***The following resources related to this article are available online at www.sciencemag.org
(this information is current as of January 10, 2010):***

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/320/5872/106

Supporting Online Material can be found at:
http://www.sciencemag.org/cgi/content/full/320/5872/106/DC1

This article **cites 14 articles**, 8 of which can be accessed for free:
http://www.sciencemag.org/cgi/content/full/320/5872/106#otherarticles

This article has been **cited by** 56 article(s) on the ISI Web of Science.

This article has been **cited by** 19 articles hosted by HighWire Press; see:
http://www.sciencemag.org/cgi/content/full/320/5872/106#otherarticles

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

Although the association of metabolic enzymes has been claimed for enzymes involved in the glycolytic pathway (*21*, *22*), to our knowledge this complex, the glycosome, has not been identified in living mammalian cells. By analogy, the present clusters observed in the de novo purine biosynthetic pathway may constitute a "purinosome." The formation of the purinosome appears to be dynamically regulated by stimulation of de novo purine biosynthesis in response to changes in purine levels. The purinosome may be a general phenomenon in all cell types during specific stages of the cell cycle, along with posttranslation modifications. Because of the relevance of de novo purine biosynthesis to human diseases, the purinosome may represent a new pharmacological opportunity for therapeutic intervention.

References and Notes

1. C. Stasolla, R. Katahira, T. A. Thorpe, H. Ashihara, *J. Plant Physiol.* **160**, 1271 (2003).
2. R. Zrenner, M. Stitt, U. Sonnewald, R. Boldt, *Annu. Rev. Plant Biol.* **57**, 805 (2006).
3. C. Z. Constantine *et al.*, *Biochemistry* **45**, 8193 (2006).
4. C. A. Caperelli, P. A. Benkovic, G. Chettur, S. J. Benkovic, *J. Biol. Chem.* **255**, 1885 (1980).
5. G. K. Smith, W. T. Mueller, G. F. Wasserman, W. D. Taylor, S. J. Benkovic, *Biochemistry* **19**, 4313 (1980).
6. J. Rudolph, J. Stubbe, *Biochemistry* **34**, 2241 (1995).
7. L. T. Gooljarsingh, J. Ramcharan, S. Gilroy, S. J. Benkovic, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 6565 (2001).
8. M. A. Becker, M. Kim, *J. Biol. Chem.* **262**, 14531 (1987).
9. T. Yamaoka *et al.*, *J. Biol. Chem.* **272**, 17719 (1997).
10. T. Yamaoka *et al.*, *J. Biol. Chem.* **276**, 21285 (2001).
11. Materials and methods are available as supporting material on *Science* Online.
12. G. F. Wasserman, W. T. Mueller, S. J. Benkovic, W. S. Liao, J. Taylor, *Biochemistry* **23**, 6704 (1984).
13. M. Howarth, K. Takao, Y. Hayashi, A. Y. Ting, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7583 (2005).
14. M. Kondo *et al.*, *J. Biochem. (Tokyo)* **128**, 57 (2000).
15. H. Iwahana *et al.*, *Biochim. Biophys. Acta* **1261**, 369 (1995).
16. L. F. Thompson, R. C. Willis, J. W. Stoop, J. E. Seegmiller, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 3722 (1978).
17. M. A. Becker, M. J. Losman, M. Kim, *J. Biol. Chem.* **262**, 5596 (1987).
18. S. X. Li *et al.*, *J. Mol. Biol.* **366**, 1603 (2007).
19. L. Pagliaro, D. L. Taylor, *J. Cell Biol.* **107**, 981 (1988).
20. M. Pozuelo Rubio *et al.*, *Biochem. J.* **379**, 395 (2004).
21. M. E. Campanella, H. Chu, P. S. Low, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2402 (2005).
22. F. R. Opperdoes, P. Borst, *FEBS Lett.* **80**, 360 (1977).
23. We thank M. Kyoung (Pennsylvania State University) for valuable suggestions on data collection, imaging analysis, and statistics; E. Kunze (Pennsylvania State University) for assistance with confocal laser-scanning microscopy; A. A. Heikal (Pennsylvania State University) for sharing HTB-125 and -126 cell lines; and R. Y. Tsien (University of California, San Diego) for providing the pRSET_mOrange plasmid as a gift. This work was supported, in part, by Pennsylvania State University and the Center for Optical Technologies (E.D.S.). Additional acknowledgment is made to the Donors of the American Chemical Society Petroleum Research Fund, for partial support of this research (E.D.S.).

