# Applying Rapid Genome Sequencing Technologies To Characterize Pathogen Genomes

ARTICLE *in* ANALYTICAL CHEMISTRY · MARCH 2008

Impact Factor: 5.64 · DOI: 10.1021/ac086027z · Source: PubMed

David Okou
Emory University
**20** PUBLICATIONS **605** CITATIONS

Michael E Zwick
Emory University
**58** PUBLICATIONS **1,847** CITATIONS

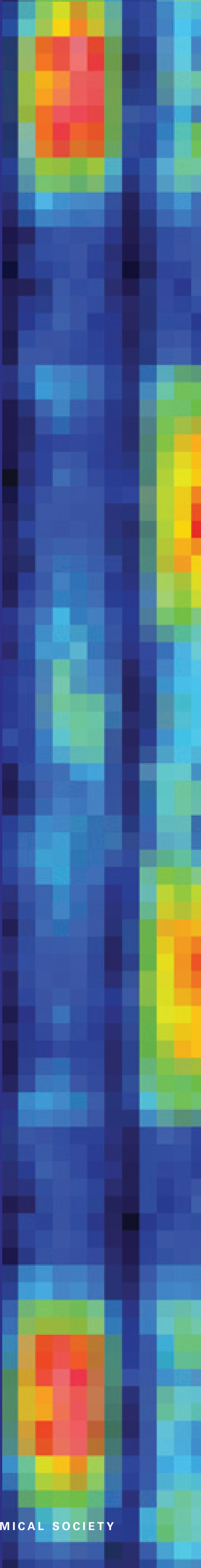
# Applying **Rapid Genome Sequencing** Technologies To Characterize **Pathogen** Genomes

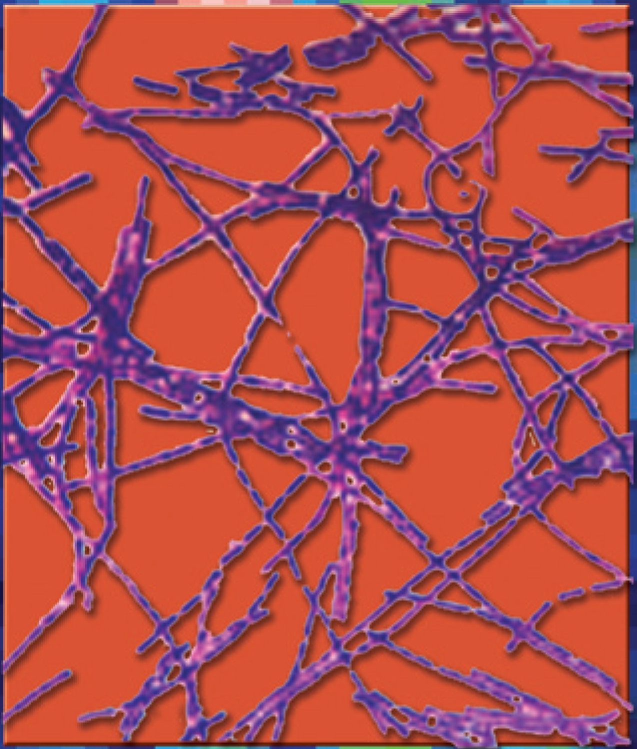
*Innovations in DNA sequencing shed light on pathogen genomes.*

Karyn Meltz Steinberg
David T. Okou
Michael E. Zwick
EMORY UNIVERSITY

Recent predictions of the worldwide burden of diseases under a variety of scenarios suggest that infectious diseases will continue to negatively impact the health of human populations well into the future (*1*). Although economic, social, and technological developments will surely influence these projections, the impact of the constantly changing interaction between hosts and infectious agents is rarely considered. Throughout evolutionary history, human populations were selected to develop immune responses to microbial challenges. The resulting human adaptations then caused pathogen populations to develop resistance via mechanisms such as antigenic variation to evade the immune system. The human host, in turn, counteradapted to the pathogens, not only with genetic changes but also with technologies, such as vaccines and therapeutics, that better recognize and combat the more virulent pathogens. This pattern of host–pathogen coevolution, where changes in allele frequencies in one population exert selection pressures affecting allele frequencies in the other population, creates a complex evolutionary arms race whose course may be difficult to predict (*2*).

Analyzing pathogen and host genomes can elucidate the mechanisms underlying novel adaptations that arise as a consequence of this ongoing evolutionary arms race. Pathogen genomes, because of their relatively small size, were the first to be completely sequenced and have provided researchers fundamental insights into the biology of the organisms, evolutionary relationships, and the determinants of virulence (*3*).





