4, 2009

values. These pause sites corresponded to regions with predicted stable secondary structure in the template and matched well with bulk capillary electrophoresis data (fig. S7). The major pause point seen at position 40 did not result in an increased frequency of dissociation events. The enzymatic rate of incorporation increased immediately after passing through the putative hairpin for experiments performed at 100 nM dNTP (from

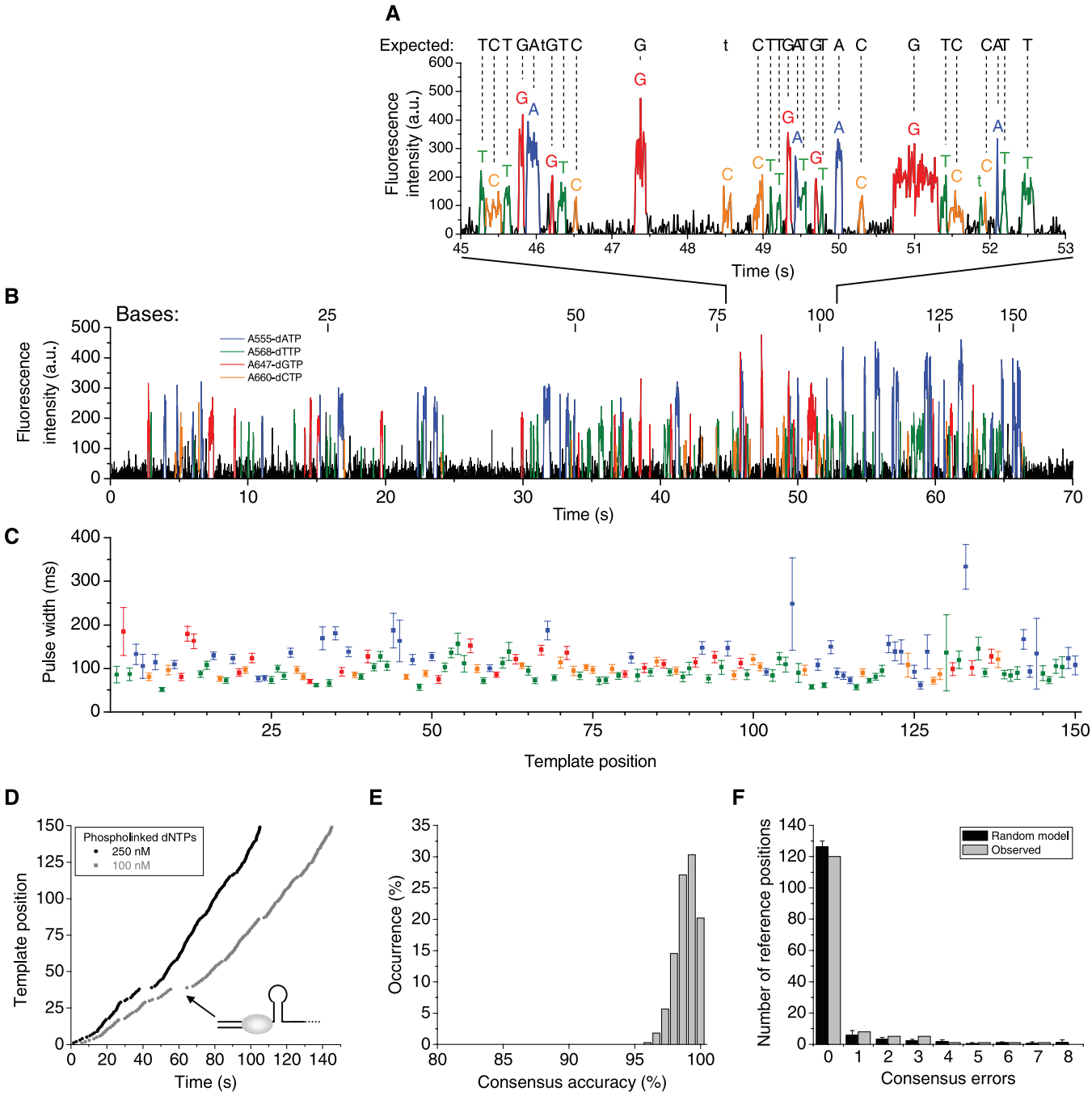


Fig. 4. Single-molecule, real-time, four-color DNA sequencing. (**A**) Total intensity output of all four dye-weighted channels, with pulses colored corresponding to the least-squares fitting decisions of the algorithm. This section of a fluorescence time trace shows 28 bases of incorporations and three errors. The expected template sequence is shown above, with dashed lines corresponding to matches; errors are in lowercase. (**B**) The entire read that proceeds through all 150 bases of the linear template. On average, ~63% of reads proceeded through the entire length of the DNA template. (**C**) Average pulse width as a function of template position (extracted from n = 449 reads). (**D**) Cumulative interpulse duration plotted as a function of template position for two different phospholinked dNTP concentrations (250 nM, n = 449 reads; 100 nM, n = 868 reads). The arrow indicates a pause site observed for both conditions at position 40, corresponding to predicted secondary structure in the template at position 46 (fig. S7), taking into account the enzyme's footprint on the template (*42*). (**E**) Histogram of the sequence accuracy of 100 consensus sequences created by subsampling from 449 single-molecule reads to 15-fold average coverage. The median accuracy of the distribution is 99.3%. (**F**) Observed systematic bias compared with prediction from a random model free of sequence context bias. The error frequencies for observed (gray bars) and bias-free model data (black bars) are plotted in a histogram with the number of errors on the x axis and the number of different reference positions showing this many errors in 100 trials on the y axis. The random model is based on the observed error frequencies (table S3) (*26*).