Single-Molecule DNA Sequencing of a Viral Genome

Timothy D. Harris,[1]* Phillip R. Buzby,[1] Hazen Babcock,[1] Eric Beer,[1] Jayson Bowers,[1] Ido Braslavsky,[2] Marie Causey,[1] Jennifer Colonell,[1] James DiMeo,[1] J. William Efcavitch,[1] Eldar Giladi,[1] Jaime Gill,[1] John Healy,[1] Mirna Jarosz,[1] Dan Lapen,[1] Keith Moulton,[1] Stephen R. Quake,[3] Kathleen Steinmann,[1] Edward Thayer,[1] Anastasia Tyurina,[1] Rebecca Ward,[1] Howard Weiss,[1] Zheng Xie[1]

The full promise of human genomics will be realized only when the genomes of thousands of individuals can be sequenced for comparative analysis. A reference sequence enables the use of short read length. We report an amplification-free method for determining the nucleotide sequence of more than 280,000 individual DNA molecules simultaneously. A DNA polymerase adds labeled nucleotides to surface-immobilized primer-template duplexes in stepwise fashion, and the asynchronous growth of individual DNA molecules was monitored by fluorescence imaging. Read lengths of >25 bases and equivalent phred software program quality scores approaching 30 were achieved. We used this method to sequence the M13 virus to an average depth of >150× and with 100% coverage; thus, we resequenced the M13 genome with high-sensitivity mutation detection. This demonstrates a strategy for high-throughput low-cost resequencing.

DNA sequencing and the attendant genetic manipulation it enables have fundamentally altered life science, with the completion of the human genome sequence as a major milestone of this work (*1*, *2*). However, large sample sets—thousands of genomes—are required to analyze many phenomena in which genetics plays a role. With current sequencing technologies, the cost and complexity of such experiments remains limiting (*3*). Having the consensus human genome sequence in hand fundamentally changes the technology requirements for resequencing human genomes. In particular, one can use low-cost techniques with much shorter read lengths and higher parallelism than found with the Sanger capillary electrophoresis methods used to generate the reference genome (*4*).

Several recent reports emphasize the progress in short-read sequencing strategies (*5–8*). Although those methods have been used successfully to sequence microbial genomes, their current cost of sequencing, the complexity associated with DNA library preparation, and their use of polymerase chain reaction (PCR) amplification may limit broad application to human genome resequencing. The use of PCR is problematic for three reasons. First, because amplification efficiencies vary as a function of template properties, PCR introduces an uncontrolled bias in template representation. Second, short-read techniques require many more templates than conventional sequencing, and the in vitro manipulations to create libraries with defined sequences at the ends of templates are onerous and expensive in terms of DNA manipulation. Third, errors can be

introduced; in a recent large-scale cancer resequencing effort, PCR errors alone accounted for about one-third of initially detected "mutations" (*3*). The fidelity of PCR polymerases is widely reported at 0.5 to 1.0×10^{-4} (*9*), a substantial error rate for amplification of single-molecule targets. These limitations can be ameliorated by single-molecule sequencing approaches.

