Cramming More Sequencing Reactions onto Microreactor Chips

John H. Leamon*,† and Jonathan M. Rothberg‡

RainDance Technologies, Inc., 530 Whitfield Street, Guilford, Connecticut 06437, and The Rothberg Institute for Childhood Diseases, 530 Whitfield Street, Guilford Connecticut 06437

Received December 19, 2006

Contents

1. Introduction

Conventional Sanger sequencing has been the mainstay for DNA sequencing in the research community, generating information for countless DNA fragments and whole genomes from bacterial to human. Despite the amount of sequences available to date, the traditional sequencing process has failed to meet an increasing demand due to the technology's high costs and low throughput.[1] Since the inception of the Human Genome Project, concerted efforts have been made to increase the rate and throughput of the sequencing process. These efforts have focused almost exclusively on Sanger-based sequencing techniques; capillary electrophoresis increased the rate at which individual samples could be sequenced[2−4] beyond that of the traditional slab gels, and these advances were furthered by the assembly of individual arrays into parallel capillary arrays.[5,6] Unfortunately, continued throughput increases in capillary array systems may be limited due to the physical requirements for manufacturing, inefficiencies in the injection process,[7] and time-consuming sample preparation processes. Many of the manufacturing issues can be addressed through miniaturization,[8] which could permit additional increases in sequencing throughput and speed as the capillary path is reduced in length and more capillaries or microfabricated channels are packed onto a single device. However, these devices are still in the development phase and the substantial issues of multisample preparation and injection remain unaddressed.

Remarkably, the highest gains in throughput and sequencing speed to date have been realized by a technology that is not based on Sanger chemistry. 454 Sequencing describes a highly miniaturized sequencing system, based on pyrophosphate sequencing technology,[9] where the reactions are conducted in a PicoTiterPlate, a partially etched plate of fused fiber-optic bundles containing roughly 1.6 million, 75 picoliter reaction wells.[10] Sequencing reactions are conducted simultaneously under a flow regime with the open face of the PicoTiterPlate exposed to the stream of flowed reagents and the smooth, unetched face of the plate pressed against a CCD camera. The fiber-optic linkage between the reaction wells in the PicoTiterPlate and the CCD camera permits precise quantification of the amount of light generated within each well.[9]

454 Sequencing represents the evolution of Pyrosequencing[11−13] from individual reactions conducted within microliter-scale volumes in 96- or 384-well plates to the simultaneous sequencing of several hundred thousand picoliter scale reactions, generating an average of 35 Mb of sequence per 5.5 h run, at an average rate of over 5 Mb per hour.[9] Pyrosequencing reactions have been previously conducted in the miniature,[14−18] yet these reactions were limited to single reactions conducted at microliter volumes. The novel development of high density, miniaturized pyrophosphate-based sequencing requires the integration of several substantive changes in the traditional Pyrosequencing process[19,20] from which 454 Sequencing was originally derived.

2. Traditional Pyrosequencing: Theory and Practice

Traditional Pyrosequencing relies upon the principle of sequencing by synthesis,[21] wherein the sequence of a given nucleic acid strand is ascertained through DNA polymerase-mediated replication. A Pyrosequencing reaction is composed of a pool of identical DNA templates with DNA primers annealed to known sites on each strand, to which a DNA polymerase is subsequently bound. Each of the four nucleotide triphosphates is added individually to the reaction in

* To whom correspondence should be addressed. Telephone: 203-458-2947. Fax: 203-458-2514. E-mail: leamonj@raindancetechnologies.com.
† RainDance Technologies, Inc.
‡ The Rothberg Institute for Childhood Diseases. Telephone: 203-458-7100. Fax: 203-458-2514. E-mail: jrothberg@childhooddiseases.org.

10.1021/cr068297s CCC: $65.00
Published on Web 07/10/2007

John H. Leamon was born in 1966 in Lewiston, ME. He obtained a B.A. in Zoology from Connecticut College in 1989, his M.S. in Biological Oceanography from the University of Connecticut in 1994, and a Ph.D. in Physiology and Neurobiology from the University of Connecticut in 1999. John conducted his postdoctoral research at Yale School of Medicine under the direction of Dr. Paul Lizardi, working on rolling circle amplification and single molecule detection. He then spent 5 $^1/_2$ years at 454 Life Sciences, where he developed and productized the emulsion PCR (emPCR) process. He is currently the Project Leader for Nucleic Acid Applications at RainDance Technologies in Guilford, CT. John's research interests revolve around microfluidics, nucleic acid amplification, and emulsion chemistry, specifically centering on defining the practical benefits and limitations to single molecule amplifications in emulsified reactions.

Jonathan M. Rothberg was born in 1963 in New Haven, CT. He earned a B.S. in Chemical Engineering with an option in Biomedical Engineering from Carnegie Mellon University and an M.S., M. Phil, and Ph.D. in Biology from Yale University. Most recently, Dr. Rothberg completed the first sequence of an individual human being (James D. Watson) and initiated the Neanderthal Genome Project. Dr. Rothberg is the founder of CuraGen Corporation, 454 Life Sciences, Clarifi Corporation, and The Rothberg Institute for Childhood Diseases, and the cofounder and Chairman of RainDance Technologies. Dr. Rothberg was named an Ernst and Young Entrepreneur of the Year and is the recipient of *The Wall Street Journal*'s Gold Medal for Innovation for his invention of 454 Sequencing, and The Irvington Institute's Corporate Leadership Award in Science. Dr. Rothberg has appeared on CNBC for his pioneering work in the field of genomics medicine, and his scientific work has been featured on the covers of leading scientific journals including *Cell*, *Science*, and *Nature*. While at CuraGen, Dr. Rothberg developed a series of new medicines, now in over 14 human clinical trials, for the treatment of a wide range of cancers. Dr. Rothberg's invention of a new way to sequence DNA on a chip—454 Sequencing, first motivated by his son's visit to the emergency room—has ushered in the era of personal medicine and is now in use at major pharmaceutical companies, universities, genome centers, and medical centers around the world. Dr. Rothberg was invited to the World Economic Forum in Davos Switzerland as a Technology Pioneer. Dr. Rothberg is a member of the National Academy of Engineering and the Connecticut Academy of Science and Engineering, and he serves on the board of trustees of Carnegie Mellon University.

series, with the series of nucleotide additions repeated cyclically for the length of the experiment. Should the introduced nucleotide base pair with the template strand at the position adjacent to the 3′ end of the primer, it is incorporated into the nascent priming strand by the polymerase, and the length of the priming strand is increased by a single nucleotide (or more, if the template contains a homopolymer stretch). If the introduced nucleotide fails to form a complementary base pair with the template, the nucleotide is not incorporated into the priming strand and the polymerase pauses until the complementary nucleotide is added.

2.1. Nucleotide Incorporation

The polymerase typically utilized in Pyrosequencing is the exo- version of the Klenow fragment from *E. coli* DNA polymerase I, which lacks 3′ → 5′ exonuclease activity. The loss of exonuclease activity is essential for retaining synchronized sequencing of the template strands in the reaction, as it prevents 3′ degradation of the priming strand while the polymerase idles during flows when nucleotides are absent.[20] Additionally, the enzyme exhibits a high level of strand displacement, which enables effective sequencing through regions of the template with high levels of secondary structure.[22]

2.2. Enzyme Cascade

Sequence information is determined by the release of inorganic pyrophosphate (PP_i) with every nucleotide incorporation,[23,24] where the PP_i release initiates an ATP-sulfurylase/luciferase enzyme cascade.[25] This cascade utilizes D-luciferin and adenosine phosphosulfate (APS) in the Pyrosequencing reaction to generate photons proportionally to the PP_i concentration and, by extension, the number of nucleotides incorporated by the polymerase (see Figure 1). The formula for this reaction is described by the following equations,[23] where n = number of DNA residues or nucleotides:

$$(\mathrm{DNA})_n + \mathrm{dNTP} \xrightarrow{\text{DNA polymerase}} (\mathrm{DNA})_{n+1} + \mathrm{PP_i} \quad (1)$$

$$\mathrm{PP_i} + \mathrm{APS} \xrightarrow{\text{ATP-sulfurylase}} \mathrm{ATP} + \mathrm{SO_4^{2-}} \quad (2)$$

$$\mathrm{ATP} + \text{luciferin} + \mathrm{O_2} \xrightarrow{\text{luciferase}} \mathrm{AMP} + \mathrm{PP_i} + \text{oxyluciferin} + \mathrm{CO_2} + h\nu \quad (3)$$