CDC

CDC/LARRY STAUFFER

CDC/RICHARD FACKLAM

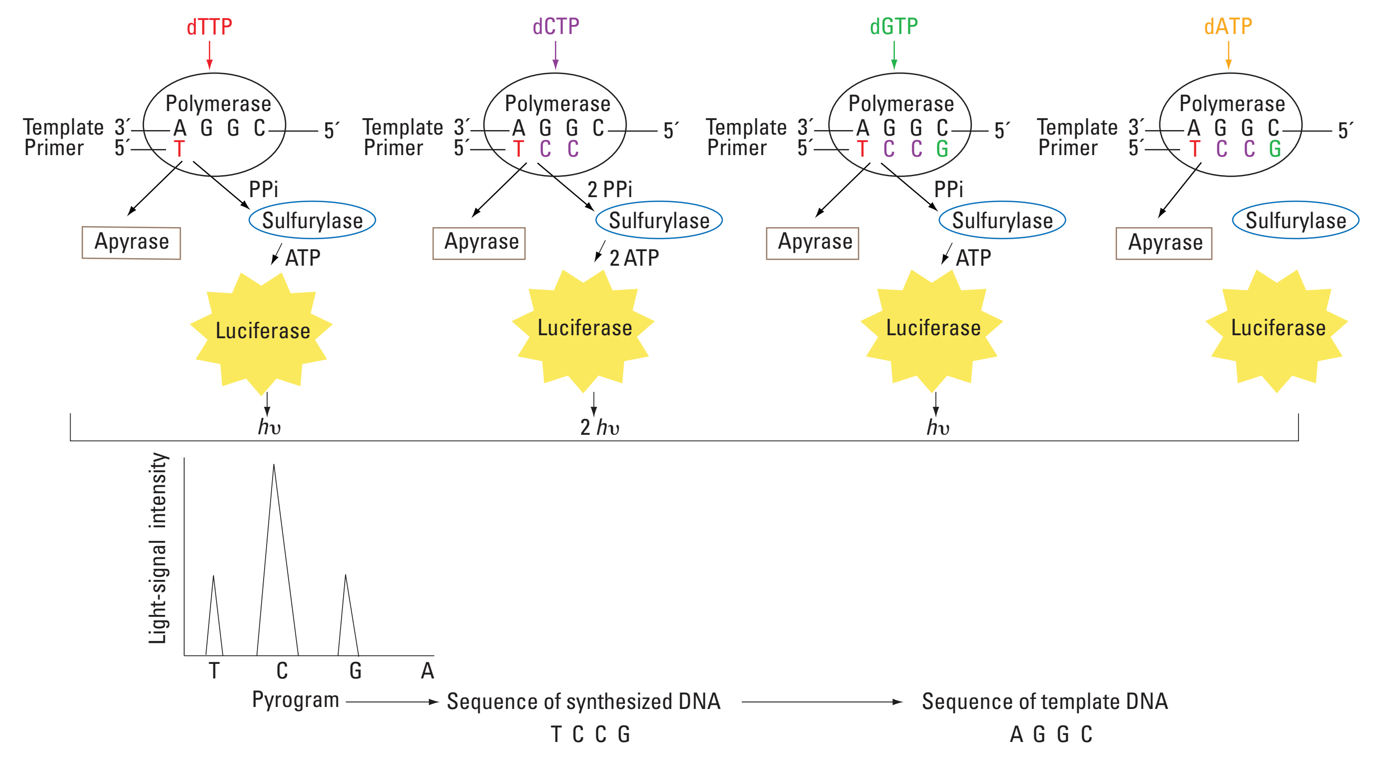


**FIGURE 1.** Determining a fragment sequence by pyrosequencing.

In the enzyme reactions, each dNTP is added individually in time; if an added dNTP forms a base pair with the template DNA, it is incorporated into the extending DNA strand by Klenow polymerase. This reaction releases pyrophosphate (PPi), and then ATP sulfurylase converts PPi into ATP, which is a substrate for luciferase. Luciferase production leads to a light signal, which is evidence that the nucleotide was incorporated. Apyrase then removes the unincorporated dNTPs. The light-signal intensities are recorded in a pyrogram, which provides the sequence of the synthesized DNA. The sequence of the template DNA is the complement to the synthesized sequence. (Adapted with permission from Ref. 22.)

Initial microbial projects focused on the de novo sequencing of a single representative of a specific pathogen species to generate a reference genome sequence. Given the vast population sizes of pathogens and the global distribution of many pathogens, it is becoming increasingly clear that if we are to truly dissect the complex biology of host–pathogen coevolution, sampling and sequencing of multiple genomes for various pathogens will be important (*4*). Furthermore, rapid pathogen-detection assays will become increasingly sequence-based as DNA sequencing technologies become less expensive and easier to use. For example, during outbreaks, complete genome sequencing could allow epidemiologists to more accurately type pathogens, track the evolution of antibiotic resistance, and identify novel pathogens, whether they are naturally occurring or bioengineered.

The process of sequencing a genome can be divided into four steps—break a large DNA polymer into smaller fragments; isolate and amplify single fragments; determine the fragment sequence; and assess data quality and assemble the sequence to reconstruct the original DNA polymer sequence. During the past decade, the implementation of Sanger sequencing chemistry and gel electrophoresis in capillaries have enabled large industrial genome-sequencing centers to automate these steps and increase throughput 50-fold while reducing costs 100-fold (*5*). This technical achievement has been remarkable.

Today, we stand on the cusp of a revolution in DNA sequencing. Novel chemistries that offer drastic cost reductions, increased data production, and high accuracy are now available in single instruments that require far fewer people and less laboratory space to operate (*6*). Collectively, these recent innovations raise the question of whether the traditional industrial genome-sequencing model is at the end of its utility.

In this article, we will review Sanger sequencing chemistry and how it has been applied to the sequencing of pathogen genomes. We will then describe three recently developed sequencing technologies that have been applied to the study of pathogens. The prospect of a genome-sequencing center in every lab is now within reach, and we can expect that this revolution will significantly alter our understanding of pathogens and their interactions with their hosts (*7*).

## Sanger sequencing

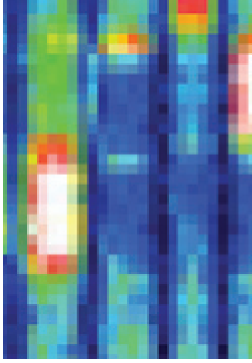
In traditional Sanger sequencing, the genome is fragmented and clonal libraries are produced to isolate and amplify single fragments. Determining the fragment sequence with Sanger chemistry, also known as dideoxy sequencing, involves using deoxynucleotides (dNTPs) to synthesize a DNA strand that is complementary to the template. When DNA polymerase incorporates unnatural 2′,3′-dideoxynucleotides (ddNTPs), synthesis is terminated (*8*). Fragments of various lengths are produced and separated by gel electrophoresis inside capillaries. This process has been automated by tagging each of the four ddNTPs with a different fluorescent dye. As labeled fragments pass through the DNA sequencer, the dye is excited by a laser, and the resulting fluorescence emission of one of the four colors is used for base-calling and sequence assembly (*9*). The integration of multiple capillary arrays per instrument has allowed the sequencing to proceed in parallel (*10*) with an output capability of up to 2 million bases during a 24 hour (h) period.

