Revista Mexicana de Biodiversidad 85: 1249-1264, 2014
DOI: 10.7550/rmb.43498

The study of biodiversity in the era of massive sequencing

El estudio de la biodiversidad en la era de la secuenciación masiva

Ana E. Escalante[1]✉, Lev Jardón Barbolla[2], Santiago Ramírez-Barahona[3] and Luis E. Eguiarte[3]

[1]*Laboratorio Nacional de Ciencias de la Sostenibilidad, Departamento de Ecología de la Biodiversidad, Instituto de Ecología, Universidad Nacional Autónoma de México. Apartado postal 70-275, 04510 México, D. F., Mexico.*
[2]*Centro de Investigaciones Interdisciplinarias en Ciencias y Humanidades, Universidad Nacional Autónoma de México. Torre II de Humanidades 4° piso, Circuito Interior, Ciudad Universitaria. Coyoacán 04510, México, D.F., Mexico.*
[3]*Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México. Apartado postal 70-275, 04510 México, D. F., Mexico.*
✉ *anaelena.escalante@gmail.com*

Abstract. Recent years have witnessed the advent and rapid development of massive sequencing technology, commonly known as Next Generation Sequencing (NGS). This technology allows for rapid, massive and inexpensive sequencing of genome regions or entire genomes, making possible genomic studies of non-model organisms and has seen great progress in metagenomic studies. The promise of this information-rich era is to expand the molecular approach of ecological and evolutionary studies towards urgent issues related with conservation and management of biological diversity in the face of global change. Among the current NGS technologies, there are fundamental differences that impact DNA sequence accuracy, length and range of applications. Key differences among platforms are the procedure for library preparation (when needed) and the sequencing process itself (e.g., pyrosequencing, synthesis). In this review we describe the technical details of commercially available platforms for massive sequencing. We discuss their potential applications for specific biodiversity analyses, from model to non-model organisms, from Single Nucleotide Polymorphism (SNPs) to entire genome analysis and metagenomic approaches of microbial communities, including possible taxonomic, phylogenetic, conservation biology and ecosystem applications of NGS methods in the study of biodiversity. We also provide a to-date estimation of the associated costs for each approach and the computational implications for the analyses of sequences derived from these platforms.

Key words: massively parallel sequencing, next-generation sequencing, genomics, metagenomics.

Resumen. Recientemente se han desarrollado nuevas tecnologías de secuenciación masiva, conocidas como secuenciación de siguiente generación (NGS, por sus siglas en inglés). Estas tecnologías permiten secuenciación rápida, masiva y a bajo costo de regiones genómicas o genomas completos, haciendo posible estudios genómicos de organismos no modelo y estudios metagenómicos. Estas tecnologías prometen expandir las aproximaciones moleculares de estudios ecológicos y evolutivos hacia asuntos relacionados con conservación y manejo de la diversidad biológica ante retos como cambio climático. Entre las plataformas NGS disponibles hay diferencias fundamentales que resultan en diferente precisión en la determinación de las secuencias, así como diferencias en la longitud de las mismas. Algunas diferencias clave entre plataformas son los procedimientos para la preparación de bibliotecas (cuando son necesarias) y el proceso de secuenciación *per se* (e.g., pirosecuenciación, síntesis). En esta revisión se describen las plataformas comercialmente disponibles para NGS y se discuten sus aplicaciones en estudios de biodiversidad de organismos modelo o no modelo, como son análisis de polimorfismos únicos (SNPs), así como análisis de genomas completos y aproximaciones metagenómicas para el estudio de comunidades microbianas. También revisamos posibles aplicaciones de los métodos NGS para resolver problemas taxonómicos, filogenéticos, de biología de la conservación y de ecosistemas, todos relevantes en el estudio de la biodiversidad. Adicionalmente se presenta una estimación actual de los costos asociados para cada plataforma, así como las implicaciones computacionales para los análisis de secuencias derivadas de estas tecnologías.

Palabras clave: secuenciación masiva en paralelo, secuenciación de siguiente generación, genómica, metagenómica.