The nucleotides used in Pyrosequencing include the dATP analogue deoxyadenosine α-thiotriphosphate (dATPαS) as opposed to natural dATP. Although luciferase processes dATP as a substrate inefficiently, generating 1−2% of the light produced by an equimolar concentration of ATP,[12] this reaction contributes unacceptable background to the sequencing signal. dATPαS is used instead, as it is utilized by luciferase at only 0.05% the efficiency of dATP[12] but is efficiently incorporated by the Klenow polymerase.[23]

2.3. Apyrase

Accumulation of unincorporated or excess nucleotides in the reaction presents an obvious hurdle for effective sequencing. Accurate sequence determination depends on the correlation between the quantities of light produced and the

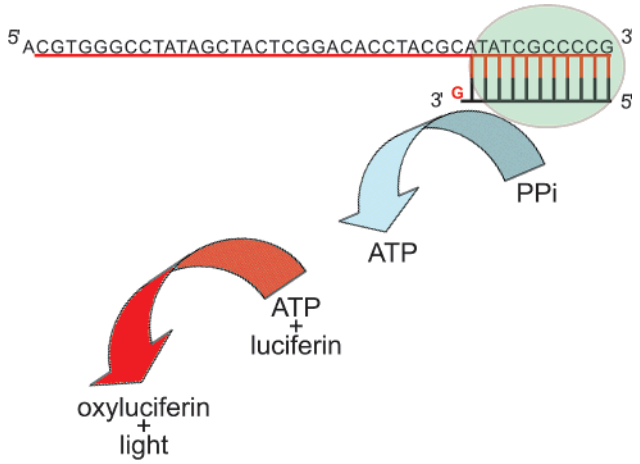

Figure 1. Diagrammatic representation of the Pyrosequencing process. The template strand is represented in red, the annealed primer is shown in black, and the DNA polymerase is shown as the green oval. Incorporation of the complimentary base (the red "G") generates PP_i, which is converted to ATP by the sulfurylase (blue arrow). Luciferase (the red arrow) uses the ATP to convert luciferin to oxyluciferin, producing light as a side product.

particular nucleotides added; simultaneous, or nonsynchronous, incorporation of residual nucleotides from previous additions degrades this correlation. Original embodiments surmounted this problem by immobilizing the DNA on solid supports and inserting a wash step after every nucleotide addition.[12] The wash step was obviated and solution-phase Pyrosequencing enabled with the inclusion of the enzyme apyrase in the reaction mix,[11] degrading both residual ATP and dNTPs to monophosphates and inorganic phosphate via the following reaction:

$$\mathrm{ATP} + \mathrm{dXTP} \xrightarrow{\text{apyrase}} \mathrm{AMP} + \mathrm{dXMP} + 4\mathrm{P_i} \qquad (4)$$

2.4. Read Length Limitations

Even with the inclusion of apyrase, the sequence read lengths generated via traditional Pyrosequencing are typically less than 35 bases,[26] lengths useful for resequencing and SNP identification. While some experimental methods utilizing modified nucleotides have produced increased read lengths,[27,28] standard Pyrosequencing encounters four main limitations to read length that prevent its use in several applications, including *de novo* sequencing and whole genome comparisons.

In the standard Pyrosequencing application, the reaction is conducted in a microplate well. With the addition of multiple nucleotide cycles, the total volume in the well increases, diluting reagent concentrations and reducing enzyme concentration and efficieny.[20] Longer sequencing runs require more nucleotide additions and experience correspondingly greater dilution.

Another limitation is found in an excessive background signal (a decreased signal-to-noise ratio) that prevents accurate sequence determination. The background signal is believed to be the caused by simultaneously sequencing multiple templates (or multiple regions of a single template) within a single reaction. This can result from nonspecific primer annealing, primer−dimer formation, or the presence of a 3′ self-priming hairpin in the template.[26,29] Research has demonstrated that the background can be reduced by the addition of single-stranded binding proteins[26] or elevated sequencing temperatures.[30]

Read lengths are also limited by carry forward, or plus frameshifts, which occur when residual nucleotides from previous additions are not completely degraded, resulting in nonsynchronous incorporation and polymerase advancement on some templates relative to the rest of the population. As the relative number of nonsynchronous polymerases increases, the amount of background also increases. Carry forward is typically caused by the accumulation of substances which inhibit apyrase[20,28] or enzymatic contaminants such as nucleoside diphosphate kinase.[31]

Conversely, read lengths may also be limited by incomplete extension or minus frameshifts, characterized by the lack of available nucleotides or inefficient nucleotide incorporation by the polymerase. Incomplete extension can be caused by a relative excess of apyrase, degrading nucleotides before they can be incorporated, or reduced efficiencies in nucleotide incorporation, such as those seen with some modified bases.[28] This effect is particularly evident after long homopolymer stretches when the polymerase is unable to complete the region due to localized nucleotide depletion or an insufficient incorporation period.[20]

3. Modifications to Standard Pyrosequencing

Considerable research has been conducted on the various enzymatic components of the Pyrosequencing process.[32] Alternative DNA polymerases, such as φ29[22] and Sequenase,[33] have been investigated with a range of results. While Sequenase exhibited improved dATPαS incorporation, lower susceptibility to primer−dimers, and self-priming templates,[33] the enzyme has yet to obtain widespread use in Pyrosequencing, presumably due to increased cost relative to exo-Klenow. Other research[34] has focused on optimizing the enzyme cascade at the heart of Pyrosequencing, utilizing pyruvate orthophosphate dikinase (PPDK) to convert PP_i to ATP. This modification increased sequence read lengths to 70 bases while improving the assay's lower limit of detection and decreasing the requisite amount of template DNA by 2 orders of magnitude. In addition, the PPDK generated substantially more light per nucleotide incorporation, raising the possibility of replacing cooled CCD cameras or PMT-based detectors with less expensive photodiode arrays.[34] The benefit of linking the luciferase to DNA-binding proteins such as polymerases or single-stranded binding proteins (SSB) has also been explored,[35] wherein luciferase could be directly bound to the DNA template rather than added to every reagent flow, providing potential cost reductions.

3.1. Miniaturizing Pyrosequencing

Significant progress has been made in miniaturizing the underlying technology, reducing the required reagent volumes to nanoliters. Some of the methodologies entail shrinking the size of the reaction and reagent wells to the point where they can be contained on a single microdevice.[14] Other researchers have conducted the reaction in a flow regime[15−18] utilizing immobilized DNA templates, thereby reducing the reliance upon apyrase for ATP and dNTP degradation.

Standard Pyrosequencing has been successfully utilized for a variety of applications that require accurate resequencing or SNP detection over short (30 base) regions. These include population level investigations of organisms ranging from protists[36] to pigs,[37] SNP discovery and confirmation,[38,39] promoter methylation studies,[40−42] and bacterial strain iden-

tification.[43,44] Increases in application efficiency have been achieved through combining multiplex PCR and Pyrosequencing, permitting the simultaneous sequencing of several discrete SNPs with a single reaction.[42,45] Despite the potential advances in miniaturization and enzymology, pyrophosphate-based sequencing was unable to deliver the read densities, lengths, and accuracies required for *de novo* DNA sequencing of exons or whole genomes until it was adapted to the confines of a PicoTiterPlate well.