Automated data-quality assessment and sequence-assembly algorithms are then used to reconstruct the original genome sequence. Because large projects sequence millions to billions of bases, error rates must be exceptionally low. The Bermuda standard, the community-accepted quality level for finished genome sequencing, corresponds to <1 error per 10,000 bases sequenced (99.99% accuracy). For Sanger sequencing, single sequence reads may often fail to meet this stringent requirement. To achieve very high accuracy, multiple reads of the same base are often necessary. To determine the probability of a base-calling error, Phred quality scores are calculated for each sequenced base (Phred score = $-10\log_{10}$(error probability); *11*, *12*).

The Bermuda standard corresponds to a Phred score of 40. Because the initial library construction randomly fragments the DNA polymer, a genome must be oversampled to obtain sufficient coverage of all the bases. Achieving a quality score of 40 for an entire genome with Sanger technology typically requires a random 10-fold coverage that increases the cost. Currently, the cost is ~$0.001/sequenced base, and a complete draft sequence at 4-fold coverage (corresponding to a Phred score of 20, which is far below the Bermuda standard) costs ~$0.008/base (*13*).

Sanger sequencing, implemented on a large industrial scale, has been used successfully to sequence 478 microbial genomes, including *Streptococcus pneumoniae, Yersinia pestis, Neisseria meningitidis,* and *Bacillus anthracis* (*14–17*). An additional 778 microbial genomes are in the process of being sequenced (www.ncbi.nlm.nih.gov). These studies have provided the first reference genome sequences, allowing for the development of novel approaches for characterizing the numbers and locations of genes, predicted proteins, and pseudogenes. In addition, insertion sequences, deletions, and horizontally transferred genetic elements, such as bacteriophages and plasmids, were identified. Providing the sequences to the public allows researchers to compare the genomes of related pathogens, leading to a better understanding of pathogenesis and evolutionary history.

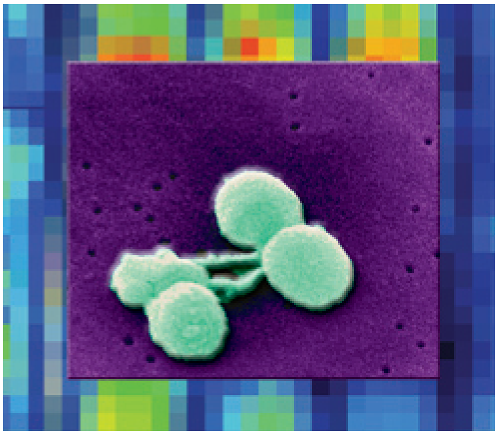
Although the industrial implementation of Sanger sequencing is a proven technology and is still considered the gold standard of DNA sequencing, the relatively high cost of obtaining the final sequence suggests that this methodology is not sufficiently economical for routine genome resequencing (*6*). Even a recent nanoliter-scale microfabricated processor that incorporates all of the steps involved in Sanger sequencing (sample amplification, sample purification, and CE), uses only 1 fmol of initial template, and generates read lengths of 425–550 bases with 99% accuracy is unlikely to reduce costs to the degree necessary (*18*). Furthermore, the vast infrastructure requirements and costs to establish a genome-sequencing center in the first place preclude the ability of individual laboratories to perform routine sequencing on this scale. As a consequence, the desire for ever-greater throughput is driving the development of other, more efficient technologies that may offer revolutionary reductions in costs.



## Pyrosequencing

Pyrosequencing is a sequencing-by-synthesis protocol in which the DNA template is first hybridized to a complementary sequencing primer. The DNA sequence of the fragment is then determined in a stepwise fashion through a series of four enzymatic reactions (Figure 1). First, the Klenow fragment of DNA polymerase adds a single nucleotide to the end of the sequencing primer. If the complementary base is added, the polymerase extends the primer, but if a noncomplementary base is encountered, the reaction pauses until the proper complementary base is added (*19–21*).

Second, inorganic pyrophosphate is released by Klenow polymerase and acts as a substrate for the adenosine triphosphate (ATP) sulfurylase that produces ATP. In the third reaction, ATP is converted to light by luciferase. In the fourth reaction, apyrase is used to remove the unincorporated nucleotides and ATP. The removal of nucleotides by apyrase is a necessary step for cyclic sequencing by synthesis. The second, third, and fourth reactions occur only if the appropriate nucleotide is added and incorporated into the complementary strand. The light signal produced by luciferase when a particular base is incorporated is detected with a photon detector and recorded on a pyrogram, a graph that shows the intensity of the light signal for each base added. Because the identity of the added nucleotide is known, the template sequence can be inferred by analyzing the pyrogram (Figure 1; *19*, *22*).



Pyrosequencing can be used for de novo sequencing, resequencing, genotyping, and sequence determination of difficult secondary DNA structures. This technique is currently one of the most rapid methods for sequencing a PCR product, although read lengths are often quite short, averaging ~100–200 bases (*6*, *19*). However, typing microbial strains requires only relatively short read lengths, making pyrosequencing more efficient than Sanger sequencing for this application (*22*).