Single-molecule sequencing was proposed as early as 1989 (*10*). Recent work, however, has demonstrated the feasibility of single-molecule sequencing using DNA polymerase to sequence by synthesis (*11*), and a subsequent study of single–RNA polymerase activity shows DNA sequence can be inferred from the serial observation of four identical single-molecule templates (*12*). We have used single-molecule DNA sequencing to resequence the M13 phage genome (*13*). Our sequencing-by-synthesis scheme is diagrammed in Fig. 1. The library preparation process is simple and fast and does not require the use of PCR; it results in single-stranded, poly(dA)-tailed templates. Poly(dT) oligonucleotides are covalently anchored to glass cover slips at random positions. These oligomers are first used to capture the template strands, and then either as a primer for the template-directed primer extension that forms the basis of the sequence reading (Fig. 1) or, optionally, for a template replication step before sequencing (Fig. 2A). Up to 224 sequencing cycles are performed; each cycle consists of adding the polymerase and labeled nucleotide mixture (containing one of the four bases), rinsing, imaging multiple positions, and cleaving the dye labels. For the M13 data reported below, this sequencing process was performed simultaneously on more than 280,000 primer-template duplexes.

This single-molecule sequencing method allows a number of innovations that are not possible with bulk sequencing by synthesis (*5*, *8*). Most of these are related to the principle of asynchronous synthesis; that is, because each template molecule

[1]Helicos BioSciences Corporation, One Kendall Square, Cambridge, MA 02139, USA. [2]Department of Physics and Astronomy, Ohio University, Athens, OH 45701, USA. [3]Department of Bioengineering, Stanford University, and Howard Hughes Medical Institute, Stanford, CA 94305, USA.

*To whom correspondence should be addressed. E-mail: tharris@helicosbio.com
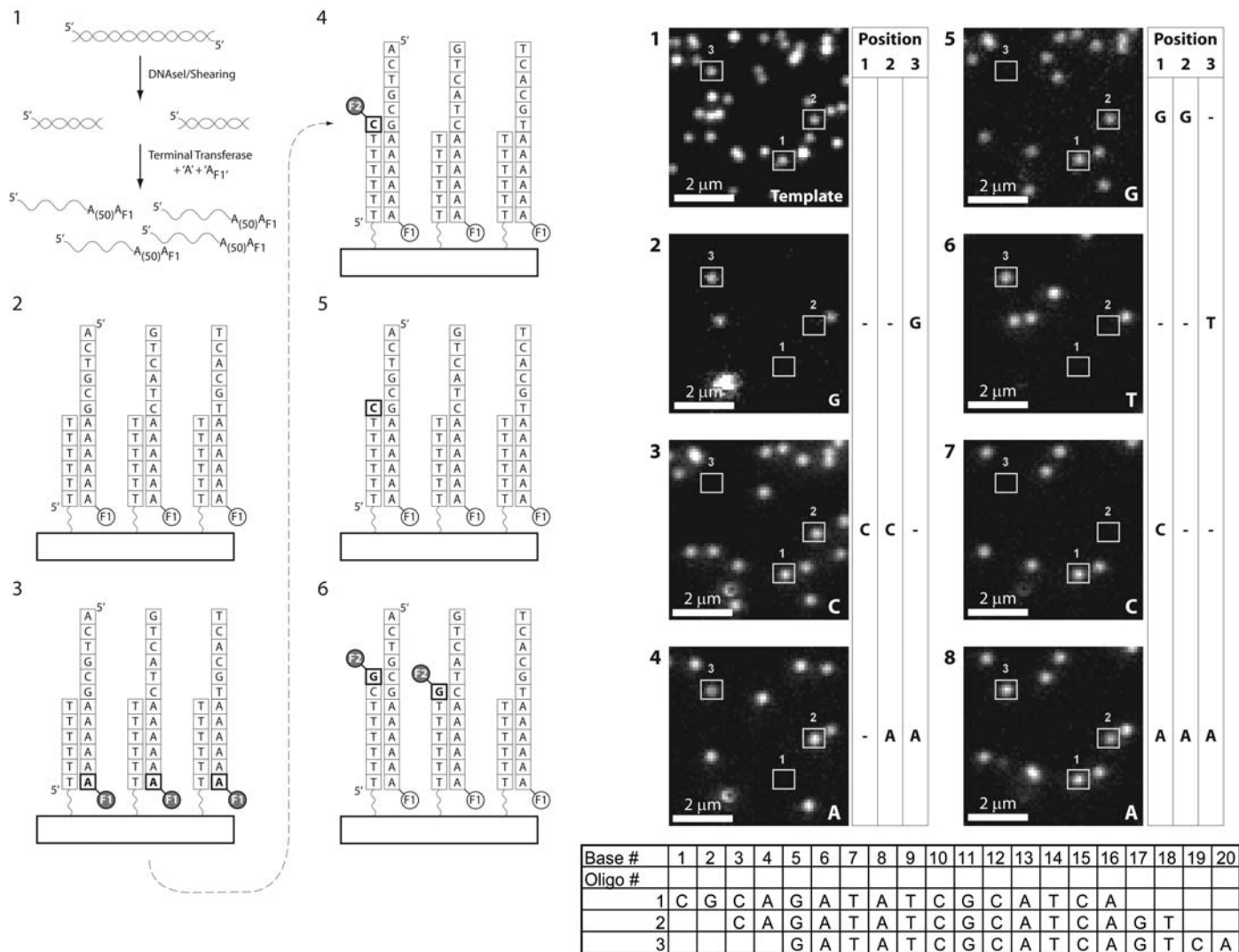
Base #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Oligo #																				
1	C	G	C	A	G	A	T	A	T	C	G	C	A	T	C	A				
2			C	A	G	A	T	A	T	C	G	C	A	T	C	A	G	T		
3					G	A	T	A	T	C	G	C	A	T	C	A	G	T	C	A

