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

radioactively marked ones (Smith et al., 1986). Today, automated sequencing by Sanger method requires the generation of relatively small fragments (1 000 bp), either by selective amplification by PCR of the regions of interest (amplicons), or by physical or chemical fragmentation of entire genomes, a strategy known as shotgun sequencing (Staden, 1979). Once the fragments are generated, these can be sequenced. If the fragments are obtained through amplification of a single homozygous individual, sequencing can proceed through synthesis. In the case of heterozygous or a mixture of individuals (populations) or species (communities, metagenomics), and also in the case of shotgun generated fragments, individualization of the different copies is needed and is usually accomplished by clone library construction (Escalante, 2008).

The shotgun strategy was of great relevance and increased the rate at which genomes were sequenced, becoming the most used approach for genome sequencing for many years (Anderson, 1981; Roach et al., 1995; Adams et al., 2000). In fact, this technology allowed sequencing of the first complete genomes, including the human genome (International Human Genome Sequencing Consortium 2001). However, this massive approach has two main difficulties: *i)* each sequencing reaction has to be performed separately, which results in a limited number of base pairs obtained per day, and *ii)* the cost per base pair is relatively high, which at a genomic scale can be prohibitive for most laboratories and research institutions.

As we will describe below, massive sequencing methods couple clone library construction with sequencing in different ways. In other words, one common trait of all platforms of massive sequencing is a 2 level parallelization (Liu et al., 2012): generation and separation of millions of individual fragments, and their subsequent parallel sequencing (Mardis, 2008). It should be noted that some next generation sequencing strategies promise real time sequencing of single molecules, eliminating the clone library construction step (Thompson and Milos, 2011; Carneiro et al., 2012).

Another common element that has been important in the development of massive sequencing technologies is the progressive incorporation of advances in materials science. This allows the use of thin layers, nanopores, nanoscale emulsions or microspheres as physical means to adhere DNA fragments and probes. This miniaturization of the scale of work is a key aspect of the massive parallelization of the processes involved in simultaneously obtaining of the nucleotide sequences of thousands and millions of fragments.

Besides the specific details of clone library construction and sequencing strategies, to date all sequencing platforms have the same basic stages of sample processing. These stages are: *i)* total DNA or RNA extraction; *ii)* amplification of specific genomic regions or fragmentation (shotgun) of total DNA. Fragmentation can be done enzymatically or physically (e.g., nebulization, sonication). In case of RNA sequencing, retrotranscription is required to obtain cDNA which in turn is fragmented; *iii)* selection of fragments by size (usually 200 bp); *iv)* adaptor sequence addition to the fragment mix; *v)* library construction (except real time sequencing technologies, and *vi)* sequencing.

It is important to note that in all cases during the preparation of fragments it is possible to incorporate specific sample-labels to simultaneously sequence multiple samples, a procedure that is technically known as multiplexing or barcoding. The uniqueness of the next generation sequencing methods resides in the fact that addition of a new nucleotide generates some type of physicochemical change (i.e., light emission, voltage change, fluorescence) that can be detected by an ultrasensitive device coupled to the sequencing system itself.

In the following section we will describe the strategies of library construction and sequencing in detail for the currently available sequencing platforms.

Next generation sequencing methods or massive parallel sequencing

Coverage is a central concept to the description of capacities, limitations and error rates of massive sequencing methods. It refers to the number of times that a specific nucleotide is read during the sequencing process (DePristo et al., 2011). Coverage of 30X means that, on average each base has been read 30 times. However, coverage if not uniform across regions, with regions having low or null coverage and regions being well covered. This variation in coverage across regions follows a certain distribution, sometimes called coverage depth or depth histogram (http://www.illumina.com/truseq/quality_101/coverage/coverage_distribution.ilmn).