Monstein and colleagues demonstrated that pyrosequencing could rap-

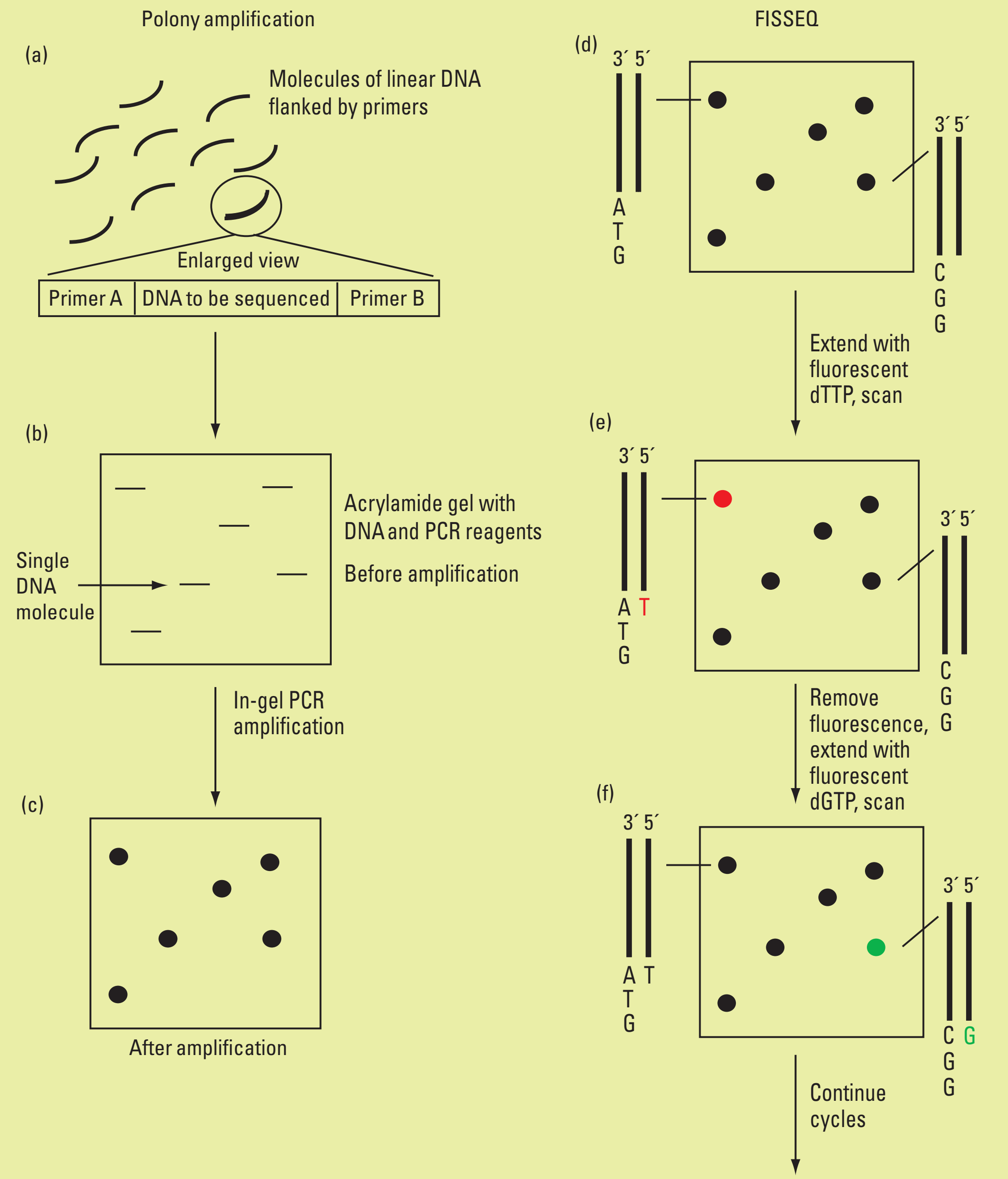


**FIGURE 2.** Polony amplification and FISSEQ.

(a) To amplify polonies, a library of linear DNA molecules with universal priming sites is PCR amplified (b) in a polyacrylamide gel. (c) The template molecule amplifies to a PCR colony. (d) For each polony, a universal primer is annealed to the primer binding sites. (e, f) The polony is sequenced by serial addition of single fluorescent dNTPs and primer extension. (Adapted with permission from Ref. 31.)

idly identify 23 clinical isolates of *Helicobacter pylori* from gastric biopsy samples with alleles from 16S ribosomal DNA (*23*). Subtle sequence variation was detected accurately, and the isolates were subtyped into different lineages for phylogenetic analysis. In a related study, pyrosequencing accurately identified all of the Gram-positive isolates and the aerobic Gram-negative pathogens from PCR products (*24*). Instead of extracting DNA, the researchers carried out PCR directly on picked bacterial colonies; this eliminated a time-consuming step necessary for other sequencing technologies. Gram-negative samples, however, required longer signatures and further analysis to type accurately. Pyrosequencing also has been applied to the study of antimicrobial resistance in pathogens. The sequencing of 40 clinical isolates of *Neisseria gonorrhoeae* detected a point mutation that decreased susceptibility to ciprofloxacin in the DNA gyrase gene *gyrA* (*25*). Of those isolates, all had at least one mutation in the *gyrA* gene.

Although the original pyrosequencing technology was not competitive with the Sanger method in time and cost-efficiency for sequencing whole genomes, the recent development of massively parallel pyrosequencing has allowed this technology to surpass Sanger sequencing (*26*). Recently, a method was introduced for whole-genome sequencing with a pyrosequencing chemistry (454 Corp.). The approach avoids traditional library construction and results in a substantial savings of time and cost. This technique shears an entire genome to generate 300-bp-long DNA fragments, ligates specialized common adapters to each fragment, captures individual fragments on beads, and then clonally amplifies each fragment within an oil emulsion. This particular methodology is highly scalable because of PicoTiterPlate technology (*27*), which can amplify 300,000 templates generated by clonal amplification of single DNA molecules that are bound to Sepharose beads.

Identification of the fragment sequence occurs as reagents are passed over the plate. The quality of the resulting sequences is determined, followed by assembly. Using this massively parallel sequencing platform, a researcher can sequence genomes up to 10 Mbp (*22*). It takes 1 day to prepare the DNA library, 1 day for emulsion PCR, and ~4.5 h for actual sequencing (www.454.com). The price per draft sequence base pair is $0.00015 for 10-fold coverage, which is 26.7× more cost-effective than Sanger sequencing (*13*). However, pyrosequencing has a few disadvantages. For example, read lengths are typically 100–200 bp, and obtaining high accuracy in calling homopolymeric sequences can be challenging.

Margulies et al. used this technology to sequence a *Mycoplasma genitalium* genome (*26*). They reported 96% coverage of the genome with an error rate of 3.0 × $10^{-6}$, which corresponds to a Phred score of 55. Combining the PicoTiterPlate with Sanger sequencing, Goldberg et al. sequenced six microbial genomes to fill in gaps created by a strong secondary structure and hard stops left by Sanger sequencing methods (*28*). The most striking features of combining pyrosequencing with PicoTiterPlate technology are the dramatic cost and infrastructure reductions that could enable single laboratories to perform whole-genome pathogen sequencing. With such increased availability, epidemiological tasks such as strain typing, identification, and tracking will become much more feasible.