4. The 454 Sequencing System

The advantages of PicoTiterPlate-based sequencing are similar to those seen in the development of the integrated circuit, where the integration of numerous, small transistors into a single device was able to provide an enormous performance improvement over the previous devices assembled from larger discrete circuits or vacuum tubes. As with the transistors in the integrated circuits, the small size and close proximity of the reaction wells comprising the PicoTiterPlate permit increased speed, density, and throughput by utilizing diffusion to solve issues associated with simultaneously loading and washing more than a million reaction wells. Sequencing runs conducted on this platform generate over 200 000 individual sequencing reads, producing an average of 35 MB of sequence data.[9] (Since this manuscript was written, a second generation sequencer, the GS-FLX, has been released, producing 400 000 individual reads, with lengths of 200−300 bases per read and an average of 100 MB of sequence data per run. This latest product is not discussed in this work.) Additionally, the integration of millions of reaction wells into a single, mass produced platform provides inherent manufacturing advantages in terms of production capability, capacity, and reliability relative to capillary arrays at even a fraction of the throughput. As Gordon Moore in his article "Cramming more components onto integrated circuits"[46] predicted the decrease in size and increase in number of features on a microchip in the past, the size and quantity of wells in a PicoTiterPlate show similar room for future improvement.

4.1. High Throughput Sample Preparation

Although sequencing within the PicoTiterPlate truly enables 454 Sequencing, the conversion from macroscale Pyrosequencing to picoscale 454 Sequencing requires significant engineering of almost all aspects of the original technique. One of the most critical hurdles for any viable high throughput methodology to overcome is the ability to mate the high throughput system with an equally high throughput sample preparation process. Regardless of the potential throughput a new system may provide, it is essentially meaningless if sufficient samples cannot be generated for processing in a reasonable period of time. For example, it would be highly impractical to employ the standard Pyrosequencing sample preparation process, where biotin-labeled PCR primers are used to bind DNA templates to streptavidin beads,[47] to generate several hundred thousand discrete samples. The 454 Sequencing system sample preparation process, summarized in Figure 2, utilizes emPCR,[9] based upon the emulsion PCR[48,49] process, to generate up to several million clonally amplified beads for sequencing at a rate sufficient to support multiple PicoTiterPlates. The emPCR process emulsifies a PCR reaction mix in microreactors suspended in a thermostable oil matrix. The asymmetric PCR process conducted inside the droplets serves to both amplify and immobilize the template DNA on 28 μm DNA capture beads to which complementary primers have been covalently bound. Emulsification of the PCR reaction permits clonal, unbiased amplification of complex samples[50] including entire genomes,[9,50] by segregating each template in a discrete reaction droplet where biases due to amplification efficiency,[51,52] G+C content,[53] etc. are minimized. With this amplification system, a single PCR tube can generate hundreds of thousands of unique templates for sequencing, representing a substantial improvement in throughput over traditional Pyrosequencing sample preparation. The clonal nature of the amplification process is preserved by the capture of the products on the DNA capture beads. The solid-phase capture also permits enrichment[9] of the PCR positive beads (arising from droplets where the PCR reaction was successful) from PCR negative or "null beads" (from droplets which lacked template DNA due to limiting template dilutions); this increases the percentage of DNA capture beads that generate usable sequences during a sequencing run. Additionally, the solid-phase amplification process in emPCR provides the template DNA in a bead-immobilized format, precisely the format required for sequencing in a flow regime.

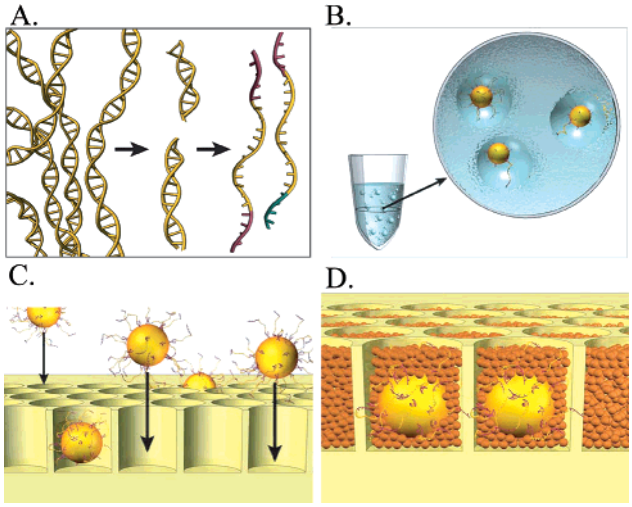

Figure 2. Overview of the 454 Sequencing sample preparation process. (A) Template DNA is broken into small random fragments through nebulization. Oligonucleotide adaptors are ligated to the fragments, and the fragments are then separated into single strands. (B) Single-stranded, adapted fragments are bound to beads at a limiting dilution, resulting in at most a single copy of template DNA bound to a given bead. The beads are emulsified in a thermostable PCR-reaction-mixture-in-oil emulsion, and the individual template is amplified several million-fold via PCR. The amplified template is retained on the DNA capture bead by extension along reverse primers covalently bound to the surface of the bead. (C) Following amplification, the emulsion is broken, the DNA strands are rendered single-stranded, and beads are deposited into wells of a fiber-optic slide. (D) The DNA capture beads in each well are subsequently covered by a layer of smaller beads upon which enzymes required for pyrophosphate sequencing have been immobilized. Adapted by permission from Macmillan Publishers Ltd: *Nature* (Margulies et al. *Nature* **2005**, *437*, 376.), copyright 2005 (http://www.nature.com).

4.2. Sequencing in a Flow Regime

Sequencing within the PicoTiterPlate also requires substantial engineering to reflect the fact that, given the small volume of the wells (~75 pL) and the constant flow regime across the open well face, diffusion rates control almost every

aspect of the reaction. Reagent flow provides a mechanism for rapid mixing within the well, obviating the mechanical shaking required in traditional Pyrosequencing.[32] However, successful sequencing depends upon striking the proper balance between diffusion rates both into and out of the well and the rate of enzymatic reactions occurring within. Diffusion of nucleotides and substrate into the PicoTiterPlate wells enables a large number of sequencing reactions to be conducted simultaneously, as manual reagent loading into every well is impossible. Diffusion of unutilized reagents and reaction byproducts out of the well is also essential, as this supports the generation of long sequences (over 100 bases) without signal degeneration due to increased signal background,[28] increased reaction volumes, or accumulation of inhibitory substances. Sequencing within a flow regime can pose challenges as well, however. Within this diffusion-dominated environment, the three enzymatic processes governing nucleotide incorporation, PPi conversion to ATP, and the conversion of luciferin to oxyluciferin must have sufficient time to reach completion before the essential reagents are washed from the well.

Successful PicoTiterPlate-based sequencing is accomplished by combining bead-immobilization with high efficiency enzymes to achieve a balance between reaction kinetics and diffusion. The theoretical kinetics for a single well, which predict that the nucleotide incorporation by the Bst polymerase is completed in 1 to 2 s and that the time required to convert the majority of the PP_i to light is less than 10 s, roughly equivalent to the 10 s rate of diffusion in the well, are described in detail later in this manuscript.

4.2.1. Immobilized Enzymes

The constant reagent flow in 454 Sequencing necessitates changes in how critical enzymes are retained in the well and in the enzymes used in the cascade. The standard Pyrosequencing reactions utilize sulfurylase and luciferase in solution: a practice which becomes costly in flow reactions, as the enzymes have to be flowed into the well simultaneously with every nucleotide. Instead, the 454 Sequencing system immobilizes the enzymes required for the light producing cascade on solid microparticles and loads those enzyme-bearing beads into each of the PicoTiterPlate wells.[9] This is accomplished by preparing both the sulfurylase and luciferase as biotin carboxyl carrier protein (BCCP) fusions.[9] A lysine residue in the BCCP region permits the covalent linkage of a biotin molecule during *in vivo E. coli* expression, and the purified fusion proteins are then bound as a 3:1 ratio (mgs luciferase to mgs sulfurylase) to streptavidin-coated 2.8 μm Dynal M280 paramagnetic microparticles. Deposition of the beads into the PicoTiterPlate wells place the enzymes in close proximity to the DNA capture beads, decreasing the distance between the source of the PP_i and the enzyme cascade. The enzyme beads also serve to physically retain the DNA capture beads, preventing the beads from washing out of the well during reagent flow.

