Published in final edited form as:
Nature. 2005 September 15; 437(7057): 376–380.

Genome Sequencing in Open Microfabricated High Density Picoliter Reactors

Marcel Margulies[1,*], **Michael Egholm**[1,*], **William E. Altman**[1], **Said Attiya**[1], **Joel S. Bader**[1], **Lisa A. Bemben**[1], **Jan Berka**[1], **Michael S. Braverman**[1], **Yi-Ju Chen**[1], **Zhoutao Chen**[1], **Scott B. Dewell**[1], **Lei Du**[1], **Joseph M. Fierro**[1], **Xavier V. Gomes**[1], **Brian C. Goodwin**[1], **Wen He**[1], **Scott Helgesen**[1], **Chun He Ho**[1], **Gerard P. Irzyk**[1], **Szilveszter C. Jando**[1], **Maria L.I. Alenquer**[1], **Thomas P. Jarvie**[1], **Kshama B. Jirage**[1], **Jong-Bum Kim**[1], **James R. Knight**[1], **Janna R. Lanza**[1], **John H. Leamon**[1], **Steven M. Lefkowitz**[1], **Ming Lei**[1], **Jing Li**[1], **Kenton L. Lohman**[1], **Hong Lu**[1], **Vinod B. Makhijani**[1], **Keith E. McDade**[1], **Michael P. McKenna**[1], **Eugene W. Myers**[3], **Elizabeth Nickerson**[1], **John R. Nobile**[1], **Ramona Plant**[1], **Bernard P. Puc**[1], **Michael T. Ronan**[1], **George T. Roth**[1], **Gary J. Sarkis**[1], **Jan Fredrik Simons**[1], **John W. Simpson**[1], **Maithreyan Srinivasan**[1], **Karrie R. Tartaro**[1], **Alexander Tomasz**[4], **Kari A. Vogt**[1], **Greg A. Volkmer**[1], **Shally H. Wang**[1], **Yong Wang**[1], **Michael P. Weiner**[2], **Pengguang Yu**[1], **Richard F. Begley**[1], and **Jonathan M. Rothberg**[1]

1 *454 Life Sciences Corp., 20 Commercial St., Branford, CT 06405, USA.*

2 *The Rothberg Institute For Childhood Diseases, 530 Whitfield St., Guilford, CT 06437, USA.*

3 *University of California, Berkeley, CA 94720, USA.*

4 *Laboratory of Microbiology, The Rockefeller University, New York, NY 10021, USA.*

Abstract

We describe a scalable, highly parallel sequencing system with raw throughput significantly greater than that of state-of-the-art capillary electrophoresis instruments. The apparatus uses a novel 60×60 mm^2 fibreoptic slide containing 1,600,000 individual wells and is able to sequence 25 million bases, at 99% or better accuracy (*phred* 20), in a 4 hour run. To provide sequencing templates, we clonally amplify DNA fragments on beads in the droplets of an emulsion. The template-carrying beads are loaded into the wells to convert each into a picoliter-scale sequencing reactor. We perform sequencing by synthesis using a pyrosequencing protocol optimized for solid support and the small dimension of the open reactors. Here we show the utility, throughput, accuracy and robustness of this system by shotgun sequencing and *de novo* assembling the *Mycoplasma genitalium* genome with 96% coverage at 99.96 % accuracy in one run of the machine.

DNA sequencing has dramatically changed the nature of biomedical research and medicine. Reductions in the cost, complexity and time required to sequence large amount of DNA, including improvements in the ability to sequence bacterial and eukaryotic genomes will have significant scientific, economic and cultural impact. Large scale sequencing projects, including whole genome sequencing, have usually required the cloning of DNA fragments into bacterial vectors, amplification and purification of individual templates, followed by Sanger sequencing [1] using fluorescent chain-terminating nucleotide analogues [2] and either slab gel or capillary electrophoresis. Current estimates put the cost of sequencing a human genome between $10

Correspondence and requests for materials should be addressed to J. M. Rothberg (jrothberg@454.com). Sequences for *M. genitalium* and *S. pneumoniae* were deposited at DDBJ/EMBL/GenBank under accession numbers AAGX01000000 and AAGY01000000 respectively..

[*] These authors contributed equally to the work.

Supplementary Information accompanies the paper on www.nature.com/nature.

and $25 million [3]. Alternative sequencing methods have been described [4, 5, 6, 7, 8] however, no technology has displaced the use of bacterial vectors and Sanger sequencing as the main generators of sequence information.

In this paper we describe an integrated system whose throughput routinely enables applications requiring millions of bases of sequence information, including whole genome sequencing Our focus has been on the co-development of an emulsion-based method [9, 10, 11] to isolate and amplify DNA fragments *in vitro*, and of a fabricated substrate and instrument that performs pyrophosphate-based sequencing (“pyrosequencing” 5, [12]) in picoliter-sized wells.