Higher coverage guaranties that almost all regions have been sequenced at least once. However, it increases laboratory and bioinformatics costs while increasing certainty of base assignment. Given that DNA fragmentation generates a collection of pieces of different size and with some level of overlap among them, a single nucleotide of the same sequence will be represented many times during the amplification and sequencing processes. In general terms, platforms with longer reads (more bp) tend to generate less copies of the same nucleotide, thus have less coverage. On the other end, platforms with smaller fragments tend to deliver more copies of the same nucleotide, improving coverage but requiring more computational resources for adequate sorting and assembly.

The level of coverage gains relevance because of the elevated error rate of the available methods of massive sequencing in comparison to Sanger sequencing (Table 1). More reads give confidence to the base assignment during the reading of sequences and will also simplify the assembly and sorting during the bioinformatics processing. Moreover, biases during library construction may result in an overrepresentation (more coverage) of particular fragments or amplicons, which will alter considerably the results and their interpretation. These biases can have technical (procedural) causes, may be produced by intrinsic sample characteristics (CG content, DNA extraction biases) or can be associated with platform-inherent and poorly understood causes.

Among the currently available platforms for massive sequencing, the most used are the second generation platforms, which include 454, Illumina, SOLID and Ion-Torrent. We will describe each one in terms of their procedures for library construction and sequencing strategies.

Roche / 454 GLS FLX Titanium Pyrosequencer. This platform is commonly referred to as pyrosequencer. The fragments to be sequenced, as we previously mentioned, have to be selected by size (400-600 bp), and then processed to ligate the adaptors necessary for library construction. The procedure for library construction relies on PCR amplification of each DNA fragment in the sample. Each amplification reaction happens inside a micelle that contains a single DNA molecule, a nanosphere or bead, a polymerase molecule and the rest of the components necessary for the reaction to proceed (dNTPs, buffer, $MgCl_2$, primers). This process is known as emulsion PCR. Each bead has oligonucleotides or probes attached to its surface, the sequence of these probes is complementary to the fragment adaptors. It is important to quantify with precision the components of the PCR reaction, because each micelle has to contain only 1 DNA molecule in order to create a clonal amplification attached to each bead. After library construction, the resulting emulsion, which contains the beads (each bead is a clone), is deposited in a picotiter plate (PTP). The PTP allows only 1 bead in each of the hundred thousand wells that are individually monitored during the pyrosequencing process. Data for an individual DNA sequence will be produced in each well.

Pyrosequencing reactions consist on the synthesis of the complementary strand of the DNA molecules attached to the beads. This reaction is coupled with the activity of enzyme containing beads that are also added to the PTP and surround the DNA beads. Enzymatic activity of luciferase and sulfurilase catalyze the downstream pyrosequencing reaction steps. The PTP serves as a flow cell into which single nucleotide solution is added 1 at a time. The addition of nucleotides results in light emission and no enzymatic termination. The PTP is coupled with a high-resolution camera (CCD) that takes pictures every nucleotide round. Nucleotide addition is not followed by termination and thus the first nucleotides of the adaptor serve as a light emission calibration equivalent to 1 nucleotide, allowing the calculation of the number of nucleotides added in a single round of synthesis (Fig. 1). However, the calibration for long stretches of a single nucleotide has its limits, and it is usually inaccurate when > 6 bases of the same nucleotide are added in a row (Mardis, 2008). Although in the last years the length of reads obtained by

Table 1. Next generation sequencing platform traits (modified from Glenn, 2011 and Quail et al., 2012)