4.2.2. Selection for Specific Enzyamtic Properties

The luciferase selected for the 454 Sequencing system exhibits lower K_m values than those typically used in Pyrosequencing. The 454 system utilizes the UltraGlow Luciferase (Promega, Madison WI), a thermostable glow-type luciferase, as opposed to the commonly used flash luciferase obtained from *Photinus pyralis* American fireflies. The K_m for UltraGlow (0.7 μM ATP) has been reported to be on the order of 100-fold lower than that of wild type *P. pyralis*.[29] The reduced K_m could prove limiting in standard Pyrosequencing reactions where the total number of DNA template molecules is high and there is no diffusion of ATP or PP_i out of the well.[29] In 454 Sequencing, however, there are several important distinctions: the total number of DNA molecules is lower, reducing the amount of PP_i produced by nucleotide incorporations, and both ATP and PP_i continuously diffuse into the reagent flow and are lost from the well. With the interwell concentrations of PP_i and ATP available for the 454 Sequencing enzyme cascade thus reduced, optimal sequencing relies upon utilizing the maximum amount of the PP_i and ATP. This, in turn, favors the most efficient enzymes, those with the lowest K_m values. UltraGlow's substantially lower K_m allows increased light production at lower levels of ATP than the less efficient wild type *P. pyralis* enzyme, and allows the system to maintain a linear relationship between the photons produced and homopolymer size up to 8 bases in length.[9]

The thermostable properties of UltraGlow luciferase permit an additional feature of 454 Sequencing impossible with traditional Pyrosequencing; 454 Sequencing can occur at temperatures up to 35 °C. The 28 °C temperature required in standard Pyrosequencing is necessitated by the thermolabile *P. pyralis* luciferase. Pyrosequencing at elevated temperatures has been shown to be beneficial, decreasing background from primer−dimers and hairpin loops,[29] contributing to extended sequence read length. Additionally, increased temperatures enable the use of polymerases with higher optimal temperature ranges, such as Bst polymerase, with a 65 °C incubation temperature.

4.2.3. Bst Polymerase

The ability to sequence through problematic sequence motifs using the 454 system[54] is partly due to the use of Bst polymerase, which was substituted for the Klenow polymerase used in traditional pyrosequencing. The Bst polymerase is an exo-enzyme[55] that has been shown to possess a high degree of strand displacement activity[56] and extremely low rates of replication slippage.[57] It has also been shown to be highly suitable for whole genome amplification,[58] able to successfully replicate a diverse range of structural motifs and GC contents.[59] The enzyme exhibits a high degree of processivity,[60] ensuring that it will remain bound to the priming strand during the course of the sequencing reaction. This eliminates the costly need to replenish the polymerase concentration by periodically flowing additional enzyme into the well.

4.3. Corrective Software Algorithms

Despite the enzymatic improvements that enable successful sequencing within the confines of a PicoTiterPlate well, interactions remain that are most effectively corrected by the application of software algorithms. All of the sequencing reads generated in a 454 sequencing run undergo a two stage process of algorithmic filtering: image processing and signal processing. Together, image and signal processing output filtered reads, and bases within each of the reads are then called and assigned a quality score.

4.3.1. Image Processing

Image processing performs initial pixel-level calculations on the raw data captured by the CCD. These calculations

include such processes as background subtraction and pixel-level signal normalization. These data are used to identify and group clusters of pixels as active PicoTiterPlate wells, those in which sequencing is occurring. For each active well, the raw signal data is extracted from the CCD images corresponding to all nucleotide or PP_i flows, and stored for subsequent signal processing. By identifying and extracting the active well data, the volume of raw data is reduced to amounts that can be efficiently processed by the subsequent filtering algorithms in a practical time frame.

4.3.2. Signal Processing

The signal processing algorithms perform well-level calculations across the whole series of images generated in a run for each of the active wells. These processes include the classification of signal height obtained for a given flow in a well, correction for loss of polymerase synchrony within a run, and correcting well to well interactions.

4.3.2.1. Signal Peak Classification. Automated base-calling algorithms used for traditional Sanger sequencing instruments determine the signal strength for each peak on the chromatograph by calculating the area under each peak. In contrast, 454 Sequencing does not need to integrate the detected signal, because the total amount of signal (light) generated during the nucleotide extension phase of the sequencing cycle is captured quantitatively on the CCD. The strength of the signal from each well is determined by the number of photons that strike the CCD chip at that position. The signal intensity is proportional to the number of bases incorporated during that flow, which consists of both the number of individual template strands currently sequencing in the well (typically around 10M) and the number of nucleotides incorporated on each strand.

Since base-calling accuracy is directly dependent on signal intensity, it is important to understand potential sources of error for signal intensity measurement. The variability in signal strength, or the standard error of the signal, is proportional to the signal strength obtained from any well. Thus, as the length of homopolymers increases, so does the signal strength, and so too the variability, giving more potential for error in accurately determining the true signal strength. In a typical shotgun *E. coli* run (data not shown), a single base can be called with greater than 99.9% accuracy, homopolymers of length 2 at 99.5% accuracy, and homopolymers of length 3 with 99.0% accuracy. The accuracy of base calling in a single read falls off so that base calls of 9-base homopolymers are approximately 64% accurate. It is important to note, however, that 99% of 9-base homopolymers are called either as 8, 9, or 10. This allows the 454 Sequencing algorithms to address homopolymer variability through oversampling, as the variability in each signal is random, so multiple reads of the same homopolymeric region of the template allow averaging of the error associated with that particular homopolymer stretch, yielding consensus base calls of close to 99% accuracy for 9-base homopolymers.

For Sanger-based systems, there are a number of well-known sequence related problems that can occur; these are discussed elsewhere (for example, see PE/Applied Biosystems *Automated DNA Sequencing Guide*). These problems are caused by either the template DNA or the sequencing chemistry, and include false stops in dye−primer chemistry, compressions, GC-rich (>70%) regions of sequence, overall GC-rich sequences, regions of pronounced secondary structure, GT-rich regions in BigDye terminator chemistry, repetitive DNA, and homopolymeric regions.[61] Of these, only homopolymer regions are held in common. Sanger reads of homopolymers, especially those that contain long (10+) T or A homopolymer regions, are susceptible to "slippage",[57] which generates a pool of products in the sample with different lengths, resulting in "out of phase" type errors. Slippage in the region of a homopolymer is dependent on the length of the homopolymer and can lead to noisy data following the homopolymer as well as inaccurate sequencing at the homopolymer. In many cases, it is possible to clone the homopolymeric region for more accurate sequencing. However, cloning of difficult-to-sequence homopolymeric regions tends to be expensive and time-consuming. In addition, multiple clones must be sequenced to ensure that cloning artifacts are excluded.

4.3.2.2. Carry Forward and Incomplete Extension. As with traditional Pyrosequencing, 454 Sequencing is affected by the gradual loss of synchronicity among the individual templates on a given bead during a sequencing run.[20] In these cases, a small number of the polymerases either fall behind the others due to incomplete extension of the template or get ahead of the other polymerases due to the presence of residual nucleotides in the well, permitting "carry forward". The 454 Sequencing process has been shown to experience approximately 0.1−0.3% incomplete extension and 1−2% carry forward[9] during an entire run. As this loss of synchronicity is cumulative, it must be corrected to enable accurate sequencing at long read lengths. The two sources of error are corrected by a single algorithm referred to as CAFIE (Carry Forward, Incomplete Extension), which relies upon a combination of physical modeling of intrawell diffusion and washing rates, as well as empirical data collected over several years and gigabases of sequence analysis. As the loss of synchronicity accumulates with time and the length of the sequencing run, the bases sequenced toward the end of run are more affected than those at the start. Figure 3 illustrates the effect of the CAFIE correction on a test fragment sequence obtained during a 100 cycle run, comprised of 400 individual nucleotide additions. The early cycles of both the pre- and post-CAFIE correction flowgrams display good-quality sequencing: the signal-to-noise obtained for positive and negative peaks is high, and the background for the negative reads stays low and shows no sign of steady increase. Sequenced across only 50 bases, the error rate for the TF is 2.0% (1 error in 50 bases) without CAFIE and 0.0% with the CAFIE correction. When the entire 216 base read is analyzed, however, the effect of the CAFIE correction is much more obvious; the error rate for the uncorrected sequence 12.0% (26 errors in 216 bases), but the read remains error-free following application of the CAFIE algorithm. This is visually apparent when the flowgrams are examined; the background increases and the signal decreases along the length of the run when the CAFIE correction is not employed.