In a typical run we generate over 25 million bases with a *phred* 20 or better quality score (predicted to have an accuracy of 99% or higher). While this *phred* 20 quality throughput is significantly higher than that of Sanger sequencing by capillary electrophoresis, it is currently at the cost of substantially shorter reads and lower average individual read accuracy [13]. We further characterize the performance of the system, and demonstrate that it is possible to assemble bacterial genomes *de novo* from relatively short reads, by sequencing a known bacterial genome, *Mycoplasma genitalium* (580 kbp), and comparing our shotgun sequencing and *de novo* assembly with the results originally obtained for this genome [14]. The results of shotgun sequencing and *de novo* assembly of a larger bacterial genome, *Streptococcus pneumoniae*[15] (2.1 Mbp), are presented in Supplementary Table 4.

Emulsion based sample preparation

We generate random libraries of DNA fragments by shearing an entire genome and isolating single DNA molecules by limiting dilution (Supplementary Methods: Library Preparation). Specifically, we randomly fragment the entire genome, add specialized common adapters to the fragments, capture the individual fragments on their own beads and, within the droplets of an emulsion, clonally amplify the individual fragment (Figure 1A and 1B). Unlike in current sequencing technology, our approach does not require subcloning in bacteria or the handling of individual clones; the templates are handled in bulk within the emulsions [9, 10, 11].

Sequencing in fabricated picoliter sized reaction vessels

We perform sequencing by synthesis simultaneously in open wells of a fibreoptic slide using a modified pyrosequencing protocol that is designed to take advantage of the small scale of the wells. The fibreoptic slides are manufactured by slicing of a fibreoptic block that is obtained by repeated drawing and fusing of optic fibres. At each iteration, the diameters of the individual fibres decrease as they are hexagonally packed into bundles of increasing cross-sectional sizes. Each fibreoptic core is 44 μm in diameter and surrounded by 2–3 μm of cladding; etching of each core creates reaction wells approximately 55 μm in depth with a centre-to-centre distance of 50 μm (Figure 1C), resulting in a calculated well size of 75 pL and a well density of 480 wells/mm^2. The slide, containing approximately 1.6 million wells[16], is loaded with beads and mounted in a flow chamber designed to create a 300 μm high channel, above the well openings, through which the sequencing reagents flow (Figure 2, A and B). The unetched base of the slide is in optical contact with a second fibreoptic imaging bundle bonded to a CCD sensor, allowing the capture of emitted photons from the bottom of each individual well (Figure 2, C, and Supplementary Methods: Imaging System).

We developed a three-bead system, and optimized the components to achieve high efficiency on solid support. The combination of picoliter-sized wells, enzyme loading uniformity allowed by the small beads and enhanced solid support chemistry enabled us to develop a method that extends the useful read length of sequencing-by-synthesis to 100 bp (Supplementary Methods: Sequencing).

In the flow-chamber cyclically delivered reagents flow perpendicularly to the wells. This configuration allows simultaneous extension reactions on template carrying beads within the open wells and relies on convective and diffusive transport to control the addition or removal of reagents and by-products. The time scale for diffusion into and out of the wells is on the order of 10 seconds in the current configuration and is dependent on well depth and flow channel height. The time scales for the signal-generating enzymatic reactions are on the order of 0.02–1.5 seconds (Supplementary Methods: Interwell Diffusion). The current reaction is dominated by mass transport effects and improvements based on faster delivery of reagents are possible. Well depth was selected based on a number of competing requirements: (i) wells need to be deep enough for the DNA-carrying beads to remain in the wells in the presence of convective transport past the wells, (ii) they must be sufficiently deep to provide adequate isolation against diffusion of by-products from a well in which incorporation is taking place to a well where no incorporation is occurring, and (iii) they must be shallow enough to allow rapid diffusion of nucleotides into the wells, and rapid washing out of remaining nucleotides at the end of each flow cycle to enable high sequencing throughput and reduced reagent use. Following the flow of each nucleotide, a wash containing a nuclease is used to ensure that nucleotides do not remain in any well prior to the next nucleotide being introduced.