Platform	*Library construction/ sequencing*	*Millions of reads per run*	*Bases per read*	*Yield Gb/run*	*Error rate*	*Error type*	*Service cost (USD) per Gb*
Sanger	PCR/ synthesis	0.000096	650 bp	0.00006	0. 1%	Substitution	6000
454 FLX Titanium	Emulsion PCR/ pyrosequencing	1 million	800 bp	0.5	1%	Indels	12 000
Illumina GAIIx	Bridge/ synthesis	Approximately 200 million	2 X 150 bp	30	0.76%	Substitution	148
Illumina HiSeq 2000	Bridge/ synthesis	Approximately 4 000 million	2 X 150 bp	600	0.26%	Substitution	41
Illumina MiSeq	Bridge/ synthesis	Approximately 10 million	2 X 150 bp	2	0.80%	Substitution	502
SOLID (5500xl)	Emulsion PCR/ synthesis	Up to 1 400 million	Up to 100	155	0.01%	AT bias	40
Ion-Torrent (PGM, 318 Chip)	Emulsion PCR/ ligation	Up to 5 million	200 bp	Up to 1	1.7%	Indels	1 000 (318 Chip)
Pac-Bio	None/ SMRT	0.1 million	800 bp	0.1	12.86%	CG deletions	2 000

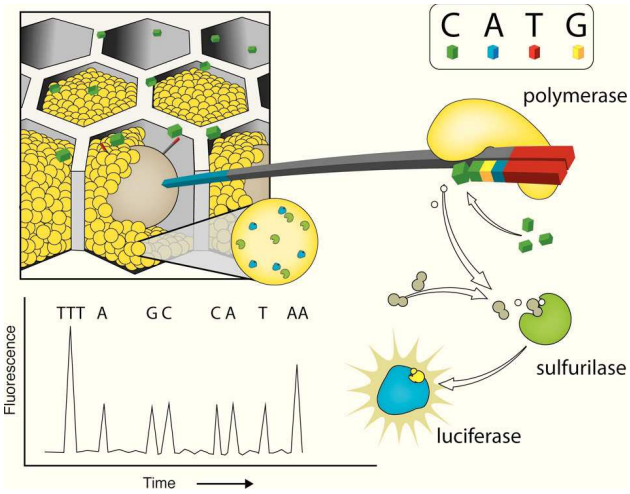

Figure 1. Pyrosequencing using Roche/454 Titanium platform. After loading the DNA-amplified beads (libraries) into individual Pico Titer Plate (PTP) wells, other type of beads, coupled with luciferase and sulphurilase, are added. The figure shows just one type of 2´-deoxyribonucleoside triphosphate (dNTP) – cytosine - that flows through the wells. Once the polymerase adds one nucleotide, a sulfurilase-luciferase reaction occurs, emitting light. A fiber optic slide is attached to a microfluidics camera allowing the reagents to reach the wells packed with beads. Underneath the fiber optic slide there is a direct connection to a high-resolution camera (charge coupled device or CCD), which allows detection of light emitted by each PTP when a pyrosequencing reaction occurs. Modified from Metzker (2010).

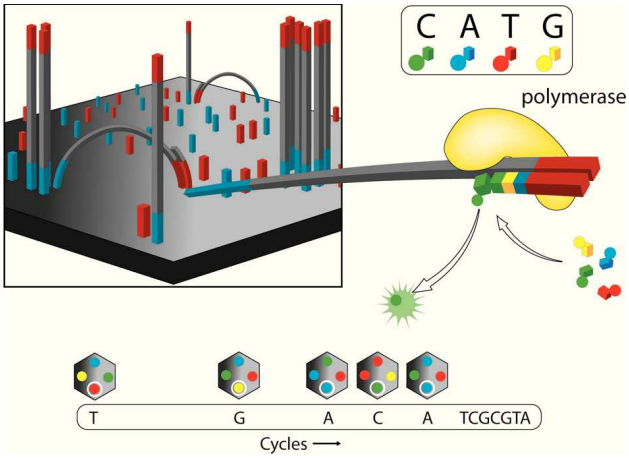

Figure 2. Illumina. *Solid phase amplification* (upper left). The initial step is the coupling of oligonucleotides (priming) and the extension of single-strand molecules. The next step is bridge amplification of the immobilized template with the adjacent primer to form clusters. *Four fluorescent color reversible termination* (upper right). This platform uses a terminator and fluorescent label that differs for each base type. After the addition of a new base, a picture is taken and fluorescence removed leaving the hydroxyl group free for a new addition (bottom). As an example, the bottom part of the figure shows 4 color images that represent the sequence data from one template. Modified from Metzker (2010).

pyrosequencing has increased considerably (Shokralla et al., 2012), the coverage for this approach is still the lowest for the currently available platforms, ranging from 7-10X, depending on genome size (i.e. Barbazuk et al., 2007; Wheeler et al., 2008; Vera et al., 2008).

