REVIEW 10.1111/j.1469-0691.2009.02962.x

How will next-generation sequencing contribute to the knowledge concerning *Helicobacter pylori*?

L. Engstrand

Department of Bacteriology, Swedish Institute for Infectious Disease Control & Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden

Abstract

Molecular microbiology has revolutionized the landscape of microbiology and will continue to do so by providing new solutions for microbe identification and characterization. This applies also to the study of *Helicobacter pylori* where current genotypic (molecular) methods are important complements or alternatives to phenotypic methods. Besides providing sensitivity and specificity and an enhancement of the detection process, they also reduce much of the subjectivity inherent in the interpretation of morphological and biological data. Another key advantage of molecular methods is that they allow the identification of novel virulence factors of pathogenic bacteria. For example, such gene products enable *H. pylori* to establish itself within the gastric environment and enhance its potential to cause disease. Next-generation sequencing will open up new areas of research for those involved in the field of *Helicobacter* research and will also provide information that will help to develop novel treatment strategies and increase our understanding of the mechanisms behind chronic inflammations in the gut. The analysis of data resulting from a large-scale sequencing project requires the use of bioinformatics, including standard BLAST analysis, annotation or clustering, and assembly competence. However, the amount of data produced by the next-generation sequencing platforms will require a bioinformatics capacity at the industrial scale, which may limit the availability of such technologies. Consequently, building effective new approaches to data analysis must be given high priority.

Keywords: Bioinformatics, data analysis, *Helicobacter pylori*, next-generation sequencing, technology platforms

Clin Microbiol Infect 2009; **15:** 823–828

Corresponding author and reprint requests: L. Engstrand, Department of Bacteriology, Swedish Institute for Infectious Disease Control & Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden
E-mail: lars.engstrand@smi.se

Technology Platforms

Growing demand in both the research and clinical markets has fueled the development of more efficient genomic sequencing methods [1,2]. Such methods are already several orders of magnitude more efficient than the Sanger capillary-array electrophoresis machines that were used in the human genome project. Massively parallel DNA sequencing platforms have not only reduced the cost of DNA sequencing, but also have moved the technology from major genome centres to individual investigators. The new platforms will dramatically accelerate biological and biomedical research, by enabling the comprehensive analysis of genomes to become inexpensive, routine and widespread. Below, three commercial systems are briefly described (Table 1) and each of them has the potential to contribute to our knowledge concerning *Helicobacter pylori*.

Roche/454 FLX pyrosequencer

Multiple whole prokaryote genomes can easily be sequenced using the 454 FLX system (Roche Diagnostics, Basel, Switzerland.) [3,4]. This high-throughput technology that sequences in real time provides long reads (400 bp) that facilitate the completion of near-finished draft sequences in a single instrument run. Large-size genomic DNA samples are randomly fragmented into small 300- to 800-bp fragments for shotgun sequencing. Addition of adapters to the fragments creates a library of DNA fragments, which is immobilized on DNA capture beads, whereafter PCR amplification takes place in water-in-oil microreactors, resulting in millions of copies of the template. Finally, the microreactor is broken and beads

TABLE 1. Comparison of next-generation sequencing technologies

	454 FLX	Solexa	SOLiD
Read length	250–400 bp	25–35 bp	25–35 bp
Reads	1 M	30 M	90 M
Data	400 Mb	3 Gb	30 Gb
Scale-up of number of reads	+	+++	+++
Future increase of read length	++	+	+
Access to instruments	+++	+++	+++
Drawbacks	High error rate for homopolymers	Error rate increases with read length	Error rate increases with read length
Advantages	Long read length	Easy to scale up	Easy to scale up

carrying single-stranded DNA templates are individually sequenced on a picotitre plate device. The generated sequences are then assembled into a number of unordered contigs using specific assembler software. Finally, a consensus sequence is generated.

The sequencing depth achieved with 454 FLX titanium sequencing systems ensures the accurate characterization of microbial or bacterial diversity, the sensitive detection of even rare mutations, and the rapid discovery of the disease-causing agents [5]. Furthermore, this system for ultra-high-throughput DNA sequencing is used for *de novo* sequencing and resequencing of genomes, for metagenomics, and for targeted sequencing of DNA regions of interest. The newest version generates up to 400 million bases per 10-h instrument run. The key advantage of this technology is read-length (up to 400 bp, which is necessary in *de novo* assembly and metagenomics). However, a major limitation is that no prevention of multiple incorporations at a given cycle is provided, which leads to homopolymer errors.

The technology has enabled a number of peer-reviewed studies in diverse research fields, such as cancer and infectious diseases, drug discovery, marine biology, anthropology, paleontology, and many more. The value of the FLX System for bacterial sequencing applications is emphasized by a number of important studies, including a study of *Mycobacterium tuberculosis* that resulted in the identification of the first tuberculosis-specific drug candidate in 40 years [6]. FLX System pyrosequencing has so far been the method of choice for sequencing of *H. pylori* and for exploring the human stomach microbiota [7,8].

Illumina/Solexa genome analyzer

Illumina sequencing technology (Illumina Inc., San Diego, CA, USA), or the Solexa platform, allows for the selection of any single-nucleotide polymorphism or probe, enabling dense, uniform coverage across the genome and the ability to target any genomic region [9,10]. This platform is based on massively parallel sequencing of millions of fragments using a reversible terminator-based sequencing chemistry. The technology, together with a software application, allows a scalable system that many consider cost-effective and accurate. It relies on the attachment of randomly fragmented genomic DNA to an optically transparent surface. Attached DNA fragments are extended and subjected to bridge amplification to create an ultra-high-density sequencing flow cell with ≥50 million clusters, each containing approximately 1000 copies of the same template. These templates are sequenced using a four-colour DNA sequencing-by-synthesis technology that employs reversible terminators with removable fluorescent dyes. This approach ensures high accuracy and true base-by-base sequencing, eliminating sequence context-specific errors and enabling sequencing through repetitive sequences. After completion of the first read, the templates can be regenerated *in situ* to enable a second >36-bp read from the opposite end of the fragments. A paired-end module directs the regeneration and amplification operations to prepare the templates for the second round of sequencing. Once the original templates are cleaved and removed, the reverse strands undergo sequencing-by-synthesis. The second round of sequencing occurs at the opposite end of the templates, generating >36 bp reads for a total of >3 Gb of data, which is an obvious advantage when sequencing large genomes. The short read-length (35 bp) is a limitation but, compared with 454 FLX pyrosequencing, homopolymer errors are less of an issue with this technology.

Applied Biosystems SOLiD™ system

Sequencing by ligation generates DNA by measuring the serial ligation of an oligonucleotide. This technology is used in the SOLiD system (Applied Biosystems, Foster City, CA, USA) [11,12]. All fluorescently labelled oligonucleotide probes are present simultaneously and compete for incorporation. After each ligation, the fluorescence signal is measured and then cleaved before another round of ligation takes place. The SOLiD system is a massively parallel genomic analysis platform that supports a wide range of applications. The flexibility of two independent flow cells allows multiple experiments in a single run. The SOLiD system can cost effectively complete large-scale sequencing and, with a reference sequence for a microorganism, it is possible to