Base Calling of Individual Reads

Nucleotide incorporation is detected by the associated release of inorganic pyrophosphate (PPi) and the generation of photons [5, 12]. Wells containing template-carrying beads are identified by detecting a known four-nucleotide “key” sequence at the beginning of the read (Supplementary Methods: Image Processing). Raw signals are background-subtracted, normalized and corrected. The normalized signal intensity at each nucleotide flow, for a particular well, indicates the number of nucleotides, if any, that were incorporated. This linearity in signal is preserved to at least homopolymers of length 8 (Supplementary Figure 6). In sequencing by synthesis a very small number of templates on each bead lose synchronism (i.e. either get ahead of, or fall behind, all other templates in sequence [17]). The effect is primarily due to leftover nucleotides in a well (creating “carry forward”) or to incomplete extension. Typically, we observe a carry forward rate of 1–2% and an incomplete extension rate of 0.1–0.3%. Correction of these shifts is essential because the loss of synchronism is a cumulative effect that degrades the quality of sequencing at longer read lengths. We have developed algorithms, based on detailed models of the underlying physical phenomena, that allow us to determine, and correct for, the amounts of carry forward and incomplete extension occurring in individual wells (Supplementary Methods: Signal Processing). Figure 3 shows the processed result, a 113 bp long read generated in the *M. genitalium* run discussed below. To assess sequencing performance and the effectiveness of the correction algorithms, independently of artifacts introduced during the emulsion-based sample preparation, we created test fragments with difficult-to-sequence stretches of identical bases of increasing length (homopolymers) (Supplementary Methods: Test Fragments and Supplementary Figure 4). Using these test fragments, we have verified that at the individual read level we achieve base call accuracy of approximately 99.4%, at read lengths in excess of 100 bp (Table 1).

High Quality Reads and Consensus Accuracy.

Prior to base calling or aligning reads, we select high quality reads without relying on *a priori* knowledge of the genome or template being sequenced (Supplementary Methods: High Quality Reads). This selection is based on the observation that poor quality reads have a high proportion of signals that do not allow a clear distinction between a flow during which no nucleotide was incorporated and a flow during which one or more nucleotide was incorporated. When base calling individual reads, errors can occur because of signals that have ambiguous values (Supplementary Figure 5). To improve the usability of our reads, we also developed a metric which allows us to estimate *ab initio* the quality (or probability of correct base call) of each base of a read, analogous to the *phred* score [18] used by current Sanger sequencers (Supplementary Methods: Quality Scores and Supplementary Figure 8).

Higher quality sequence can be achieved by taking advantage of the high oversampling that our system affords and building a consensus sequence. Sequences are aligned to one another using the signal strengths at each nucleotide flow, rather than individual base calls, to determine optimal alignment (Supplementary Methods: Flow-space Mapping, Consensus Accuracy and Genome Coverage). The corresponding signals are then averaged, *after* which base calling is performed. This approach greatly improves the accuracy of the sequence (Supplementary Figure 7), and provides an estimate of the quality of the consensus base. We refer to that quality measure as the Z-score; it is a measure of the spread of signals in all the reads at *one* location and the distance between the average signal and the closest base calling threshold value. In both re-sequencing and *de novo* sequencing, as the minimum Z-score is raised the consensus accuracy increases, while coverage decreases; approximately half of the excluded bases, as the Z-score is increased, belong to homopolymers of length 4 and larger. Sanger sequencers usually require a depth of coverage at any base of three or more in order to achieve a consensus accuracy of 99.99%. To achieve a minimum of three fold coverage of 95% of the unique portions of a typical genome requires approximately 7 to 8 fold oversampling. Due to our higher error rate, we have observed that comparable consensus accuracies, over a similar fraction of a genome, are achieved with a depth of coverage of 4 or more, requiring approximately 10–12x oversampling.

Mycoplasma genitalium (580,069 bp).

*Mycoplasm*a genomic DNA was fragmented and prepared into a sequencing library as described above. (This was accomplished by a single individual in 4 hours.) Following emulsion PCR and bead deposition onto a 60×60 mm^2 fibreoptic slide, a process which took one individual 6 hours, 42 cycles of 4 nucleotides were flowed through the sequencing system in an automated 4 hour run of the instrument. The results are summarized in Table 2. In order to measure the quality of individual reads, we aligned each High Quality Read to the reference genome at 70% stringency, using flow-space mapping and criteria similar to those used previously in assessing the accuracy of other base callers [18]. When assessing sequencing quality, only reads that mapped to unique locations in the reference genome were included. Since this process excludes repeat regions (parts of the genome whose corresponding flowgrams are 70% similar to one another), the selected reads did not cover the genome completely. Figure 4A illustrates the distribution of read lengths for this run. The average read length was 110 bp, the resulting oversample 40 fold, and 84,011 reads (27.4%) were perfect. Figure 4B summarizes the average error as a function of base position. Coverage of non-repeat regions was consistent with the sample preparation and emulsion not being biased (Supplementary Figure 8). At the individual read level, we observe an insertion and deletion error rate of approximately 3.3%; substitution errors have a much lower rate, on the order of 0.5%. When using these reads without any Z-score restriction, we covered 99.94% of the genome in 10 contiguous regions with a consensus accuracy of 99.97%. The error rate in homopolymers is significantly reduced in the consensus sequence (Supplementary Figure 7). Of the bases not covered by this consensus sequence (366 bp), all belonged to excluded repeat regions. Setting a minimum Z-score equal to 4, coverage was reduced to 98.1% of the genome, while consensus accuracy increased to 99.996%. We further demonstrated the reproducibility of the system by repeating the whole genome sequencing of *M. genitalium* an additional 8 times, achieving a 40 fold coverage of the genome in each of the 8 separate instrument runs (Supplementary Table 3).