Illumina genome analyzer (GAIIx). This is one of the most praised massive sequencing platforms, due to its high quality sequences and great coverage (millions of simultaneously sequenced fragments). The procedure for library construction is also based on synthesis of complementary strands of DNA through PCR. However, the specifics of such synthesis have notable differences with the emulsion PCR used by 454. As most current massive sequencing protocols, DNA fragmentation and size selection are necessary steps prior to library construction. Once fragments of a desired size are selected, platform-specific adaptors are added which allows attachment to the sequencing matrix, a flow-cell device that will support bridge amplification and sequencing *per se*. On this matrix, each attached fragment is amplified producing multiple and identical DNA copies in a cluster (Fig. 2). Illumina´s matrix has 8 separate channels (or lines) in each of which a library of clusters can be created. Each cluster is sequenced by synthesis, in which all 4 nucleotides are added simultaneously to the flow device, along with DNA polymerase for addition into the oligo-primed fragments that form the clusters. For this sequencing approach, the addition of a nucleotide blocks subsequent incorporation of nucleotides, interrupting the synthesis, and releases a fluorescent signal that is unique for each type of nucleotide. Imaging of the fluorescent signal follows the incorporation. After imaging, the blocking group is removed and leaves DNA strands ready for the next nucleotide incorporation by DNA polymerase. This series of steps continues for a specific number of rounds allowing read-lengths of 60-150 bases (Glenn, 2011).

Applied biosystems SOLiD^{TM} *sequencer*. Sequencing by oligo ligation detection (SOLiD). This platform has the best quality of sequences and the smallest error rate (Table 1), as a result of the ligation-based approach. The library construction occurs through an emulsion PCR with small magnetic beads (similar to 454/Roche). After completion of library construction, the resulting beads are attached covalently to a flow-cell glass slide. Compared to other platforms, SOLiD has a major difference in that the approach taken for sequencing the amplified fragments,

uses DNA ligase as illustrated in figure 3. The slide with attached beads is exposed to cyclic enzymatic reactions. During the first cycle a primer sequence is incorporated, that is complementary to the adaptor ligated to the fragments as well as a ligase and 4 fluorescently labeled 2-base encoded probes. Non-ligated probes are then washed away, followed by the imaging of the released fluorescence that identifies the ligated probe (Landergren et al., 1988). The cycle is repeated to remove the fluorescent dye and regenerate the 5´-PO4 groups for 10 subsequent ligation cycles (Fig. 3). A second ligation round is performed with a "n-1" primer, which resets the interrogated sequence one base downstream, and then the 10 cycle ligation proceeds.