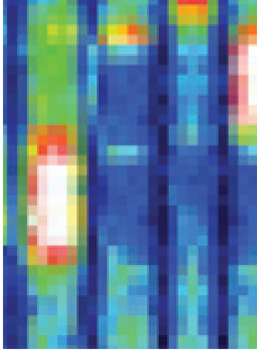
## Fluorescent in situ sequencing

In contrast to Sanger sequencing and pyrosequencing, fluorescent in situ sequencing (FISSEQ), developed by Mitra and Church, originally was intended to enable highly parallel sequencing of large collections of DNA fragments (*29*). This method uses a multiplex polymerase colony, or polony, protocol to generate a library of millions of fragments that are each 135 bp in length. Each fragment has 100 bp in common and contains two mate-pair tags of 17–18 bp; mate-pair tags are sequencing reads from opposite ends of each fragment. The tags consist of unique genome sequences selected to be ~1000 bp apart in the genome and are flanked by universal sequences that are complementary to the primers used in subsequent steps (*30*).

Although the **Sanger method** is the **gold standard of DNA** sequencing, **the high cost** suggests that it is not **economical for** routine **genome** resequencing.

This results in fragments that contain a variable region flanked by two constant regions (*31*). Each fragment is attached to a separate 1 µm bead, amplified with a water-in-oil emulsion PCR protocol, and immobilized in a 1.5 cm$^2$ acrylamide gel. Each bead contains thousands of single-stranded copies of the same PCR product. The fragments are then sequenced in parallel via a ligation protocol that uses four dyes to identify each possible base. The DNA is extended by adding one fluorescently labeled dNTP at a time until nucleotide incorporation can be measured by a fluorescence signal (*31*). This cycle is repeated with another base until one is incorporated successfully. At the end of each cycle, the fluorescence from the preceding reaction is bleached by removal of the fluorophore so that each polony is sequenced iteratively. The images generated from each reaction are compiled and evaluated by using automated software designed to compare the bright-field images (polony location) with the fluorescence images (base incorporation) for each reaction; then, the images are analyzed with a base-calling algorithm (Figure 2; *30*).

The multiplex polony sequencing strategy has been applied to the experimental evolution of *E. coli* samples grown under limited nutrition. These strains are predicted to have a limited number of differences from a known reference strain (*30*). The experiment generated 30.1 Mb of sequence over 26 cycles. Of 14 million fluorescence images, 1.6 million were highly accurate. High-confidence base calls were given for >70% of the *E. coli* genome at 4-fold coverage. Despite fluctuations in the genome coverage and PCR amplification errors, this method detected six differences from the reference sequence that were confirmed by the Sanger method. Although the read length was significantly lower than that of conventional sequencing (26 compared with ~700), the number of accurate sequences generated was astronomically higher than that of Sanger sequencing (1.6 million versus 96).



The advantages of FISSEQ are that it is inexpensive, avoids time-consuming library construction, and provides high consensus accuracies. With error rates estimated to be <1/3.3 million bases (Phred 65; *7*, *30*), FISSEQ can be used for resequencing or de novo sequencing and is a viable competitor of Sanger sequencing. However, the read lengths are short, and calling homopolymeric sequences is difficult (*31*). The problem of calling homopolymeric sequences may be solved with the use of reversibly terminating nucleotides (by Solexa) that help in the simultaneous incorporation of all four fluorophores, followed by cleavage (as opposed to FISSEQ where the fluorophores are added one at a time; *32*).

It costs ~$0.000006/base to sequence with reversibly terminating nucleotides; this process is one of the most cost-efficient technologies available at this time. However, it does have one drawback, because the read lengths are so short, if a fragment's sequence is found at various loci throughout the genome, that particular fragment can never be aligned properly with high confidence. In addition, the algorithms used for sequence assembly are unable to deal with variants due to insertions or deletions, even those indels as small as 1 base. Although this technology to sequence microbial genomes has not yet appeared in publication, it holds great promise for rapid and accurate sequencing that may prove invaluable to the fields of epidemiology, genetics, and clinical diagnostics.

## Sequencing by hybridization

In sequencing by hybridization, labeled DNA is differentially hybridized to oligonucleotide probes depending on the reference DNA template. Overlapping oligonucleotide probes that are 25 bp long are immobilized on a membrane or chip, called a resequencing array (RA). These probes are tiled at a 1 bp resolution. Each base has a total of eight features: four identi-