We assembled the *M. genitalium* reads from a single run into 25 contigs with an average length of 22.4 kbp. One of these contigs was misassembled due to a collapsed tandem repeat region of 60 bp, and was corrected by hand. The original sequencing of *M. genitalium* resulted in 28 contigs prior to directed sequencing used for finishing the sequence [14]. Our assembly covered 96.54% of the genome and attained a consensus accuracy of 99.96%. Non-resolvable repeat regions amount to 3% of the genome: we therefore covered 99.5% of the unique portions of the genome. Sixteen of the breaks between contigs were due to non-resolvable repeat regions, 2 were due to missed overlapping reads (our read filter and trimmer are not perfect and the algorithms we use to perform the pattern matching of flowgrams occasionally misses valid overlaps), and the remainder to thin read coverage. Setting a minimum Z-score of 4, coverage was reduced to 95.27% of the genome (98.2% of the resolvable part of the genome) with the consensus accuracy increasing to 99.994%.

Discussion

We have demonstrated in this paper the simultaneous acquisition of hundreds of thousands of sequence reads, 80–120 bases long, at 96% average accuracy, in a single run of the instrument using a newly developed *in vitro* sample preparation methodology, and sequencing technology. With *phred* 20 as a cutoff, we show that our instrument is able to produce over 47 million bases from test fragments and 25 million bases from genomic libraries. We used test fragments to decouple our sample preparation methodology from our sequencing technology. The decrease in single read accuracy from 99.4% for test fragments to 96% for genomic libraries is primarily due to a lack of clonality in a fraction of the genomic templates in the emulsion, and is not an inherent limitation of the sequencing technology. Most of the remaining errors result from a broadening of signal distributions, particularly for large homopolymers (7 or more), leading to ambiguous base calls. Recent work on the sequencing chemistry and algorithms that correct for crosstalk between wells suggests that the signal distributions will narrow, with an attendant reduction in errors and increase in read lengths. In preliminary experiments with genomic libraries that also includes improvements in the emulsion protocol, we are able to achieve, using 84 cycles, read lengths of 200 bp with accuracies similar to those demonstrated here for 100 bp. On occasion, at 168 cycles, we have generated individual reads which are 100% accurate over greater then 400 bp.

Using *M. genitalium*, we demonstrate that short fragments *a priori* do not prohibit the *de novo* assembly of bacterial genomes. In fact, the larger oversampling afforded by the throughput of our system resulted in a draft sequence having fewer contigs than with Sanger reads, with substantially less effort. By taking advantage of the oversampling, consensus accuracies greater then 99.96% were achieved for this genome. Further quality filtering the assembly, a consensus sequence can be selected with accuracy exceeding 99.99%, while incurring only a minor loss of genome coverage. Comparable results were seen when we shotgun sequenced and *de novo* assembled the 2.1 Mbp genome of *Streptococcus pneumoniae* [15] (Supplementary Table 4). The *de novo* assembly of genomes more complex than bacteria, including mammalian genomes, may require the development of methods, similar to those developed for Sanger sequencing, to prepare and sequence paired end libraries that can span repeats in these genome. To facilitate the use of paired end libraries we have developed methods to sequence, in an individual well, from both ends of genomic template, and plan to add paired end read capabilities to our assembler (Supplementary Methods: Double Ended Sequencing).

Future increases in throughput, and a concomitant reductions in cost per base, may come from the continued miniaturization of the fibreoptic reactors, allowing more sequence to be produced per unit area – a scaling characteristic similar to that which enabled the prediction of significant improvements in the integrated circuit at the start of its development cycle [19].

Methods

Emulsion based clonal amplification.