Downloaded from www.sciencemag.org on January 4, 2009

0.7 to 1.25 bases/s) and at 250 nM dNTP (from 1.1 to 1.5 bases/s). This increased rate resulted from a decrease in interpulse duration; the pulse widths remained nearly constant. It is not surprising that the interpulse durations, which encompass motion of the polymerase relative to the DNA template, would be strongly affected by DNA secondary structure, whereas variations in the pulse widths, which are governed by local chemical processes in the active site, are less affected.

Pulse widths showed only moderate variability with sequence context, and the interpulse durations, although highly dependent on secondary structure, always produced average values above 200 ms. Thus, sequence errors in individual reads should be predominantly uncorrelated and amenable to molecular ensemble averaging. To test this hypothesis, we formed 100 consensus sequences with reads randomly subsampled from the data set to yield 15-molecule coverage, using the center-star algorithm (*39*). The median accuracy over this set of sequences was 99.3%, with a distribution of values shown in Fig. 4E. The consensus accuracy as a function of fold coverage is shown in fig. S8. To explore the possibility of systematic error beyond the fluorophore-dependent error rates (table S3), we analyzed the dependence of consensus error frequency on sequence context via the distribution of the number of times out of the 100 trials that each reference sequence position was reported incorrectly (Fig. 4F) (*26*). This histogram is in agreement with a context bias–free random model, showing that within the sensitivity of this study there were no other biophysical sources of systematic error.

The systematic variations in pulse width and interpulse duration seen in Fig. 4 do not interfere with the development of accurate consensus sequence. In fact, such variations constitute an additional signal that is dependent on DNA primary and secondary structure that can be exploited to increase the accuracy of the consensus. Another appealing feature of this sequencing approach is that, through the strand-displacing capability of the polymerase (demonstrated in Fig. 3), closed circular templates can be sequenced multiple times by a DNA polymerase in a single run. This allows determination of a circular consensus sequence using only one DNA molecule. The resulting insensitivity to sample heterogeneity will greatly improve detection of rare mutations. This single-molecule aspect also enables simplified sample preparation and minimizes reagent consumption because only small amounts of genomic DNA are required.

In addition to the sequence, the real-time aspect of our approach generates unprecedented information about DNA polymerase kinetics that will allow other uses of the technology. Because the system reports the kinetics of every base incorporation through the pulse width and the interpulse duration, the system can be used today to investigate kinetics of DNA polymerization with unprecedented resolution and speed, providing the distribution of kinetic parameters over hundreds of different sequence contexts in a single 5-min experiment. Because polymerase kinetics is sensitive to biological perturbation, our approach would allow investigation of DNA binding proteins, DNA polymerase inhibitors, and the effects of base methylation.

Commercially available high-throughput sequencing systems that rely on stepwise flushing of a solid support with reactants and subsequent scanning to read out a single base currently operate in the regime of ~1 hour per base sequenced (*13*, *14*, *16*). This low rate of sequence production is compensated by high multiplex levels (~10^6 to 10^8). The single-molecule real-time DNA sequencing approach demonstrated here represents an increase in the speed of the underlying sequencing cycle by approximately four orders of magnitude. Stepwise sequencing systems are characterized by relatively short read lengths because of the deleterious effects of interrupting enzyme activity. Exploiting uninterrupted DNA synthesis will enable sequence reads thousands of bases in length.

We have shown that with just 15 molecules, a consensus sequence with 99.3% median accuracy can be formed with no detectable sequence context bias and a uniform error profile within reads. The present level of accuracy can produce alignment and consensus adequate for resequencing applications. However, it would create challenges for de novo assembly or alignment into highly repetitive DNA. The accuracy of the system could be enhanced by improvements in enzyme kinetics. Reducing the free energy of the nucleotide-bound state through polymerase mutation and nucleotide modification would reduce the occurrence of cognate nucleotide dissociation and the attendant insertion errors. Lowering the rate of phosphodiester bond formation would lengthen the pulses, reducing the incidence of deletion errors. Deletions could also be reduced through increases in fluorophore brightness and system optical collection efficiency. Finally, circular consensus sequencing can be used to eliminate stochastic errors in single-molecule sequencing.