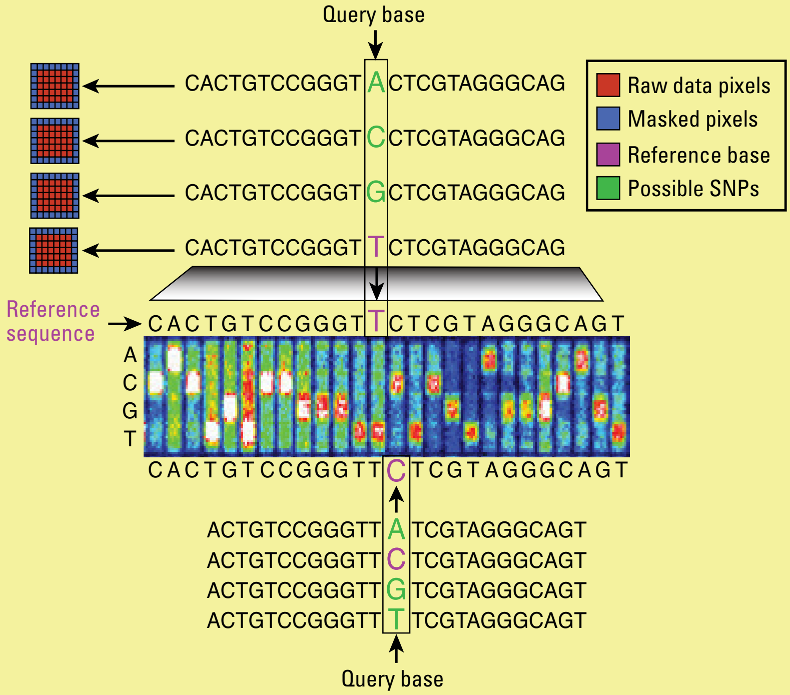

Query base
CACTGTCCGGGT A CTCGTAGGGCAG
CACTGTCCGGGT C CTCGTAGGGCAG
CACTGTCCGGGT G CTCGTAGGGCAG
CACTGTCCGGGT T CTCGTAGGGCAG
🟥 Raw data pixels
🟦 Masked pixels
🟪 Reference base
🟩 Possible SNPs
Reference sequence
CACTGTCCGGGT T CTCGTAGGGCAGT
A
C
G
T
CACTGTCCGGGTT C TCGTAGGGCAGT
ACTGTCCGGGTT A TCGTAGGGCAGT
ACTGTCCGGGTT C TCGTAGGGCAGT
ACTGTCCGGGTT G TCGTAGGGCAGT
ACTGTCCGGGTT T TCGTAGGGCAGT
Query base

**FIGURE 3.** Sequencing by hybridization.

Four forward strand and four reverse complement strand features are associated with every site in the reference sequence. A feature is a 25-base oligonucleotide, and the 13th base is the query base. Each feature contains $10^6$ copies of the unique oligonucleotide at a specific location on the RA. Labeled target DNA is hybridized to the RA overnight, allowing target fragments to diffuse and to locate and bind most strongly to features containing the complementary query base. The RA is then scanned to generate an image file that can be processed with ABACUS to determine the DNA sequence (*35*).

cal to the forward strand and four for the reverse complement strand. Position 13 is called the query base and contains either A, C, G, or T. Each feature contains ~1,000,000 copies of the particular oligonucleotide. Target DNA is isolated from genomic DNA by using long PCR for selected genomic regions or, alternatively, by using whole-genome amplification (WGA). Isolated target DNA is then fragmented by either sonication or DNAse I digestion. Fluorescently labeled sample DNA is then hybridized to the eight features, and two of these features (one forward and one reverse) that are complementary to the test sequence will provide the brightest signal. If the sample DNA is heterozygous at position 13, the two features with the appropriate complementary base will provide the highest signal (Figure 3).

One company (Affymetrix) produces high-density oligonucleotide RAs by using masked photolithography with ~2.4 million 25-mer oligos and 8 µm feature size (*33*). This enables the sequencing of ~300 Kb/chip. Another company (NimbleGen Systems) produces RAs with ~385,000 features and a 16 µm feature size (*34*) and is now producing chips with 2.1 million features with a 13 µm feature size. A flexible maskless photolithography technology allows variable probe length throughout the array; this can lead to a more uniform hybridization efficiency at each base.

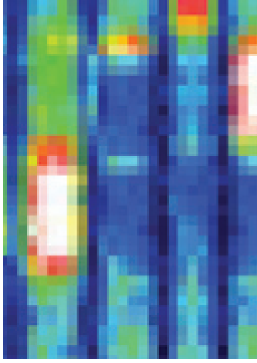
Most RA data are analyzed with the adaptive background genotype calling scheme (ABACUS) (*35*). ABACUS is a fully automated statistical algorithm that determines individual base/genotype calls with high accuracy regardless of the nature of the site. The algorithm uses likelihood models for each of the possible base calls that are tested independently for the forward and reverse strands. ABACUS then assigns a quality score based on the difference between the best-fitting model and the second-best-fitting model for each genotype. In the initial application of RAs with ABACUS, >99.9999% of >80% of the genotypes were called (*35*). Subsequent software improvements that automated grid alignment allowed highly accurate (99.9999%) calls to be made at >90% of resequenced bases. ABACUS and grid alignment software are publicly available in a package called RATools, available at www.dpgp.org/RA/ra.htm.

RAs have many different applications in microbial genomics. Comparing the genomes of various strains of the same pathogen, or of closely related species, can help researchers identify genes and provide valuable information about genome organization and evolution and insight into the virulence of strains. For example, RAs were used to resequence 56 *B. anthracis* strains (*36*). Of 118 total RAs, 115 were successful and called 92.6% of possible bases accurately. Base-call failure was attributed to amplification failure as well as probes with excess purine composition, especially guanine. When compared with conventional shotgun sequencing, single RAs performed as well as, and in some cases better than, multiple DNA sequencing reads.

Sequence analysis revealed that within *B. anthracis*, sequence variation is low. In addition, silent sites, where the protein structure is not altered, are much more variable than replacement sites, which alter the amino acid structure. In addition, an excess of rare variants (relative to a population in mutation-drift equilibrium) was observed in *B. anthracis*; this can be caused by either recent population expansion from a single clone or natural selection on variants. We suggest that because plasmid exchange has been minimal, recent population expansion from a single clonal ancestor is the most likely cause. This situation also has been proposed for *Y. pestis*, the causative agent of plague.

NimbleGen RAs were extremely useful during an outbreak of SARS-CoV (*37*). The array contained the entire 29.7 Kb genome and all known deletions and insertions. Resequencing 14 whole genomes from cell culture and from patients produced highly reproducible and accurate data comparable to that from conventional capillary sequencing. The SARS-CoV study highlights many of the advantages, including the speed at which multiple genomes of viral isolates can be sequenced, which make it an invaluable tool in epidemiology.