The simultaneous amplification of fragments is achieved by isolating individual DNA-carrying beads in separate ~100 μm aqueous droplets (on the order of 2×10^6/mL) made through the creation of a PCR-reaction-mixture-in-oil emulsion. (Figure 1B and Supplementary Methods: Preparation of DNA Capture Beads, Binding Template Species to DNA Capture Beads, PCR Reaction Mix Preparation and Formulation, Emulsification and Amplification). The droplets act as separate microreactors in which parallel DNA amplifications are performed, yielding approximately 10^7 copies of a template per bead; 800 μL of emulsion containing 1.5 million beads are prepared in a standard 2 mL tube. Each emulsion is aliquoted into 8 PCR tubes for amplification. After PCR, the emulsion is broken to release the beads, which include beads with amplified, immobilized DNA template, and empty beads (Supplementary Methods: Breaking the Emulsion and Recovery of Beads). We then enrich for template-carrying beads (Supplementary Methods: Enrichment of Beads). Typically, about 30% percent of the beads will have DNA, producing 450,000 template-carrying beads per emulsion reaction. The number of emulsions prepared depends on the size of the genome and the expected number of runs required to achieve adequate oversampling. The 580 kbp *M. genitalium* genome, sequenced on one 60×60 mm^2 fibreoptic slide, required 1.6 mL of emulsion. A human genome, oversampled 10 times, would require approximately 3000 mL of emulsion.

Bead loading into Picoliter Wells.

The enriched template-carrying beads are deposited by centrifugation into open wells (Figure 1C), arranged along one face of a 60×60 mm^2 fibreoptic slide. The beads (diameter ~ 28 μm) are sized to ensure that no more than one bead fits in most wells (we observed that 2–5% of filled wells contain more than one bead). Loading 450,000 beads (from one emulsion preparation) onto each half of a 60×60 mm^2 plate was experimentally found to limit bead occupancy to approximately 35% of all wells, thereby reducing chemical and optical crosstalk between wells. A mixture of smaller beads that carry immobilized ATP sulfurylase and luciferase necessary to generate light from free pyrophosphate are also loaded into the wells to create the individual sequencing reactors (Supplementary Methods: Bead Deposition, Preparation of Enzyme Beads and Micro-particle Fillers).

Image Capture.

A bead carrying 10 million copies of a template yields approximately 10,000 photons at the CCD sensor, per incorporated nucleotide. The generated light is transmitted through the base of the fibreoptic slide and detected by a large format CCD (4095×4096 pixels). The images are processed to yield sequence information simultaneously for all bead-template carrying wells. The imaging system was designed to accommodate a large number of small wells and the large number of optical signals being generated from individual wells during each nucleotide flow. Once mounted, the fibreoptic slide's position does not shift; this makes it possible for the image analysis software to determine the location of each well (whether or not it contains a DNA-carrying bead), based on light generation during the flow of a pyrophosphate solution which precedes each sequencing run. A single well is imaged by approximately nine 15 μm pixels. For each nucleotide flow, the light intensities collected by the pixels covering a particular well are summed to generate a signal for that particular well at that particular nucleotide flow. Each image captured by the CCD produces 32 megabytes of data. In order to perform all the necessary signal processing in real time, the control computer is fitted with an accessory board (Supplementary Methods: Field Programmable Arrays), hosting a 6 million gate FPGA [20, 21].

De novo Shotgun Sequence Assembler.

A *de novo* flow-space assembler was developed to capture all of the information contained in the original flow-based signal trace. It also addresses the fact that existing assemblers are not optimized for 80 to120 bp reads, particularly with respect to memory management due to the increased number of sequencing reads needed to achieve equivalent genome coverage. (A completely random genome covered with 100 bp reads requires approximately 50% more reads to yield the same number of contiguous regions (contigs) as achieved with 700 bp reads, assuming the need for a 30 bp overlap between reads.) [22]. This assembler consists of a series of modules: the *Overlapper*, which finds and creates overlaps between reads, the *Unitigger*, which constructs larger contigs of overlapping sequence reads, and the *Multialigner*, which generates consensus calls and quality scores for the bases within each contig (Supplementary Methods: De novo Sequence Assembler). (The names of the software modules are based on those performing related functions in other assemblers developed by Myers [23].)

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We acknowledge Peter Dacey and the support of the Operations groups of 454 Life Sciences. This research was supported in part by the US Department of Health and Human Services under NIH grants.