Fig. 1. Single-molecule sequencing sample preparation and imaging of single-nucleotide incorporation. (**Left**) Illustration of the single-molecule sequencing by synthesis process for single-pass sequencing. (**1**) Genomic DNA is prepared for sequencing by fragmentation and 3′ poly(A) tail addition, labeling, and blocking by terminal transferase. (**2**) Hybridization capture of these templates onto a surface with covalently bound 5′ "down" dT(50) oligonucleotide. (**3**) Imaging of the captured templates to establish sites for sequencing by synthesis. (**4**) Incubation of this surface with one labeled nucleotide and polymerase mixture, followed by rinsing of the synthesis mixture and direct imaging of the Cy5 labels exciting at 647 nm. (**5**) Chemical cleavage of the dye–nucleotide linker to release the dye label. (**6**) Addition of the next nucleotide and polymerase mixture. (**Right**) Image series illustrating template-specific base addition, successful rinsing, and successful linker cleavage. A mix of three templates is used to allow visual sequence assignment. Template complementary sequences are shown in the table (bottom). One example of each template is outlined in the figure. Each frame is a 6.6-μm square image of the same sample position, and shows ~35 of the 1.8×10^6 imaged templates in this experiment. Frame 1 is the image of the template labels. Template activity in three positions is shown in the columns to the right. Frame 2 is after the first synthesis and rinse cycle. Frames 3 to 8 show the effect of six more consecutive cleave, synthesis, and image cycles, using the base identity shown in the lower right corner of the frame.

is monitored individually, there is no need to keep each step of synthesis in phase. Thus, it is not necessary to drive each enzymatic incorporation step to completion. The principal benefit is that misincorporations are rare; their slow kinetics do not compete appreciably with the time to incorporate 80 to 90% of the correct base. A corollary is that this allows greater flexibility in the choice of reagents and synthesis chemistries, because the requirements on incorporation kinetics are relaxed. Asynchronicity is also used to facilitate reading of base-repeat sequences, homopolymers.

The use of a single-molecule method also enabled us to resequence each individual template

in situ, which greatly reduced the ensemble error rate. This "two-pass" sequencing process is illustrated in Fig. 2A. Captured oligonucleotide templates were copied using a high-fidelity polymerase to yield covalently attached templates with a distal