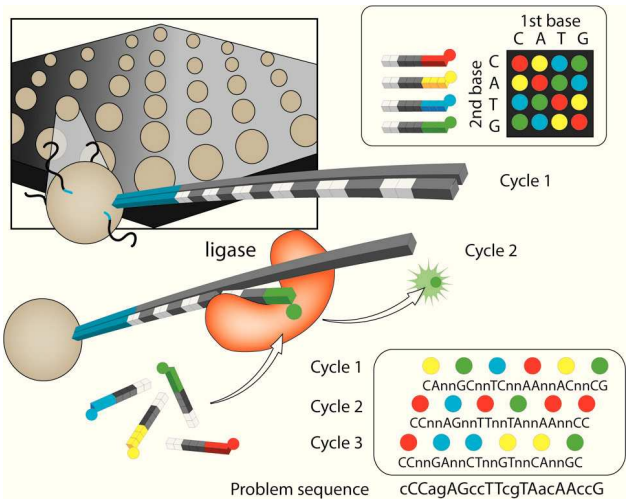

Figure 3. SOLiD. *Four color sequencing by ligation*. After annealing of a universal primer, a library of 1-2-probes is added. In contrast with sequencing by synthesis, sequencing by ligation involves ligation of the probe to a universal primer. The ligation event releases fluorescence and after imaging, ligated probes are chemically removed to generate a 5´-PO4 group. The cycle is repeated 9 more times, after which the primer is removed. Upon primer removal, another primer is added but with a one base pair shift towards the 3-´end. Four base pair shifts are completed with the corresponding 10 ligation cycles. The 1-2-probes are designed to interrogate the first (x) and second (y) positions adjacent to the hybridized primer, such that the 16 nucleotides are encoded by 4 dyes. *Two-base color scheme*. There are 4 di-nucleotide sequences associated with a color (e.g. AA, CC, GG, TT are coded with a blue dye). Each template or sequence is interrogated twice and compiled in a color space. The readings of the color space are aligned against a reference to decode the DNA sequence. An example of the double interrogation is illustrated for three out of the 5 cycles that complete the reading (lower left). Upper case letters denote double interrogated bases, lower case letters are bases interrogated once. If 5 cycles were included, all bases would have been interrogated twice. Modified from Metzker (2010).

Four more rounds of ligation cycles are performed with progressively "n-1" primers. Color calls from the 5-ligation rounds are then ordered into a linear sequence (the color space) and compared to a linear sequence to decode DNA problem sequence. The read length of SOLiD was initially 35 bp with a final output of 3 Gb per run (Liu et al., 2012). These numbers have increased up to 75 bp per read and 10 Gb per run (Shokralla et al., 2012), with high accuracy in the assignment of bases due to the 2-base sequencing method (accuracy of 99.85% after filtering; Liu et al., 2012; Table 1).

Ion-torrent (*the chip is the machine*). This platform is gaining popularity in the market mainly due to the low cost of sequencing runs and the possibility of in-house daily use without many technical requirements or maintenance. The library construction procedure is almost identical to 454 or pyrosequencing, including DNA fragmentation and adaptor ligation. The bead library from the emulsion PCR is deposited into a "chip" with millions of wells (165-600 million depending on the specific version; Shokralla et al., 2012), where only 1 bead fits in each well. The extraordinary aspect of this technology is the chip itself. Each well is integrated to the chip's ion-sensitive layer and a proprietary ion sensor to register the very small voltage changes (per well) that result from nucleotide addition during DNA sequencing by synthesis (Rothberg et al., 2011, Fig. 4). As in 454, nucleotide addition is

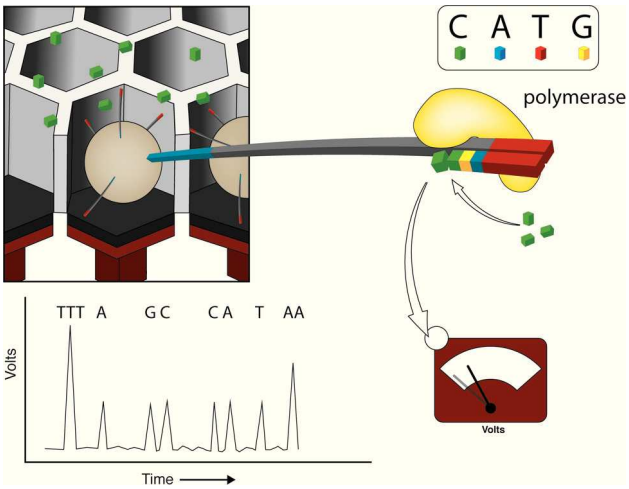

Figure 4. Ion-Torrent. The bead libraries derived from emulsion PCR are deposited into a "chip" with millions of cavities or wells that fit only one well. Each well is part of a microtransistor that integrates the chip, where voltage changes can be registered individually. Sequencing proceeds in a similar fashion as pyrosequencing, but instead of light emission, the addition of each base produces a voltage change. Modified form Niedringhaus et al. (2011).

not followed by termination and it proceeds in cycles, 1 nucleotide after another. This results in the same problem that the pyrosequencing platforms have with homopolymer detection. The read length of ion Torrent is currently about 150 bp with a final output of 3 Gb per run (Liu et al., 2012), but it has increased up to 75 additional bp per read, and up to 10 Gb per run (Shokralla et al., 2012; Table 1).