References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977;74:5463. [PubMed: 271968]
2. Prober JM, et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 1987;238:336. [PubMed: 2443975]
3. NIH News Release, October 14, 2004, http://www.genome.gov/12513210
4. Nyren P, Pettersson B, Uhlen M. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. Anal Biochem 1993;208:171. [PubMed: 8382019]
5. Ronaghi M, et al. Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 1996;242:84. [PubMed: 8923969]
6. Jacobson KB, et al. Applications of mass spectrometry to DNA sequencing. GATA 1991;8:223.
7. Bains W, Smith GC. A novel method for nucleic acid sequence determination. J Theor Biol 1988;135:303. [PubMed: 3256722]
8. Jett JH, et al. High-speed DNA sequencing: an approach based upon fluorescence detection of single molecules. Biomol Struct Dynamics 1989:301.
9. Tawfik DS, Griffiths AD. Man-made cell-like compartments for molecular evolution. Nat Biotechnology 1998;16:652.
10. Ghadessy FJ, Ong JL, Holliger P. Directed evolution of polymerase function by compartmentalized self-replication. Proc Nat Acad Sci USA 2001;98:4552. [PubMed: 11274352]
11. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Nat Acad Sci USA 2003;100:8817. [PubMed: 12857956]
12. Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. Science 1998;281:363. [PubMed: 9705713]
13. Current Sanger-based capillary electrophoresis sequencing systems produce up to 700 bp of sequence information from each of 96 DNA templates at an average read accuracy of 99.4% in one hour, or 67 thousand bases per hour, with substantially all of the bases having phred 20 or better quality. (Applied Biosystems 3730xl DNA Analyzer Specification Sheet, 2004.)
14. Fraser CM, et al. The minimal gene complement of Mycoplasma genitalium. Science 1995;270:397. [PubMed: 7569993]

15. Tettelin H, et al. Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. Science 2001;293:498. [PubMed: 11463916]

16. Leamon JH, et al. A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. Electrophoresis 2003;24:3769. [PubMed: 14613204]

17. Ronaghi M. Pyrosequencing sheds light on DNA sequencing. Genome Research 2001;11:3. [PubMed: 11156611]

18. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 1998;8:175. [PubMed: 9521921]

19. Moore GE. Cramming more components onto integrated circuits. Electronics April 19 1965;38 (Number 8)

20. Mehta K, Rajesh VA, Veeraswamy S. FPGA implementation of VXIbus interface hardware. Biomed Sci Instrum 1993;29:507. [PubMed: 8329634]

21. Fagin B, Watt JG, Gross R. A special-purpose processor for gene sequence analysis. Comput Appl Biosci 1996;9:221. [PubMed: 8481828]

22. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 1988;2:231. [PubMed: 3294162]

23. Myers EW. Toward simplifying and accurately formulating fragment assembly. J Comput Biol 1995;2:2751.

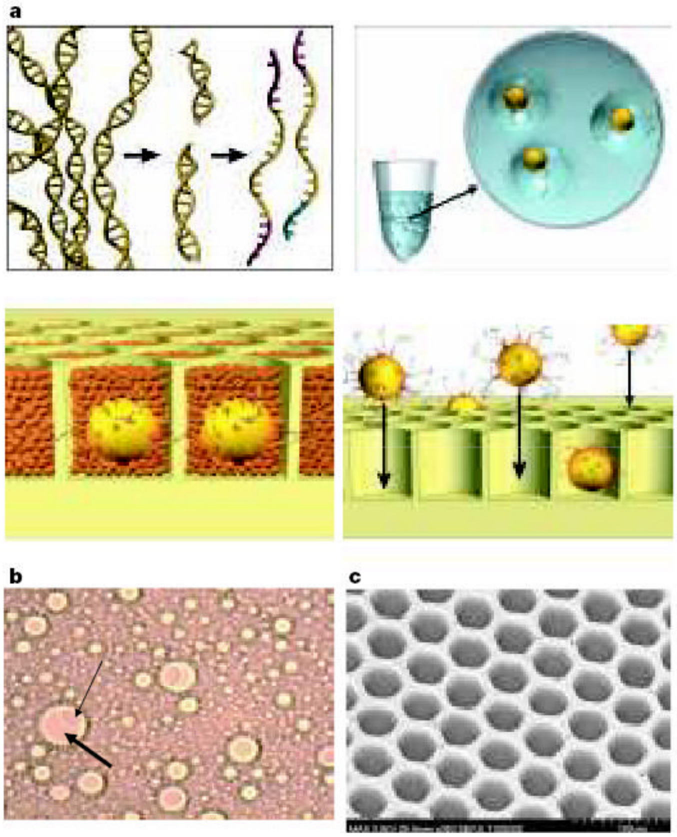

Figure 1. Sample Preparation.

(A) Clockwise from top left: (i) genomic DNA is isolated, fragmented, ligated to adapters and separated into single strands; (ii) fragments are bound to beads under conditions which favor one fragment per bead, the beads are captured in the droplets of a PCR-reaction-mixture-in-oil emulsion and PCR amplification occurs within each droplet, resulting in beads each carrying ten million copies of a unique DNA template; (iii) the emulsion is broken, the DNA strands are denatured, and beads carrying single-stranded DNA clones are deposited into wells of a fibre optic slide; (iv) smaller beads carrying immobilized enzymes required for pyrophosphate sequencing are deposited into each well. **(B)** Microscope photograph of emulsion showing both droplets containing a bead and empty droplets. The thin arrow points to a 28 μm bead, the thick arrow points to an approximately 100 μm droplet. **(C)** SEM photograph of portion of a fibre optic slide, showing fibre optic cladding and wells prior to bead deposition.