primer hybridization sequence. In the first pass, templates were primed and sequenced as described above (pass 1). The extended primers were then melted off using hot water, and the templates were primed again and sequenced a second time (pass 2).
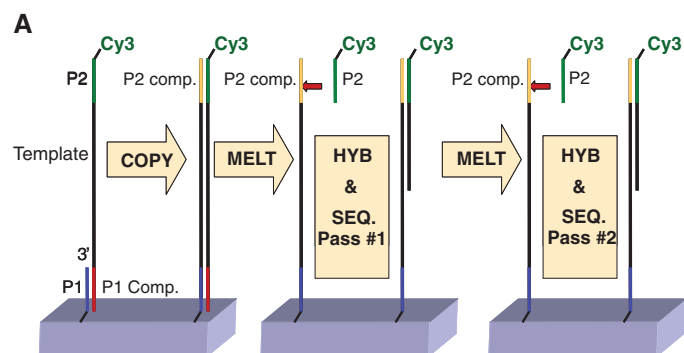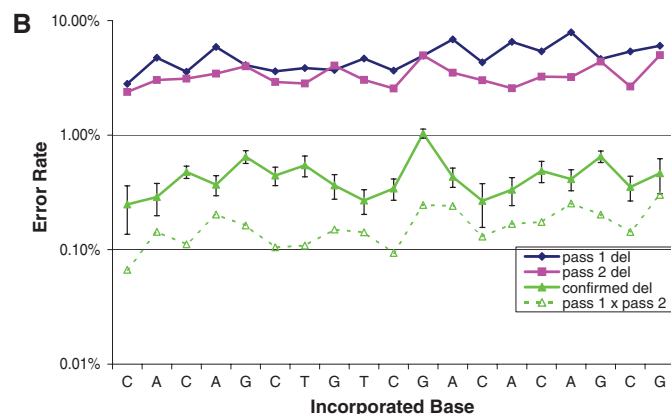
Table 1. M13 genome alignment statistics. The average read length was 23 bases, increasing to 27 bases after homopolymers were deconvolved.

Alignment	Coverage				Coverage (%)		No. of reads
	Average	Voting	Max.	Min.	>10×	>20×	
Forward	96×	75×	283×	3×	99.4	97.2	32,473
Reverse	105×	83×	301×	5×	99.5	95.8	34,109

Fig. 2. Methodology for two-pass sequencing and deletion statistics. (**A**) Process for two-pass sequencing. As with one pass, the DNA sample is hybridized to the surface capture oligomer. A copy of the DNA is made, which results in covalent anchoring of the template copy to the surface. Library construction is described in fig. S1 and provides for a primer site at the distal end of the template. This primer site is used for sequencing by synthesis, as before. After a sequencing run, the synthesized strand is melted off, and a new primer is hybridized, which allows the template to be sequenced a second time. (**B**) Average deletion rate for a single synthetic oligonucleotide, sequenced twice, from a data set of 35,000 reads; error bars show 3× the shot noise. See text and (13) for quantitative details.
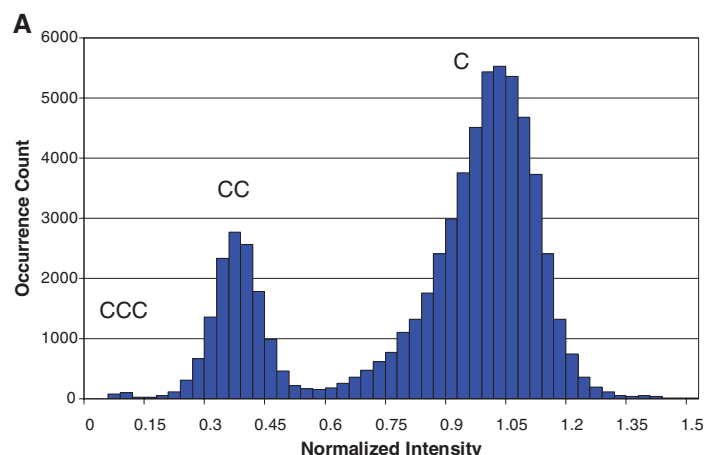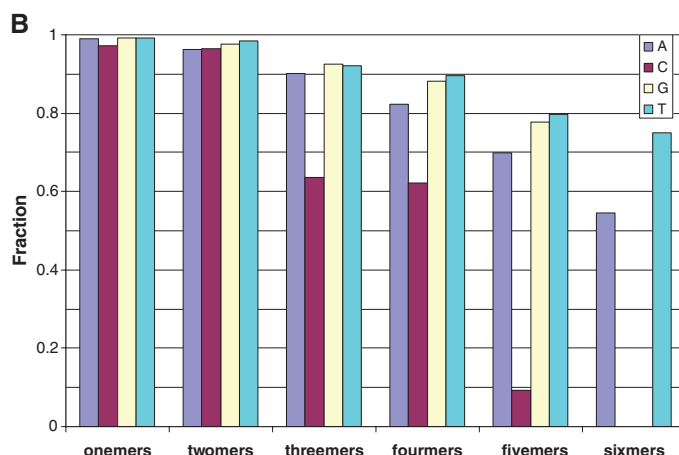