Recibido: 27 noviembre 2013; aceptado: 23 mayo 2014

Introduction

In recent years, development of massive sequencing technology has permitted rapid and relatively inexpensive sequencing of large portions or even entire genomes of different organisms. These technologies, in contrast with more traditional sequencing methods, allow sequencing non-model organisms for which limited genetic information is available (Mardis, 2008; Neale and Kremer, 2011; Cahais et al., 2012). Moreover, the field of metagenomics has flourished with the advent of massive sequencing technologies, which broadens the range of questions that can be posed and answered from ecological and evolutionary perspectives in conservation and management of natural resources (Bonilla-Rosso et al., 2008; Eguiarte et al., 2013).

The overarching goal of genomic studies is the understanding of diversity, defined as genetic variation, nucleic acid sequence variation, and the comparison of such variation among organisms (Hedrick, 2000; Eguiarte et al., 2013). This goal can be accomplished in different ways, depending on the scale at which the analysis is conducted (Mardis, 2008; Metzker, 2010; Neale and Kremer, 2011; Zhang et al., 2011). For example, the analysis can be limited to determining the sequence of a region of the genome or the complete genome of the organisms under study, including organelles in the case of eukaryotes. Furthermore, the analysis of sequences can be taken to another level by looking at the physical position of each base in a genetic map, and even further with molecular evolution and population genetics analyses that have seen considerable advances since genomic data became available (Turner and Hahn, 2007; Michel et al., 2010; Yi et al., 2010). As examples of the opportunities that genomics offer to population genetics studies, it has been shown that when comparing different lineages, it is possible to determine the physical arrangement of genes and their evolutionary conservation (e.g., synteny; Mathee et al., 2008), the evolutionary dynamics of species and genomes (e.g., *Salmonella*; Holt et al., 2008), gene and genome duplications and architecture (Ibarra-Laclette et al., 2013), the evolutionary history of species (phylogenomics; (Delsuc et al., 2005) and the genetic targets of selection and the genetic basis of adaptation (Yi et al., 2010).

Moreover, genomic studies offer great promise in advancing our understanding of complex processes associated with gene expression and gene interactions (i.e., transcriptomics, proteomics, epigenomics, developmental genetics, epistasis, pleiotropy, etc.), as well as to take a better grip into the genetic basis of phenotype and the complex relationship genotype-phenotype-environment (Mardis, 2007; Mardis, 2008; Neale and Kremer, 2011). Finally, the new field of metagenomics, in which the goal is the sequencing of all genomes from an environmental sample, has been particularly benefited from massive sequencing (Thomas et al., 2012).

Given the great insight that genomic information offers to understanding biodiversity, we consider very important to revise currently available tools that researchers in natural sciences can use if they decide to pursue a genomic approach in their studies. In this review, we describe technical details of the commercially available platforms for massive sequencing, as well as their associated costs and computational demands. We also discuss the potential applications of these platforms for specific biodiversity analyses, either ecological or purely evolutionary, from model to non-model organisms, from Single Nucleotide Polymorphism (SNPs) across the genome of regions of unknown identity or specific genes, to entire genome analysis and metagenomic approaches of microbial communities.

Sequencing basics: from Sanger to next generation sequencing (NGS)

For almost 3 decades, sequencing efforts were carried out by the classic Sanger method (Sanger et al., 1977), which is still the most used approach for routine molecular analyses. This method allows determination of nucleotide sequence of DNA fragments in the range of 1000 base pairs (bp), which approximates the length of an average gene. Sanger sequencing method signified a great improvement with respect to the Maxam-Gilbert (1977) method, which requires 32P as radioactive marker with the inconvenience of the relatively long decay rate. Moreover, and in contrast to Sanger or even some next generation methods for sequencing, the Maxam-Gilbert method is not based in the copy or synthesis of a template DNA strand, but in the inference of the sequence as a result of chemical alterations or enzymatic restrictions of the sequence of interest (Maxam and Gilbert, 1977). The simplicity of Sanger method or dideoxi, based on chain-termination synthesis made possible its popularization and establishment as the standard sequencing method. Nonetheless, Maxam-Gilbert method is still used in some cases to identify epigenetic modifications, such as methylation (Church and Gilbert, 1984; Isola et al., 1999; Ammerpohl et al., 2009).

Optimization and automation of sequencing by Sanger method was possible thanks to its coupling with PCR (polymerase chain reaction) technique (Mullis and Faloona, 1987) to obtain multiple copies of a specific DNA fragment, as well as with the introduction of fluorescently marked dideoxinucleotides instead of the originally