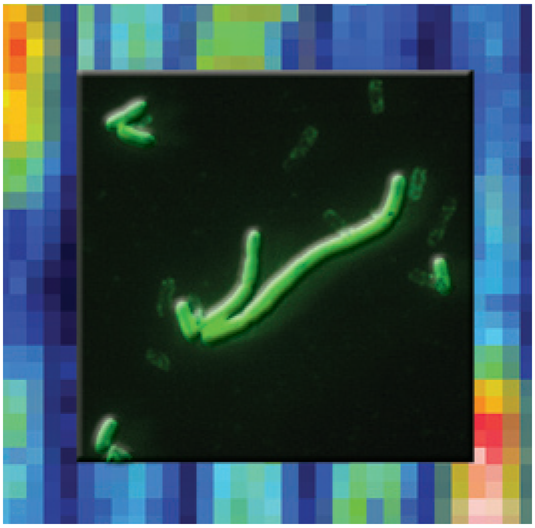
RAs are ideal for analyzing clinical samples when compared



with traditional methods that rely on multiple amplification steps and special laboratory setups. Lin et al. developed an RA called RPM v.1 that contains 57 genomic sequences from diagnostic regions of 26 different respiratory pathogens, including human adenovirus and influenza A virus (*38*). Clinical samples were then prepared by multiplex reverse transcriptase-PCR and hybridized to the chip. Results indicate that no significant cross-hybridization occurred between targets and that nearby genetic neighbors were reliably identified. The RA method could detect and type pathogens with high sensitivity and specificity from clinical samples with low titers (as low as 100 plaque forming units/mL) and samples co-infected with up to seven separate pathogens.

Sequence data were analyzed with an algorithm called ProSeq, which was developed specifically for multipathogen arrays (*39*). ProSeq references the sequence generated from the array and finds close matches in genetic databases. It then references taxonomic databases to build phylogenetic relationships between the database entries and the target sequence to properly identify the microorganism. In addition to typing pathogens, the RA provides information that can determine phylogenetic relationships among isolates, allowing researchers to track changes in rapidly mutating strains, such as influenza A. The major limitation of this RA is that the PCR primers for sample preparation must be selected for each new RA design because of the rapid mutation rate of RNA viruses during sample preparation. Also, because only a certain number of probe sets can be on the RA, a limited number of pathogens can be detected and a limited amount of information about each pathogen can be assessed at one time.

Wang et al. demonstrated that the RPM v.1 array is also effective in molecular epidemiological tracking of influenza A and B (*40*). This chip could correctly identify unique polymorphisms in a sample of 25 geographically distinct influenza isolates collected from infected humans. Influenza viruses are subject to forces of genetic drift that can rapidly cause amino acid substitutions. Almost every isolate had unique base pair substitutions that the array detected reliably, proving that this platform can be used to effectively track genetic changes in circulating influenza strains.

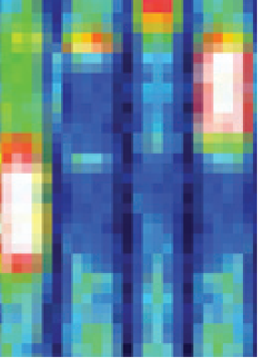


RAs can discern not just particular pathogens from a clinical sample but also important characteristics, such as antibiotic resistance profiles. Using the previously described respiratory pathogen RA RPM v.1, Davignon et al. could identify antibiotic markers from *Streptococcus pyogenes* samples (*38*, *41*). *S. pyogenes* is a community-associated pathogen that is rapidly acquiring resistance to many antimicrobials. The use of RAs is a cost-effective alternative to traditional typing methods. Another study found that microarrays were better than pulsed-field gel electrophoresis and variable number tandem repeat typing technologies at identifying antimicrobial resistance markers for the important, emerging pathogen methicillin-resistant *Staphylococcus aureus*, better known as MRSA (*42*). With an RA containing the entire bacterial chromosome at a resolution of 1 probe every 400–500 bp, the sequence variation within and between clusters of strains was identified. Community-onset MRSA clustered with major epidemic clones that had acquired virulence genes via horizontal gene transfer. The use of RAs to examine clusters of strains related to outbreaks may reveal patterns of gene transfer that are relevant to the treatment and containment of antibiotic-resistant pathogens.

RAs, like the other next-generation sequencing technologies, are highly parallel. An RA typically requires less DNA than other methods and does not require library construction for pathogen genomes. Furthermore, the pattern of hybridization as evidenced by the chip images can be directly translated into a DNA sequence with the use of ABACUS. Hence, no computationally intensive assembly step is needed. Single individuals in laboratories can generate extensive quantities of genome sequence rapidly with minimal equipment and infrastructure requirements. On the other hand, because RAs can only compare target DNA to complementary probes on the basis of a reference sequence, it is best not to use them for identifying novel sequences or sequence rearrangements that may occur in the experimental sample.

Furthermore, existing algorithms do not allow RAs to detect copy number variation (CNV) in diploid targets. The CNV class of variation, though rarer than the SNP, has nonetheless proved to be surprisingly frequent (*43*, *44*). A different class of tiled arrays, called array comparative genome hybridization chips, can detect CNVs. Although the number of features possible on an RA has grown rapidly, there are undoubtedly practical limits to the absolute number of oligos that can be synthesized on a single chip. Despite this limitation, for many genetic studies, resequencing large genomic regions, rather than entire genomes, may be sufficient to identify specific mutations. Coupled with an appropriate procedure for isolating target DNA, RAs could be very efficient at performing these types of studies.

Because of the small genome size of pathogens, isolation of target DNA is much simpler for pathogens than for vertebrates. Pathogen DNA can be isolated easily with an overnight WGA system, eliminating the time-consuming step of long PCR. Fragmentation, labeling, and hybridization steps can be accomplished in another 24 h period, followed by analysis. The entire RA process for a pathogen genome would take only



2–3 days; this would be crucial during an outbreak, when characterizing an infectious agent in a short period of time is essential. In addition, the cost of sequencing per base for RAs is $0.0001, which makes it feasible for a single lab to perform large-scale resequencing applications.

## Future directions

The pace of change in DNA sequencing continues to increase; one recent article suggested that 100-fold cost and throughput improvements would occur soon (*32*). Ultimately, comprehensively sequencing an entire pathogen genome provides vastly more information about the infectious agent as compared with conventional methods of pathogen identification. Tools such as the multiple respiratory pathogen RA may become so cost-efficient that they will replace conventional culturing and typing methods in hospital and outbreak settings (*38*). If sequencing costs continue to fall dramatically, as expected, then it seems clear that diagnostic assays and epidemiological outbreak investigations will increasingly embrace these genomics technologies to enable rapid, comprehensive pathogen characterization.