4.3.2.3. Interwell Interactions: Optical Bleed. Interwell interactions require software correction also. These interactions take two forms: one termed optical bleed, where a small amount of the light generated in one well diffuses through the fiber cladding surrounding the originating well and enters neighboring optical fibers, and the other chemical crosstalk,[9] caused by the diffusion of PP_i from the well where it is produced to another well where it generates light. The correction for optical bleed is fairly simple, as the effect is isotropic, varying only with the intensity of light generated

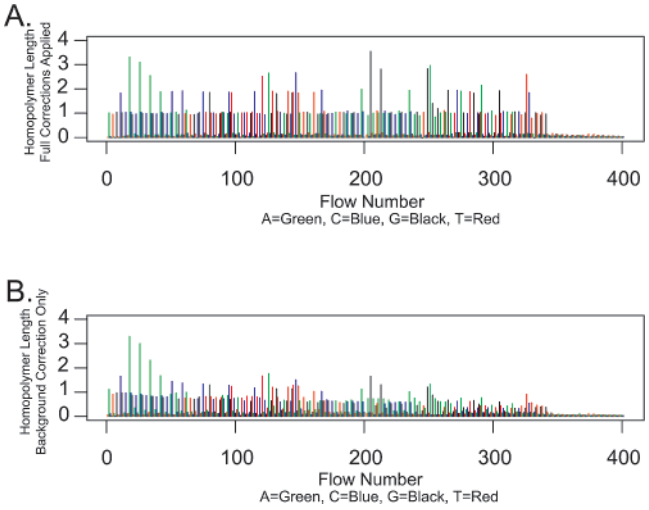

Figure 3. Effects of corrective algorithms on sequence results from a 216 base test fragment. (A) Flowgram generated from a full 100 cycle run (composed of 400 individual nucleotide flows) with software correction enabled. The sequence drops to background levels after approximately 350 flows, as the end of the fragment has been reached. Note that the background levels remain well below 1 on the *Y* axis. No errors were made in sequencing the 216 base template. (B) Flowgram generated from the same test fragment on the same sequencing run without application of any software correction. The sequence drops to background levels after approximately 350 flows, as the end of the fragment has been reached. Note that the background levels rise and the sequencing signal drops as read length increases. In the absence of software correction, the 216 base template accumulated 26 errors, a 12% error rate.

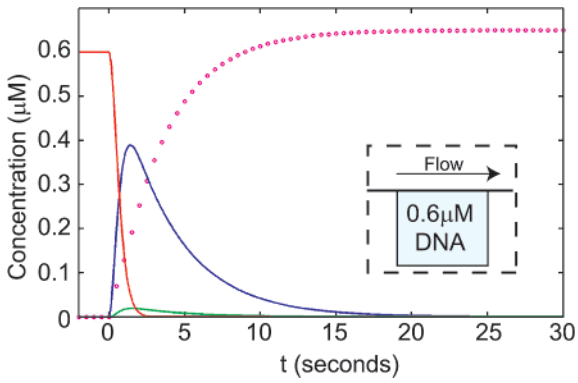

Figure 4. Kinetic modeling of 454 Sequencing within a single PicoTiterPlate well. The solid red trace represents the concentration of unextended DNA fragments, the purple dotted trace displays the dNTP concentration (scaled by 0.1× to allow simultaneous display), the solid blue trace indicates the concentration of PP_i, and the solid green trace illustrates the ATP concentration in the well. The calculations are based on the assumption of 20 million DNA copies per bead, generating an initial 0.6 μM concentration of unextended DNA fragments. The inset figure indicates the orientation of the PicoTiterPlate well with respect to the reagent flow.

within a given well and the height within the well at which the light is generated, where dispersion increases with elevation. The optical bleed correction is accomplished by identifying the contribution of a given well to the surrounding wells and applying a masking algorithm to remove that contribution from the raw signals obtained in the neighboring wells. The degree to which the adjacent well signals are reduced depends upon both the intensity of the initial well and the physical proximity of the adjacent wells.

4.3.2.4. Interwell Interactions: Chemical Crosstalk. Chemical crosstalk is a more detailed, directional interaction, controlled by the direction of reagent flow and the amount of PP_i and ATP that escape from a given well to generate light in a downstream well. The release of light-generating substrates from an individual well is illustrated by the coupled kinetic equation (eq 5) below and graphically in Figure 4:

$$\frac{d}{dt}\begin{bmatrix}DNA_n\\ dNTP\\ PPI\\ ATP\end{bmatrix}_{(1)} = \begin{bmatrix}-R_{bst(1)}\\ -R_{bst(1)} - k_c([dNTP]_{(1)} - [dNTP]_{(0)})\\ R_{bst(1)} - R_{sulf(1)} + R_{luc(1)} - k_c[PPI]_{(1)}\\ R_{sulf(1)} - R_{luc(1)} - k_c[ATP]_{(1)}\end{bmatrix} \quad (5)$$

where the subscript (0) indicates a variable particular to the reagent flow; the subscript (1) signifies variables relating to the well; the mass transfer coefficient = k_c; and R_{bst}, R_{sulf}, and R_{luc} denote the rates of Bst polymerase, sulfurylase, and luciferase respectively.

The equation increases in complexity when multiple ATP/PP_i generating wells are considered, giving rise to eq 6,[9]

$$\frac{d}{dt}\begin{bmatrix}DNA_n\\ dNTP\\ PPI\\ ATP\end{bmatrix}_{(2)} = \begin{bmatrix}-R_{bst(2)}\\ -R_{bst(2)} - k_c([dNTP]_{(2)} - [dNTP]_{(0)})\\ R_{bst(2)} - R_{sulf(2)} + R_{luc(2)} - k_c([PPI]_{(2)} - \theta[PPI]_{(1)})\\ R_{sulf(2)} - R_{luc(2)} - k_c([ATP]_{(2)} - \theta[ATP]_{(1)})\end{bmatrix} \quad (6)$$

which is applied to wells downstream (indicated by variables with the subscript (2)) from the initial well (see Figure 5). A new variable, θ, refers to the mixing ratio. The correction equation predicts that the interwell distance can be significantly reduced below the current 50 μm center-to-center separation while retaining minimal crosstalk effects on neighboring wells.[9]

The combination of a specialized sample preparation method, engineered enzymes, and a software correction system, all adapted to the demands of miniaturized high density sequencing in a flow regime, has enabled 454 Sequencing to generate accurate sequencing reads that eclipse those of traditional Pyrosequencing. Subsequent research has demonstrated the system's accuracy and sensitivity, capable of detecting single nucleotide polymorphisms (SNPs) from within a complex tissue population at frequencies as low as 0.28%,[62] compared to the 4%[63] and 5%[64] lower limits of detection ascribed to traditional Pyrosequencing. The data generated by the 454 Sequencing system is accurate enough to distill the total differences between 4 MB drug resistant and susceptible genomes to four single base mutations, one of which proved to be the source of the drug resistance.[65]