Single molecule real time (SMRT) sequencing. Technological bets on massive sequencing are focused on the possibility to sequence single molecules in real time or "single molecule real time" (SMRT) sequencing. These technological developments are referred to as the third generation sequencing and, to date, only Pacific Biosciences (PacBio) has released a commercial platform implementing SMRT sequencing. SMRT implements the possibility of attaching a polymerase to a sequencing matrix and be able to follow in real time the synthesis process of a single DNA molecule (Fig. 5). An important feature of SMRT is that it does not require library construction as a prior step to sequencing, increasing the sequence production rate. Also, this technology allows for longer reads. The read length of Pacific Biosciences system has been reported to be up to 1 500 bp with a final output of 60-75 Mb per run (Shokralla et al., 2012 ; Table 1). However, it is difficult to score single base additions in real time. The main problem in scoring real time single base additions is the high speed at which each polymerase synthesizes

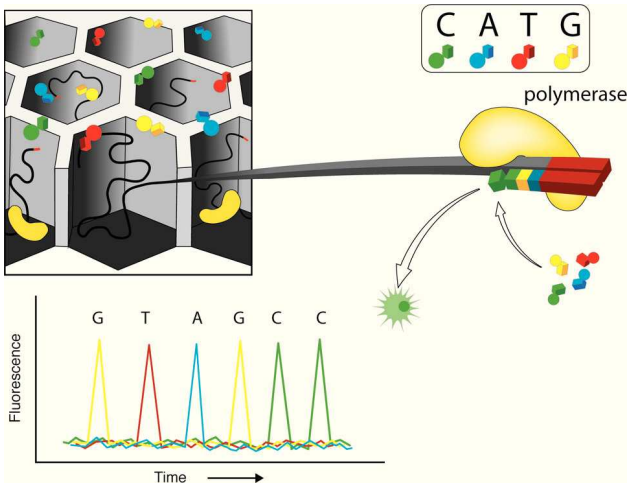

Figure 5. Single molecule real time sequencing (SMRT). PacBio. Single molecules of polymerase enzymes are attached to a sequencing matrix where just one DNA molecule will be synthesized. The enzymatic synthesis is followed individually, in real time, to reconstruct the sequence. SMRT technology does not require library construction as a prior step to sequencing, increasing the sequence production rate. Modified from Metzker (2010).

DNA exhibits stochastic fluctuations (Eid et al., 2009), thus, each enzyme has to be monitored individually and nucleotide additions registered at the appropriate speed in real time. This difficulties result in the highest error rates among NGS platforms (Table 1).

Sequence assembly and NGS

Currently, the main bottlenecks in genomic studies are in the post-sequence stages or assembly, which usually requires major computational capacities (Henson et al., 2012). Genomes sequences are variable (for instance, including heterozygosity in diploid organisms) and can be highly repetitive, and during assembly it is necessary to distinguish this "real" variation from sequencing errors and stay within reasonable computational times, making assembly a complex problem (Henson et al., 2012). The task would be much simpler if it could be determined whether a given set of reads corresponds to overlapping positions on the genome. In this sense, generally longer reads help to correctly find overlapping regions. Given the presence of short reads in all NGS platforms, it is clear that assembling genomes, metagenomes or transcriptomes is a task that faces enormous challenges, which are even more challenging due to high error rates.