Whole-genome sequencing provides detailed data for the phylogenetic characterization of microbes, tracking the global spread of particular strains of pathogens, and other applications relevant to infectious-disease epidemiology. For example, the emergence of antimicrobial resistance has created a health care crisis. Genome sequencing provides a method by which both research scientists and clinicians can better understand the mechanisms of this resistance; it can lead to improved patient outcomes. Although conventionally culturing microbes can identify the drugs to which pathogens are resistant, sequencing can aid in the identification of the mechanism of resistance at the gene and nucleotide level.

This information can be used to develop new therapeutic targets as well as to track the evolution of resistance within a population. Furthermore, inexpensive DNA sequencing might also be used to sequence pathogens that cannot be cultured and hence are currently understudied. Finally, it is now realistic to expect that soon even relatively small clinical diagnostic laboratories can use these next-generation technologies to sequence pathogen genomes with the throughput only previously seen in large genome-sequencing centers. By identifying novel variation and adaptations within pathogen populations as they occur, humans can stay one step ahead of microbes in this evolutionary arms race.

*Karyn Meltz Steinberg is a graduate student, David T. Okou is a postdoctoral fellow, and Michael E. Zwick is an assistant professor at Emory University. Steinberg's research interests include population genetics, genetic susceptibility to infectious disease, genomics, and resequencing. Okou works on human genetics, genomics, and resequencing. Zwick's research interests include human and model system genetics, population genetics, genomics, and resequencing. Address correspondence about this article to Zwick at Department of Human Genetics, 615 Michael St., Suite 301, Atlanta, GA 30322 (mzwick@genetics.emory.edu).*

## References

(1) Mathers, C. D.; Loncar, D. *PLoS Med.* **2006**, *3* (11), e442.
(2) Woolhouse, M. E.; et al. *Nat. Genet.* **2002**, *32*, 569–577.
(3) Subramanian, G.; et al. *Mol. Diagn.* **2001**, *6*, 243–252.
(4) Venter, J. C.; et al. *Science* **2004**, *304*, 66–74.
(5) Collins, F. S.; Morgan, M.; Patrinos, A. *Science* **2003**, *300*, 286–290.
(6) Shendure, J.; et al. *Nat. Rev. Genet.* **2004**, *5*, 335–344.
(7) Zwick, M. E. *Eur. J. Hum. Genet.* **2005**, *13*, 1167–1168.
(8) Sanger, F.; Nicklen, S.; Coulson, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 5463–5467.
(9) Smith, L. M.; et al. *Nature* **1986**, *321*, 674–679.
(10) Paegel, B. M.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 574–579.
(11) Ewing, B.; et al. *Genome Res.* **1998**, *8*, 175–185.
(12) Ewing, B.; Green, P. *Genome Res.* **1998**, *8*, 186–194.
(13) Chan, E. Y. *Mutat. Res.* **2005**, *573*, 13–40.
(14) Tettelin, H.; et al. *Science* **2001**, *293*, 498–506.
(15) Parkhill, J.; et al. *Nature* **2001**, *413*, 523–527.
(16) Parkhill, J.; et al. *Nature* **2000**, *404*, 502–506.
(17) Read, T. D.; et al. *Nature* **2003**, *423*, 81–86.
(18) Blazej, R. G.; Kumaresan, P.; Mathies, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 7240–7245.
(19) Ronaghi, M. *Genome Res.* **2001**, *11*, 3–11.
(20) Ronaghi, M.; et al. *Anal. Biochem.* **1996**, *242*, 84–89.
(21) Ronaghi, M.; Uhlén, M.; Nyrén, P. *Science* **1998**, *281*, 363, 365.
(22) Ahmadian, A.; Ehn, M.; Hober, S. *Clin. Chim. Acta* **2006**, *363*, 83–94.
(23) Monstein, H.; Nikpour-Badr, S.; Jonasson, J. *FEMS Microbiol. Lett.* **2001**, *199*, 103–107.
(24) Jonasson, J.; Olofsson, M.; Monstein, H. J. *APMIS* **2002**, *110*, 263–272.
(25) Gharizadeh, B.; et al. *Int. J. Antimicrob. Agents* **2005**, *26*, 486–490.
(26) Margulies, M.; et al. *Nature* **2005**, *437*, 376–380.
(27) Leamon, J. H.; et al. *Electrophoresis* **2003**, *24*, 3769–3777.
(28) Goldberg, S. M.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11,240–11,245.
(29) Mitra, R. D.; Church, G. M. *Nucleic Acids Res.* **1999**, *27* (24), e34.
(30) Shendure, J.; et al. *Science* **2005**, *309*, 1728–1732.
(31) Mitra, R. D.; et al. *Anal. Biochem.* **2003**, *320*, 55–65.
(32) Bentley, D. R. *Curr. Opin. Genet. Dev.* **2006**, *16*, 545–552.
(33) Lipshutz, R. J.; et al. *Nat. Genet.* **1999**, *21* (1 Suppl), 20–24.
(34) Albert, T. J.; et al. *Nucleic Acids Res.* **2003**, *31* (7), e35.
(35) Cutler, D. J.; et al. *Genome Res.* **2001**, *11*, 1913–1925.
(36) Zwick, M. E.; et al. *Genome Biol.* **2005**, *6*, R10.
(37) Wong, C. W.; et al. *Genome Res.* **2004**, *14*, 398–405.
(38) Lin, B.; et al. *J. Clin. Microbiol.* **2007**, *45*, 443–452.
(39) Malanoski, A. P.; et al. *Nucleic Acids Res.* **2006**, *34*, 5300–5311.
(40) Wang, Z.; et al. *Emerg. Infect. Dis.* **2006**, *12*, 638–646.
(41) Davignon, L.; et al. *J. Clin. Microbiol.* **2005**, *43*, 5690–5695.
(42) Koessler, T.; et al. *J. Clin. Microbiol.* **2006**, *44*, 1040–1048.
(43) Wong, K. K.; et al. *Am. J. Hum. Genet.* **2007**, *80*, 91–104.
(44) Sharp, A. J.; et al. *Am. J. Hum. Genet.* **2005**, *77*, 78–88.