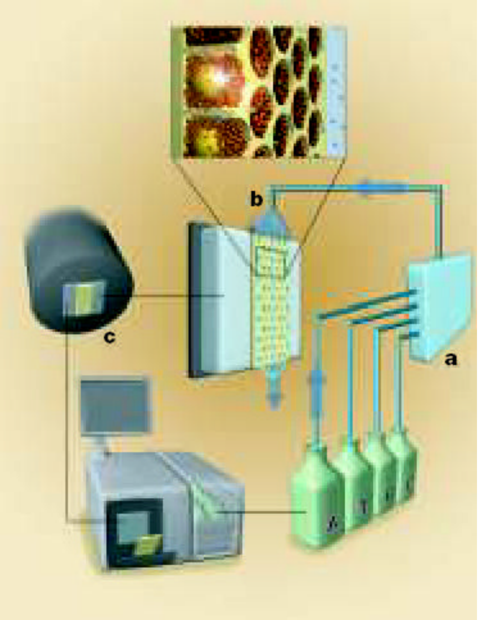

Figure 2. Sequencing Instrument.

The sequencing instrument consists of the following major subsystems: a fluidic assembly (A), a flow chamber that includes the well-containing fibre optic slide (B), a CCD camera-based imaging assembly (C) and a computer that provides the necessary user interface and instrument control.

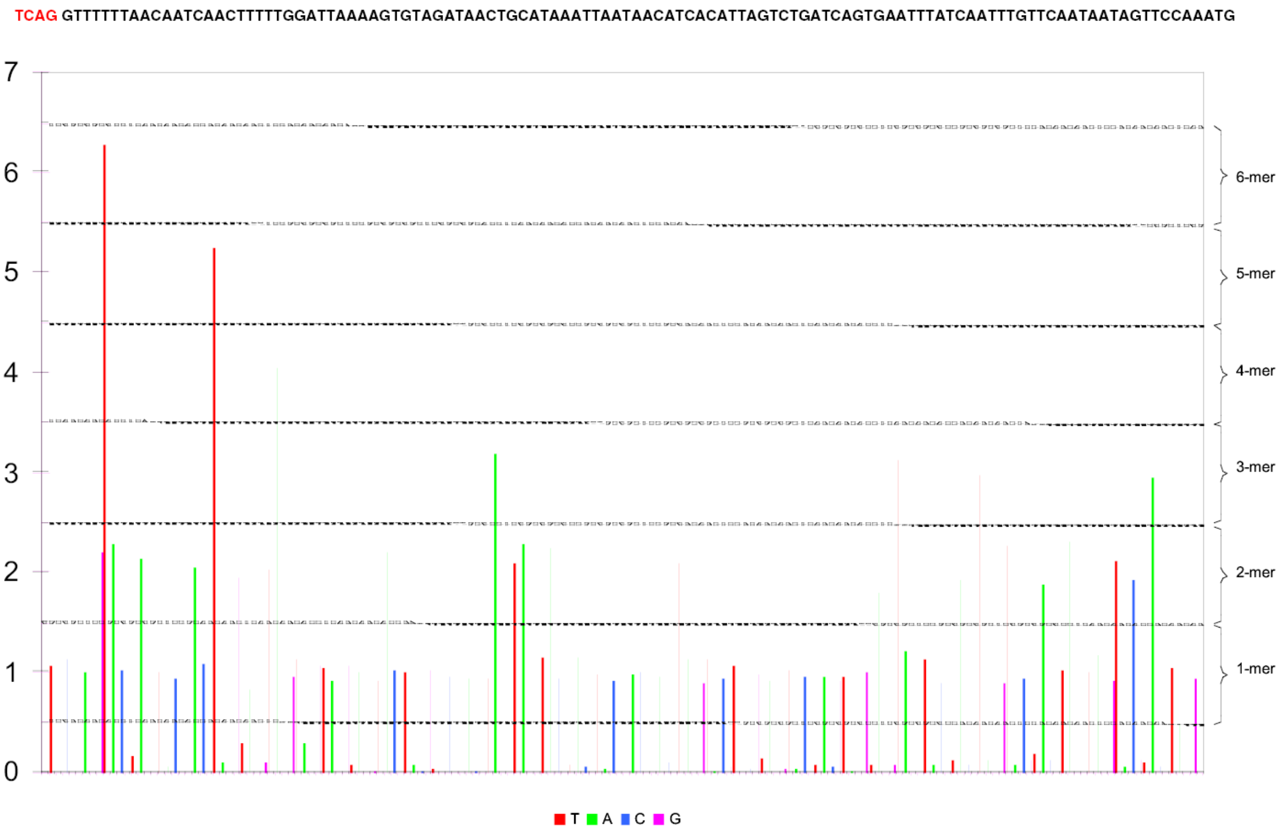

Figure 3. Flowgram of a 113 base read from an *M. genitalium* run.