Fig. 3. Intensity calling for homopolymer length. (**A**) Histogram of observed intensities for C incorporation cycles across all C after A contexts in M13. By using known sequences and the incorporation cycle count required to observe the base following the poly(C) motif, the intensity of single, double, and triple incorporations can be determined. The group centered at normalized intensity ~1.0 is single incorporations; the group at normalized intensity 0.35 is incorporation of two labeled dCTP-Cy5 analogs in a single cycle; the small peak near normalized intensity 0.10 is incorporation of three labeled dCTP-Cy5 analogs in a single cycle. The magnitude of the quenching can be adjusted by addition of nonpolar solvents to the imaging medium, in this case acetonitrile. By using thresholds chosen from plots similar to those in (A) for all four bases, an intensity-based counting process can be used. (**B**) The accuracy for confirmed two-pass length calling for all M13 homopolymer motifs except for a single 8A/8T occurrence. Onemer, twomer, and so on is short for one-nucleotide oligomer, two-nucleotide oligomer, and so forth. All contexts have a correct majority votes, exceeding 80% up to four-nucleotide oligomers, for all contexts except 5C. The 5C case and the strategy to achieve minimal false-positive–mutation calling is discussed in detail in (13).

The combined or two-pass error rate was defined as follows [discussed in more detail in (13)]: The reads from both passes of each template were separately aligned to the oligonucleotide reference. Only reads that were in agreement on both passes were given a "vote" at that position; ~80% of the bases meet this criterion. The error rate is the ratio of votes disagreeing with the reference divided by the total number of templates considered. The dominant error was deletion. Deletion errors varied from 3 to 7% in pass 1 and 2 to 5% in pass 2; the confirmed (seen in both passes) deletion error rate range was 0.2 to 1.0% (Fig. 2B); the calculated product of the first- and second-pass deletion error rates varied from 0.1 to 0.3% and is shown for comparison in Fig. 2B as open triangles. The confirmed deletion rate is roughly the same magnitude as this product, the expected result for substantially random errors.

The ultimate lower limit for single-molecule single-pass error rates is not clear, but this two-pass process produces error rates low enough to assemble contigs if adequate read length is achieved (14). The equivalent phred software program quality of these single-molecule reads ranges from 20 to 28 (15).

To demonstrate the performance of single-molecule sequencing, we resequenced the M13 phage genome. A double-stranded M13 sample, prepared as shown in fig. S1, was sequenced for a total of 224 cycles; two passes with 112 cycles of synthesis were made in each pass; each pass was 28 "quad cycles" of successive CTAG additions. The average and median read lengths were ~23 bases for this run. Increasing the cycle count increases the average read length, and we have performed sequencing runs with average read lengths as high as 30. The forward and reverse genome coverage for the M13 data here averaged 96× and 105×, respectively. We aligned the data against the known M13 reference. The alignment statistics for this run are shown in Table 1.