The limited experimental multiplex used here could be applied to sequencing small viral and bacterial genomes. Given that each ZMW is capable of producing sequence at a rate greater than 400 kb per day, just 14,000 functioning ZMWs are required to produce a raw read throughput equivalent to 1-fold coverage of a diploid human genome per day. This number is attainable using optics and detector technology available today. Even larger numbers of ZMWs could be simultaneously monitored using multi-megapixel charge-coupled device or complementary metal-oxide semiconductor cameras expected within five years (*40*, *41*). As these technologies evolve, it will be possible to provide later generations of this instrument with multiplex commensurate with current stepwise sequencing systems. Combining this level of multiplex with the high intrinsic speed and read length of single-molecule, real-time DNA sequencing will enable low-cost rapid genome sequencing.

References and Notes

1. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
2. L. Blanco *et al.*, *J. Biol. Chem.* **264**, 8935 (1989).
3. A. Kornberg, T. A. Baker, *DNA Replication* (Freeman, New York, ed. 2, 1992).
4. S. Tabor, H. E. Huber, C. C. Richardson, *J. Biol. Chem.* **262**, 16212 (1987).
5. C. Feschotte, E. J. Pritham, *Annu. Rev. Genet.* **41**, 331 (2007).
6. E. Tuzun *et al.*, *Nat. Genet.* **37**, 727 (2005).
7. S. Balasubramanian, D. R. Bentley, Patent WO 01/057248 (2001).
8. I. Braslavsky, B. Hebert, E. Kartalov, S. R. Quake, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3960 (2003).
9. M. Ronaghi, M. Uhlen, P. Nyren, *Science* **281**, 363 (1998).
10. J. Shendure *et al.*, *Science* **309**, 1728 (2005).
11. D. R. Bentley, *Curr. Opin. Genet. Dev.* **16**, 545 (2006).
12. M. L. Metzker, *Genome Res.* **15**, 1767 (2005).
13. J. B. Fan *et al.*, *Methods Enzymol.* **410**, 57 (2006).
14. T. D. Harris *et al.*, *Science* **320**, 106 (2008).
15. M. Margulies *et al.*, *Nature* **437**, 376 (2005).
16. A. Valouev *et al.*, *Genome Res.* **18**, 1051 (2008).
17. E. Y. Chan, U.S. Patent 6,210,896 (2001).
18. S. L. Cockroft, J. Chu, M. Amorin, M. R. Ghadiri, *J. Am. Chem. Soc.* **130**, 818 (2008).
19. W. J. Greenleaf, S. M. Block, *Science* **313**, 801 (2006).
20. M. J. Levene *et al.*, *Science* **299**, 682 (2003).
21. B. A. Mulder *et al.*, *Nucleic Acids Res.* **33**, 4865 (2005).
22. B. Reynolds, R. Miller, J. G. Williams, J. P. Anderson, *Nucleosides Nucleotides Nucleic Acids* **27**, 18 (2008).
23. M. Foquet *et al.*, *J. Appl. Phys.* **103**, 034301 (2008).
24. M. J. Lang, P. M. Fordyce, A. M. Engh, K. C. Neuman, S. M. Block, *Nat. Methods* **1**, 133 (2004).
25. A. M. Lieto, R. C. Cush, N. L. Thompson, *Biophys. J.* **85**, 3294 (2003).
26. See supporting material on *Science* Online.
27. J. Ju *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19635 (2006).
28. R. D. Mitra, J. Shendure, J. Olejnik, O. Edyta Krzymanska, G. M. Church, *Anal. Biochem.* **320**, 55 (2003).
29. C. C. Kao, T. Widlanski, W. Vassiliou, J. Epp, U.S. Patent 6,399,335 (2002).
30. S. Kumar *et al.*, *Nucleosides Nucleotides Nucleic Acids* **24**, 401 (2005).
31. A. Sood *et al.*, *J. Am. Chem. Soc.* **127**, 2394 (2005).
32. J. Korlach *et al.*, *Nucleosides Nucleotides Nucleic Acids* **27**, 1072 (2008).
33. F. B. Dean *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 5261 (2002).
34. J. Korlach *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1176 (2008).
35. P. M. Lundquist *et al.*, *Opt. Lett.* **33**, 1026 (2008).
36. C. Castro *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 4267 (2007).
37. K. Horne, *Publ. Astron. Soc. Pac.* **98**, 609 (1986).
38. O. Gotoh, *J. Mol. Biol.* **162**, 705 (1982).
39. D. Gusfield, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology* (Cambridge Univ. Press, Cambridge, 1997).
40. J. T. Bosiers *et al.*, *Proc. SPIE* **6996**, 69960Z (2008).
41. A. J. P. Theuwissen, *Solid State Electron.* **52**, 1401 (2008).
42. A. J. Berman *et al.*, *EMBO J.* **26**, 3494 (2007).

43. We thank the entire staff at Pacific Biosciences, and J. Puglisi, M. Hunkapiller, R. Kornberg, K. Johnson, D. Haussler, W. Webb, and H. Craighead for many helpful discussions. Supported by National Human Genome Research Institute grant R01HG003710.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1162986/DC1
Materials and Methods
Figs. S1 to S8
Tables S1 to S3
Movie S1
References

9 July 2008; accepted 20 October 2008
Published online 20 November 2008;
10.1126/science.1162986
Include this information when citing this paper.

Downloaded from www.sciencemag.org on January 4, 2009