Nucleotides are flowed in the order T, A, C, G. The sequence is shown above the flowgram. The signal value intervals corresponding to the various homopolymers are indicated on the right. The first four bases (in red, above the flowgram) constitute the “key” sequence, used to identify wells containing a DNA-carrying bead.

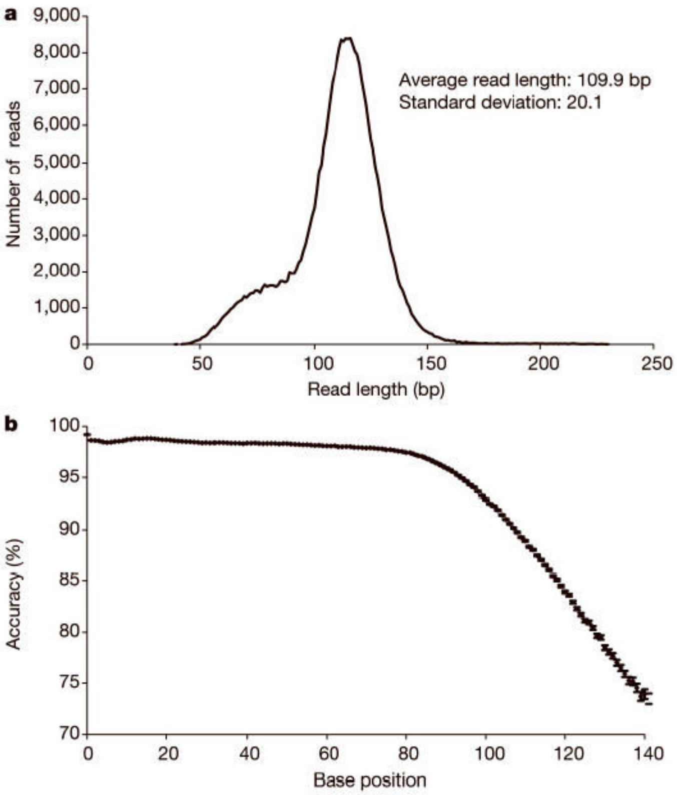

Figure 4. *M. genitalium* Data.

(**A**) Read length distribution for the 306,178 High Quality Reads of the *M. genitalium* sequencing run. This distribution reflects the base composition of individual sequencing templates. (**B**) Average read accuracy, at the single read level, as a function of base position for the 238,066 mapped reads of the same run.

Table 1

Summary of sequencing statistics for test fragments

Size of fibre optic slide	60×60 mm^2
Run Time/Number of Cycles	243 min/42
Test Fragment Reads	497,893
Average read length	108
Number of bases in test fragments	53,705,267
Phred 20 and above bases	47,181,792
Individual read insertion error rate	0.44%
Individual read deletion error rate	0.15%
Individual read substitution error rate	0.004%
All errors	0.60%

Table 2

Summary statistics for *M. genitalium*

Sequencing Summary	**M. gen. 1**
Number of Instrument Runs	1
Size of fibre optic slide	60×60 mm^2
Run Time /Number of cycles	243 min/42
High Quality Reads	306,178
Average read length	110 b
Number of bases in High Quality Reads	33,655,553
Phred 20 and above bases	26,753,540
Resequencing	
Reads mapped to single locations	238,066
Number of bases in mapped reads	27,687,747
Individual read insertion error rate	1.67%
Individual read deletion error rate	1.60%
Individual read substitution error rate	0.68%
Resequencing Consensus	
Average oversampling	40x
Coverage, all ($Z \geq 4$)	99.9% (98.2%)
Consensus accuracy, all ($Z \geq 4$)	99.97% (99.996%)
Consensus insertion error rate, all ($Z \geq 4$)	0.02% (0.003%)
Consensus deletion error rate, all ($Z \geq 4$)	0.01% (0.002%)
Consensus substitution error rate, all ($Z \geq 4$)	0.001% (0.0003%)
Number of contigs	10
***De novo* Assembly**	
Coverage, all ($Z \geq 4$)	96.54% (95.27%)
Consensus accuracy, all ($Z \geq 4$)	99.96% (99.994%)
Number of contigs	25
Average contig size	22.4 kb

The individual read error rates are referenced to the total number of bases in mapped reads.