Read lengths generated with the 454 Sequencing system have superseded those obtained with the standard Pyrosequencing also. The introductory paper on 454 Sequencing reported average read lengths of 100−120 bases on production instruments and mentioned that 400 base reads had been achieved on prototype instruments.[9] Since that time, prototype sequencers have demonstrated extended sequencing reads with median lengths beyond 400 bases; individual reads with perfect accuracy over 500 bases have also been recorded.[66] Figure 6 illustrates continued progress in read

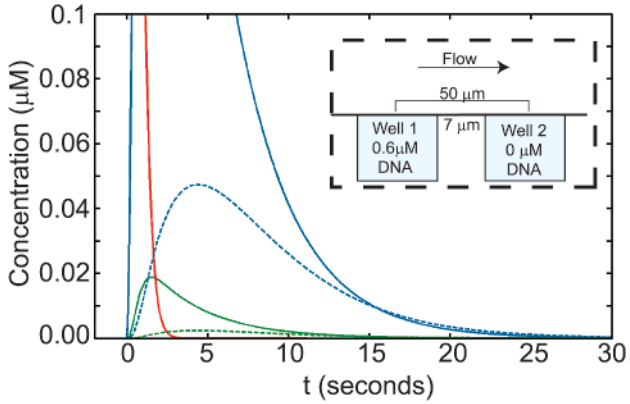

Figure 5. Kinetic modeling of the interwell effects of chemical crosstalk between well 1, an actively sequencing well, and well 2, a downstream nonsequencing well. The two well centers are separated by 50 μm, the wells themselves are separated by 7 μm cladding. As in Figure 4, the solid red trace represents the concentration of unextended DNA fragments in the original well (well 1), the solid blue trace indicates the concentration of PP_i in well 1, and the solid green trace illustrates the ATP concentration in well 1. Note that the peak concentrations for both DNA and PP_i for well 1 are identical to those shown in Figure 4 but are off the scale of this graph. The peak DNA concentration in well 1 is 0.6 μM, and PP_i peaks at approximately 0.4 μM. The dotted traces represent the ATP concentration (dotted blue trace) in well 2, a nonsequencing well containing no DNA, located 50 μm downstream. The dotted green line displays the PP_i concentration in well 2. The calculations are based on the assumption of 20 million DNA copies per bead, generating an initial 0.6 μM concentration of unextended DNA fragments. The inset figure indicates the orientation of the PicoTiterPlate wells with respect to the reagent flow.

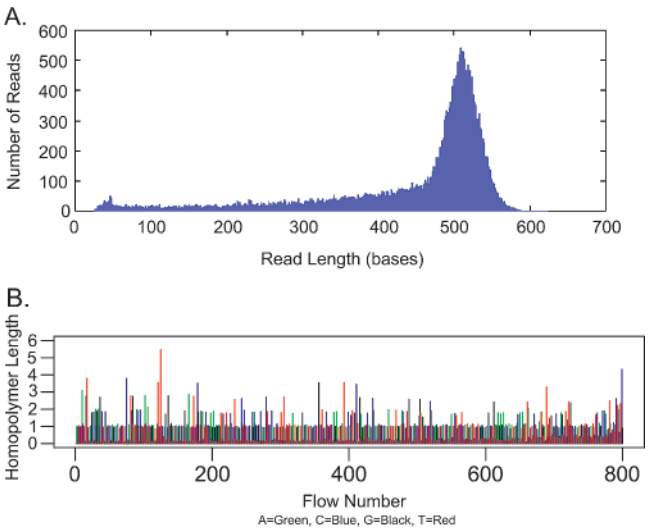

Figure 6. (A) Read length distribution of an *E. coli* sequencing run conducted on a research prototype 454 Sequencer. The peak read length distribution is approximately 510 bases, with a maximum read length of 599 bases. (B) Flowgram generated for the longest perfect read (0% error) taken from the *E. coli* run in Figure 4A. The total read length was 559 bases.

length and accuracy, with prototype sequencers generating reads with peak distributions at approximately 510 bases (Figure 6A), and an error-free, 559 base, individual read (Figure 6B).

4.3.3. Quality Scores

Each of the sequencing reads that passes the image and signal filtering is classified as a "high quality" read, and all of the bases that comprise the reads that passed the filters are classified as high quality as well. It is important to note that filtering is handled on a per-read, not a per-base, format; all signals contained in the reads that passed filtering and trimming are considered high quality. Quality scores are assigned to individual bases, however, by the following process.

4.3.3.1. Error Models and Error Correction. Errors in Sanger sequencing, regardless of the base caller used for data analysis, stem from several sources, as reviewed by Ewing et al.,[67] including poor quality reads in the beginning and end of runs, compressions, and weak or variable signals due to sequence motifs, or reagent quality issues. 454 Sequencing error is dominated by a single source, signal variability, which can be as high as 5%. This variability is probably the result of variability inherent in the enzymatic processes of synchronous enzymatic extension (see eq 6) occurring simultaneously on 10M or more immobilized DNA strands within a PicoTiterPlate well. It should be noted that although this signal variability is the dominant source of error, it is not the only error source; sequencing chemistry issues such as carry forward and incomplete extension are known to contribute error. Fortunately, these errors are largely predictable based on sequence environment and can be corrected. Hardware-induced variance can contribute error but can be measured and corrected for as well. Since signal strength variance is largely random in nature and cannot be corrected for, it is one of the most difficult factors to address in determining the quality of individual bases generated by 454 Sequencing.

Quality scores for 454 Sequencing are assigned purely on a signal level, ignoring factors such as the second order effects arising from the sequence motif in which the base in question occurs, well density, etc., although these factors contribute to overall accuracy to some degree. The correspondence between signal level and base quality has been measured empirically by mapping sequences to known genomes. The known sequence was then compared to the actual signal obtained during the run, and the respective number of correct versus incorrect base calls was determined. These data were used to create parameters for correlating signal intensities and homopolymer lengths. Using these parameters, the probability that any given signal corresponds to a homopolymer of length x can be calculated, and this probability is converted in a Phred score. Because this relationship was derived from a limited number of genomes, a model was generated and extrapolated to reflect all possible combinations of signals. The model was shown to match the empirical data collected by sequencing genomes from multiple bacterial species and comparing reads to known references (data not shown). For example, a Q40 score indicates 99.99% accuracy, or one error per 10 000 bases, requiring the collection and analysis of least 10 000 data points; logistical considerations limited the empirical confirmation of 454 Sequencing scores out to approximately Q60 (1 error in 1 000 000 bases), where 1 000 000 data points or more are required for assessment.

At the highest level, the relationship between the model and the empirical data has been confirmed. As a group, the relationship is solid, and a Q30 score reflects a signal that is correct in 999 out of 1000 base calls. By employing an empirical approach, however, potential accuracy increases obtained through understanding and correcting potential read-specific, second order effects are neglected.

4.3.3.2. Insertions and Deletions. It is important to note that, in 454 Sequencing, a single signal value is generated for a homopolymer, regardless of the number of bases contained within it. So a quality score obtained from a single signal value pertains to all bases within the homopolymer. Consider the case of sequencing a homopolymer in a single read which generates a signal intensity of 4.8 units. The probability that the homopolymer is at least 4 bases in length is quite high; the quality score assigned for bases 1 through 4 will be high. At this intensity, it is clearly a homopolymer of at least 4 bases, and quite likely 5 bases, so the software will treat it as a 5-base homopolymer. The quality score for the final fifth base will be lower, because the certainty that the homopolymer is five bases long is lower than the certainty the homopolymer is at least 4 bases long. It is less likely, though possible, to be as short as 3, or as long as 6, but 5 bases and quality scores will be output. If the homopolymer was actually 6 bases long, the output would look like it contained a deletion error because the sixth base and the correspondingly low quality score are not displayed. If the homopolymer is not 5 bases, it will look like a frame shift error. Longer homopolymers will have higher signal intensity with greater variability, and correspondingly lower probability of an accurate base call in a single read.