It is a challenge for all sequencing by synthesis methods to detect base repeats, homopolymers. Because we operate the chemistry asynchronously, we can limit incorporations in base repeats to primarily two or three bases. For example, a template with a sequence segment TGGGAT may incorporate zero, one, two, or three C's in a single synthesis cycle when reading the GGG template sequence (see the statistics in fig. S2). We observed that fluorophores in multiple incorporations interact and thus yield reduced emission. Intensity distributions from all C after A incorporations in the M13 experiment described above (Fig. 3A) show that separation between single and double incorporations is
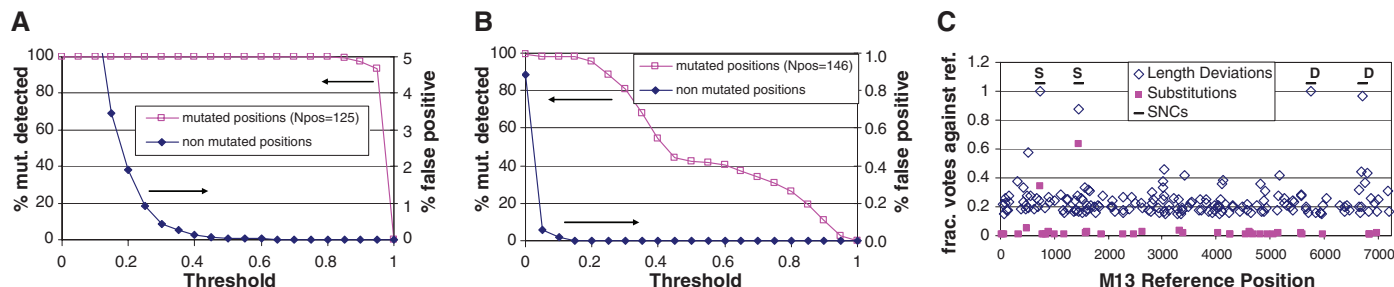
Fig. 4. Demonstration of mutation detection by alignment of experimental data against mutated M13 references. We show two single-nucleotide–change response curves. (**A**) Statistics for false-positive and mutation detection for insertions causing increase to homopolymer length. (**B**) Statistics for false-positive and mutation detection for substitutions creating all classes of sequence change. These curves show the fraction of positions in M13 that voted against the reference on both forward and reverse strands, as a function of the voting threshold. A vote against the reference is a mutation call. (**C**) A voting plot showing the results for all positions for the error types plotted in (A) and (B) against a reference with four mutations, two each for the mutation types in (A) and (B). Length mutations, those two points in the upper right, have a high false-positive rate (only votes > 0.15 are plotted for clarity) but a near 100% mutation detection efficiency (A). Substitutions have a much lower false-positive rate, all votes against the reference were plotted, but as reflected in (B), the result was lower mutation-detection efficiency. As seen in (C), substitutions are reported directly (red violet solid squares) and as length changes (open diamonds), upper left. S and D show positions for, respectively, substitution and deletion SNC changes to the reference. These curves demonstrate that it is possible to choose voting thresholds that enable successful mutation detection with very low false-positive rates. SNCs, single-nucleotide changes.

Table 2. Mutation detection via synthetic mutations in the M13 reference genome (hp, homopolymer). We tested 500 randomly chosen positions in 10 separately modified references; four of these positions had less than 10× coverage and were disqualified.

Mutation type	Number		Success (%)
	Tested	Found	
Insertion into hp	125	125	100
Deletion from hp	42	41	98
Deletion from non-hp	49	45	92
Deletion creating hp	34	33	97
Non-hp insertion	100	100	100
Substitution	146	143	98
Total mutations	496	487	98

very good and can be counted with only a small fraction of ambiguous calls. Triple incorporations are weak emitters, and a significant fraction fall below the system detection limit. Incorporation of more than three nucleotides is rare. Longer homopolymer runs were measured by adding the results of individual incorporation cycles, almost all made up of one-, two-, or three-base incorporations.

Normalized intensity thresholds delineating incorporations of one, two, or three nucleotides were determined for all four nucleotides by finding the minima between the distributions (Fig. 3A and fig. S4). For each base read in each strand, an incorporation count is generated on the basis of its intensity relative to the threshold. As with our accounting for deletion errors, we accept votes only from positions where the two passes agree on the length of a homopolymer position and ignore votes from positions where the passes disagree. The success rate for length calling for all positions in the M13 genome is shown in Fig. 3B; homopolymers range from two to six bases except for a single 8A/8T. For homopolymers with just two bases, the calling was accurate, with >95% of the homopolymer calls correct, but there were a significant fraction of incorrect length calls for longer homopolymers, particularly for C. We imposed a constraint that the called length must agree on the forward and reverse strands for our double-stranded sample, which requires sequence depth on both sense strands. Difficulties in calling C homopolymers were compensated on the complementary strand, because the corresponding positions are G homopolymers, which are called more accurately (Fig. 3B and fig. S5). For a single-stranded target, these C homopolymer length errors would result in a false-positive mutation call. The demonstrated homopolymer call accuracy is sufficient to achieve sensitive detection of mutations in the M13 genome, as described below.