4.3.3.3. 454 Quality Scores versus Sanger Quality Scores. The 454 quality score is based on signal intensity and is generated for each signal individually. As the signal intensity for any given base is random with respect to the signal intensities recorded previously and subsequently, quality scores can vary within a read, with high quality reads interspersed with lower quality reads. Unlike Sanger sequencing, where the onset of low quality scores often signals the end of quality sequencing within a read, 454 Sequencing quality scores have little or no predictive value relative to the quality of preceding or following bases.

5. Conclusions

Since completing the draft of the Human genome, the potential scale and scope of genomic sequencing projects has increased, resulting in an accelerated demand for sequence data. This desire for increased sequencing throughput has been to some degree attained on a variety of next-generation sequencing platforms.[68] The longest read lengths currently available on a next-generation sequencing system are produced by 454 Life Sciences, which uses pyrophosphate-based sequencing-by-synthesis to deliver 35 Mb of sequence data per 5.5 h run. Substantive design changes enable the 454 Sequencing system to circumvent the previous limitations in accuracy, sensitivity, and read length that challenged the original Pyrosequencing technology. The demonstrated ability to generate error-free sequencing reads over 500 bases in length, coupled with a highly scalable sample preparation, suggests that continued improvements in both throughput and accuracy are likely. Should this trend continue, it is quite possible that 454 Sequencing reads may intersect Sanger sequencing read lengths in the coming years at a substantially higher speed, higher throughput, and lower cost than those provided by Sanger-based technologies. In this event, whole genome sequencing will become a routine, rapid, affordable process for genomes from microbes to humans, and the sequencing bottleneck will pass from data generation to data analysis.

6. Acknowledgments

The authors would like to thank Mr. Chris McLeod and Drs. Michael Egholm, Mark Driscoll, and Jan Berka for their careful proofreading and suggestions, and Drs. Yi-Ju Chen, Wen He, Joe Fierro, Said Attiya, Zhoutao Chen, and Maithreyan Srinivasan and Mr. Faheem Niazi for providing manuscript figures. The authors would also like to acknowledge NHGRI for continued support under Grants R01 HG003562 and P01 HG003022 for the development of this platform, as well as all of the enthusiastic employees of 454 Life Sciences Corp. who developed the sequencing system.

7. References

(1) Metzker, M. L. *Genome Res.* **2005**, *15*, 1767.
(2) Swerdlow, H.; Gesteland, R. *Nucleic Acids Res.* **1990**, *18*, 1415.
(3) Luckey, J. A.; Drossman, H.; Kostichka, A. J.; Mead, D. A.; D'Cunha, J.; Norris, T. B.; Smith, L. M. *Nucleic Acids Res.* **1990**, *18*, 4417.
(4) Cohen, A. S.; Najarian, D. R.; Karger, B. L. *J. Chromatogr.* **1990**, *516*, 49.
(5) Huang, X. C.; Quesada, M. A.; Mathies, R. A. *Anal. Chem.* **1992**, *64*, 2149.
(6) Huang, X. C.; Stuart, S. G.; Bente, P. F., 3rd; Brennan, T. M. *J. Chromatogr.* **1992**, *600*, 289.
(7) Kheterpal, I.; Mathies, R. A. *Anal. Chem.* **1999**, *71*, 31A.
(8) Woolley, A. T.; Mathies, R. A. *Anal. Chem.* **1995**, *67*, 3676.
(9) Margulies, M.; Egholm, M.; Altman, W. E.; Attiya, S.; Bader, J. S.; Bemben, L. A.; Berka, J.; Braverman, M. S.; Chen, Y. J.; Chen, Z.; Dewell, S. B.; Du, L.; Fierro, J. M.; Gomes, X. V.; Godwin, B. C.; He, W.; Helgesen, S.; Ho, C. H.; Irzyk, G. P.; Jando, S. C.; Alenquer, M. L.; Jarvie, T. P.; Jirage, K. B.; Kim, J. B.; Knight, J. R.; Lanza, J. R.; Leamon, J. H.; Lefkowitz, S. M.; Lei, M.; Li, J.; Lohman, K. L.; Lu, H.; Makhijani, V. B.; McDade, K. E.; McKenna, M. P.; Myers, E. W.; Nickerson, E.; Nobile, J. R.; Plant, R.; Puc, B. P.; Ronan, M. T.; Roth, G. T.; Sarkis, G. J.; Simons, J. F.; Simpson, J. W.; Srinivasan, M.; Tartaro, K. R.; Tomasz, A.; Vogt, K. A.; Volkmer, G. A.; Wang, S. H.; Wang, Y.; Weiner, M. P.; Yu, P.; Begley, R. F.; Rothberg, J. M. *Nature* **2005**, *437*, 376.
(10) Leamon, J. H.; Lee, W. L.; Tartaro, K. R.; Lanza, J. R.; Sarkis, G. J.; deWinter, A. D.; Berka, J.; Weiner, M.; Rothberg, J. M.; Lohman, K. L. *Electrophoresis* **2003**, *24*, 3769.
(11) Ronaghi, M. Pyrosequencing: A tool for sequence-based DNA analysis. Doctoral Thesis, The Royal Institute of Technology, Stockholm, Sweden, 1998.
(12) Ronaghi, M.; Karamohamed, S.; Pettersson, B.; Uhlen, M.; Nyren, P. *Anal. Biochem.* **1996**, *242*, 84.
(13) Ronaghi, M.; Uhlen, M.; Nyren, P. *Science* **1998**, *281*, 363.
(14) Zhou, G.; Kamahori, M.; Okano, K.; Chuan, G.; Harada, K.; Kambara, H. *Nucleic Acids Res.* **2001**, *29*, E93.
(15) Andersson, H.; van der Wijngaart, W.; Stemme, G. *Electrophoresis* **2001**, *22*, 249.
(16) Ahmadian, A.; Russom, A.; Andersson, H.; Uhlen, M.; Stemme, G.; Nilsson, P. *Biotechniques* **2002**, *32*, 748.
(17) Russom, A.; Tooke, N.; Andersson, H.; Stemme, G. *J. Chromatogr., A* **2003**, *1014*, 37.
(18) Russom, A.; Tooke, N.; Andersson, H.; Stemme, G. *Anal. Chem.* **2005**, *77*, 7505.
(19) Ahmadian, A.; Ehn, M.; Hober, S. *Clin. Chim. Acta* **2006**, *363*, 83.
(20) Ronaghi, M. *Genome Res.* **2001**, *11*, 3.
(21) Melamede, R. J. Automatable process for sequencing nucleotide. US Patent 4,863,849, Sept 5, 1989.
(22) Ronaghi, M.; Nygren, M.; Lundeberg, J.; Nyren, P. *Anal. Biochem.* **1999**, *267*, 65.
(23) Nyren, P. *Anal. Biochem.* **1987**, *167*, 235.
(24) Hyman, E. D. *Anal. Biochem.* **1988**, *174*, 423.
(25) Nyren, P.; Lundin, A. *Anal. Biochem.* **1985**, *151*, 504.
(26) Ehn, M.; Ahmadian, A.; Nilsson, P.; Lundeberg, J.; Hober, S. *Electrophoresis* **2002**, *23*, 3289.
(27) Eriksson, J.; Gharizadeh, B.; Nourizad, N.; Nyren, P. *Nucleosides, Nucleotides Nucleic Acids* **2004**, *23*, 1583.
(28) Gharizadeh, B.; Nordstrom, T.; Ahmadian, A.; Ronaghi, M.; Nyren, P. *Anal. Biochem.* **2002**, *301*, 82.
(29) Eriksson, J. Advancements in Firefly Luciferase-Based Assays and Pyrosequencing Technology. Doctoral Thesis, Royal Institute of Technology, KTH, Stockholm, Sweden, 2004.
(30) Eriksson, J.; Gharizadeh, B.; Nordstrom, T.; Nyren, P. *Electrophoresis* **2004**, *25*, 20.
(31) Karamohamed, S.; Nordstrom, T.; Nyren, P. *Biotechniques* **1999**, *26*, 728.