To determine the M13 sequence quality, we explored how well mutations in the reference sequence could be detected. When single-nucleotide changes (6) are created in the reference genome, one expects to find poor agreement with the sequence read alignments at those positions, and they therefore serve as an unbiased test of the alignment quality and sensitivity. We aligned the data to 10 mutated M13 reference genomes, each with 50 sequence changes representing all classes of single-nucleotide change (insertions, deletions, and substitutions in varying contexts). Using these alignments, we measured the fraction of votes against the reference; locations where the aligned reads vote significantly against the reference are possible mutations. Single-nucleotide–change response curves indicated true-positive mutation detection and false-positive detection for various choices of vote thresholds (Fig. 4, A and B, and fig. S10). To score a positive, we stipulated that the read sequence must have above-threshold votes against the reference on both the forward and reverse strands. The curves show that it is possible to achieve excellent mutation detection with very low false-positives for every class of mutation. We found thresholds that gave zero false-positives and enabled discovery of more than 98% of all mutations (Table 2). The error rate and homopolymer run-through in the sequencing chemistry reported here do limit the mutation detection sensitivity—i.e., the thresholds need to be set low. Large genomes, heterogeneous samples, and genomic structural variations will likely require longer reads, reduced homopolymer run through, and enhanced alignment tools.

In summary, we report a method to sequence single molecules of genomic DNA. The consensus alignment of this sequence data is able to accurately recapitulate the M13 phage genome with 100% coverage, while demonstrating robust and efficient detection of all single base–mutation types. The simplicity of the methods described here, the freedom from cloning or amplification, and the low reagent volumes used to produce sequence from over 280,000 strands simultaneously opens a path to very high throughput sequencing.

References and Notes

1. E. S. Lander et al., Nature **409**, 860 (2001).
2. J. C. Venter et al., Science **291**, 1304 (2001).
3. T. Sjoblom et al., Science **314**, 268 (2006).
4. J. Shendure, R. D. Mitra, C. Varma, G. M. Church, Nat. Rev. Genet. **5**, 335 (2004).
5. M. Margulies et al., Nature **437**, 376 (2005).
6. J. Shendure et al., Science **309**, 1728 (2005).
7. L. W. Hillier et al., Nat. Methods **5**, 183 (2008).
8. J. Ju et al., Proc. Natl. Acad. Sci. U.S.A. **103**, 19635 (2006).
9. W. M. Barnes, Gene **112**, 29 (1992).
10. J. H. Jett et al., J. Biomol. Struct. Dyn. **7**, 301 (1989).
11. I. Braslavsky, B. Hebert, E. Kartalov, S. R. Quake, Proc. Natl. Acad. Sci. U.S.A. **100**, 3960 (2003).
12. W. J. Greenleaf, S. M. Block, Science **313**, 801 (2006).
13. Materials and methods are available as supporting material on Science Online.
14. R. Staden, Nucleic Acids Res. **6**, 2601 (1979).
15. B. Ewing, L. Hillier, M. C. Wendl, P. Green, Genome Res. **8**, 175 (1998).
16. We thank M. Mrksich, G. Church, D. Crothers, S. Chu, F. Romesberg, N. Afeyan, P. Milos, E. Lander, and S. Lapidus for many valuable discussions. This work was supported in part by NIH grant R01 HG004144-01. S.R.Q. and I.B. are paid consultants for and own equity in HelicosBioSciences Corp. S.R.Q. and I.B. are also inventors of a licensed patent on single-molecule sequencing that may yield royalty payments.

109