(32) Agah, A.; Aghajan, M.; Mashayekhi, F.; Amini, S.; Davis, R. W.; Plummer, J. D.; Ronaghi, M.; Griffin, P. B. *Nucleic Acids Res.* **2004**, *32*, e166.
(33) Gharizadeh, B.; Eriksson, J.; Nourizad, N.; Nordstrom, T.; Nyren, P. *Anal. Biochem.* **2004**, *330*, 272.
(34) Zhou, G.; Kajiyama, T.; Gotou, M.; Kishimoto, A.; Suzuki, S.; Kambara, H. *Anal. Chem.* **2006**, *78*, 4482.
(35) Ehn, M.; Nourizad, N.; Bergstrom, K.; Ahmadian, A.; Nyren, P.; Lundeberg, J.; Hober, S. *Anal. Biochem.* **2004**, *329*, 11.
(36) Takala, S. L.; Escalante, A. A.; Branch, O. H.; Kariuki, S.; Biswas, S.; Chaiyaroj, S. C.; Lal, A. A. *Infect., Genet. Evol.* **2006**, *6*, 417.
(37) Fang, M.; Andersson, L. *Proc. Biol. Sci.* **2006**, *273*, 1803.
(38) Ahmadian, A.; Gharizadeh, B.; Gustafsson, A. C.; Sterky, F.; Nyren, P.; Uhlen, M.; Lundeberg, J. *Anal. Biochem.* **2000**, *280*, 103.
(39) Damaraju, S.; Murray, D.; Dufour, J.; Carandang, D.; Myrehaug, S.; Fallone, G.; Field, C.; Greiner, R.; Hanson, J.; Cass, C. E.; Parliament, M. *Clin. Cancer Res.* **2006**, *12*, 2545.
(40) Dupont, J. M.; Tost, J.; Jammes, H.; Gut, I. G. *Anal. Biochem.* **2004**, *333*, 119.
(41) Uhlmann, K.; Brinckmann, A.; Toliat, M. R.; Ritter, H.; Nurnberg, P. *Electrophoresis* **2002**, *23*, 4072.
(42) van der Straaten, T.; Kweekel, D.; Tiller, M.; Bogaartz, J.; Guchelaar, H. J. *J. Mol. Diagn.* **2006**, *8*, 444.
(43) Jonasson, J.; Olofsson, M.; Monstein, H. J. *APMIS* **2002**, *110*, 263.
(44) Nilsson, I.; Shabo, I.; Svanvik, J.; Monstein, H. J. *Helicobacter* **2005**, *10*, 592.
(45) Pourmand, N.; Elahi, E.; Davis, R. W.; Ronaghi, M. *Nucleic Acids Res.* **2002**, *30*, e31.
(46) Moore, G. E. *Electronics* **1965**, *38*.
(47) Nyren, P.; Pettersson, B.; Uhlen, M. *Anal. Biochem.* **1993**, *208*, 171.
(48) Tawfik, D. S.; Griffiths, A. D. *Nat. Biotechnol.* **1998**, *16*, 652.
(49) Ghadessy, F. J.; Ong, J. L.; Holliger, P. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 4552.
(50) Williams, R.; Peisajovich, S. G.; Miller, O. J.; Magdassi, S.; Tawfik, D. S.; Griffiths, A. D. *Nat. Methods* **2006**, *3*, 545.
(51) Kwok, S.; Kellogg, D. E.; Spasic, D.; Goda, L.; Levenson, C.; Sninsky, J. J. *Nucleic Acids Res.* **1990**, *18*, 999.
(52) Jordan, B.; Charest, A.; Dowd, J. F.; Blumenstiel, J. P.; Yeh Rf, R. F.; Osman, A.; Housman, D. E.; Landers, J. E. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 2942.
(53) Benita, Y.; Oosting, R. S.; Lok, M. C.; Wise, M. J.; Humphery-Smith, I. *Nucleic Acids Res.* **2003**, *31*, e99.
(54) Goldberg, S. M.; Johnson, J.; Busam, D.; Feldblyum, T.; Ferriera, S.; Friedman, R.; Halpern, A.; Khouri, H.; Kravitz, S. A.; Lauro, F. M.; Li, K.; Rogers, Y. H.; Strausberg, R.; Sutton, G.; Tallon, L.; Thomas, T.; Venter, E.; Frazier, M.; Venter, J. C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11240.
(55) Aliotta, J. M.; Pelletier, J. J.; Ware, J. L.; Moran, L. S.; Benner, J. S.; Kong, H. *Genet. Anal.* **1996**, *12*, 185.
(56) Thomas, D.; Nardone, G.; Randall, S. *Arch. Pathol. Lab. Med.* **1999**, *123*, 1170.
(57) Viguera, E.; Canceill, D.; Ehrlich, S. D. *EMBO J.* **2001**, *20*, 2587.
(58) Lage, J. M.; Leamon, J. H.; Pejovic, T.; Hamann, S.; Lacey, M.; Dillon, D.; Segraves, R.; Vossbrinck, B.; Gonzalez, A.; Pinkel, D.; Albertson, D. G.; Costa, J.; Lizardi, P. M. *Genome Res.* **2003**, *13*, 294.
(59) Mead, D. A.; McClary, J. A.; Luckey, J. A.; Kostichka, A. J.; Witney, F. R.; Smith, L. M. *BioTechniques* **1991**, *11*, 76.
(60) Lizardi, P. M.; Huang, X.; Zhu, Z.; Bray-Ward, P.; Thomas, D. C.; Ward, D. C. *Nat. Genet.* **1998**, *19*, 225.
(61) Keith, J. M.; Cochran, D. A. E.; Lala, G. H.; Adams, P.; Bryant, D.; Mitchelson, K. R. *Nucleic Acids Res.* **2004**, *32*, e35.
(62) Thomas, R. K.; Nickerson, E.; Simons, J. F.; Janne, P. A.; Tengs, T.; Yuza, Y.; Garraway, L. A.; LaFramboise, T.; Lee, J. C.; Shah, K.; O'Neill, K.; Sasaki, H.; Lindeman, N.; Wong, K. K.; Borras, A. M.; Gutmann, E. J.; Dragnev, K. H.; DeBiasi, R.; Chen, T. H.; Glatt, K. A.; Greulich, H.; Desany, B.; Lubeski, C. K.; Brockman, W.; Alvarez, P.; Hutchison, S. K.; Leamon, J. H.; Ronan, M. T.; Turenchalk, G. S.; Egholm, M.; Sellers, W. R.; Rothberg, J. M.; Meyerson, M. *Nat. Med.* **2006**, *12*, 852.
(63) Oki, Y.; Jelinek, J.; Beran, M.; Verstovsek, S.; Kantarjian, H. M.; Issa, J. P. *Haematologica* **2006**, *91*, 1147.
(64) Jelinek, J.; Oki, Y.; Gharibyan, V.; Bueso-Ramos, C.; Prchal, J. T.; Verstovsek, S.; Beran, M.; Estey, E.; Kantarjian, H. M.; Issa, J. P. *Blood* **2005**, *106*, 3370.
(65) Andries, K.; Verhasselt, P.; Guillemont, J.; Gohlmann, H. W.; Neefs, J. M.; Winkler, H.; Van Gestel, J.; Timmerman, P.; Zhu, M.; Lee, E.; Williams, P.; de Chaffoy, D.; Huitric, E.; Hoffner, S.; Cambau, E.; Truffot-Pernot, C.; Lounis, N.; Jarlier, V. *Science* **2005**, *307*, 223.
(66) Leamon, J. H.; Braverman, M. S.; Rothberg, J. M. *Gene Ther. Regul.* **2007**, *Mar* (3), 15.
(67) Ewing, B.; Hillier, L.; Wendl, M. C.; Green, P. *Genome Res.* **1998**, *8*, 175.
(68) Jarvie, T. *Drug Discovery Today: Technol.* **2005**, *2*, 255.

CR068297S