The ease and accuracy of assembly depends on the degree of overlapping between reads. Two reads are considered as overlapping when there is a sequence match between reads that is long enough to be reliably distinguished from a random event (Henson et al., 2012). Thus, high uncertainty in assembly arises from locations in which not enough overlap is present to extend the genome sequence. In combination, coverage and read length increase confidence levels of the assembly process. Experience and models show that a good assembly using Sanger reads requires each base to be covered on average by at least 3 reads (3X, Lander and Waterman, 1988). However, for the short reads of NGS platforms, this number can rise up to 30X (Farrer et al., 2009; Meyer et al., 2012). High error rates also affect assembly and thus higher coverage is needed.

Published genomes have used between 5X and 10X with Sanger (Adams et al., 2000; Venter et al., 2001; Venter et al., 2004), but new publications have begun to report 50X, 100X and higher coverage with NGS (Diguistini et al., 2009; Quail et al., 2012). However, even with high coverage, overcoming the problem of repeats and derived assembly gaps sometimes need to be spanned by paired reads (2 reads generated from a single fragment of DNA and separated by known distance), which are available for most NGS platforms (Schatz et al., 2010). Much research has been published on the assembly problem, as well as development of new assembly algorithms and platforms

(for recent reviews see Schatz et al., 2010; Nagarajan and Pop, 2013). Nevertheless, it has been suggested that the best approach is to use a reference genome sequence as a guide to resolve repeats, an approach known as "comparative assembly" (Pop et al., 2004).

A detailed analysis by Schatz et al. (2010) on the assembly of large genomes using NGS (Illumina and 454) showed that assemblies with these platforms are inferior than those accomplished using Sanger technology, but recognized the appeal of the lower costs of NGS. In this paper, guidelines are given to decide the best way to proceed in terms of choosing the platform and assembly method when pursuing a genome sequencing project. As we mention above, the keys to good assembly results from deep coverage by reads with lengths longer than common repeats, paired-end reads from short (0.5-3 kb) and long (> 3 kb) DNA fragments. Using NGS platforms, the most cost-effective way to obtain sequence coverage is to use pair-end sequencing by Illumina with at least 20X coverage. Schatz et al. (2010) also mention that with the assembly software available today, it is technically feasible and cost-effective to build a good assembly entirely from short reads (Illumina 300 bp). A reliable genome project should be planned with the aim of producing deep coverage (30X) in paired-end sequences from short DNA fragments (0.5-1 kb) and additional coverage (10-20X) in paired ends from longer DNA fragments (3-10 kb). A similar strategy was followed to produce the panda genome (Ruiqiang Li et al., 2010). Another robust option would be to combine platforms, as implemented in the turkey genome assembly (Dalloul et al., 2010), in which lower coverage (5X with 454) and higher coverage (25X with Illumina) approaches are simultaneously used to overcome the coverage-ease of assembly paradox.

Considerations on information management

The size of data outputs from NGS platforms represent an important challenge in terms of information processing and analysis (Schuster, 2008; Liu et al., 2012; Yoccoz, 2012), which has triggered advancement in information and data management technologies. In a recent opinion published by Yoccoz (2012), a critical view is presented on the analytical challenges posed by massive data in the study of microorganism biodiversity (inventories) from environmental samples. This view is also applicable to other types of studies, such as whole genome and transcriptome sequencing. The author identifies 2 major obstacles when dealing with such data: *i)* data storage and management, and *ii)* implementation and assessment of statistical models.

Regarding data storage and management, the challenges are very clear: the amount of data that are produced by each platform per run ranges from 1 to 600 GB (Table 1). Thus, the processing, management and storage of data from multiple projects becomes a major task and needs the development of a sophisticated system to handle information, from sample labeling to library construction (multiplexing labeling), as well as a system of sequencing and informatic analysis (Liu et al., 2012). Most laboratories will outsource library construction and sequencing. However, sample and data management, as well as bioinformatic processing and analyses will usually be in-house tasks. The size of data sets that are generated in NGS is also an issue for sorting specific information and for the implementation and assessment of statistical models. Thus, following Cardenas and Tiedje (2008) we emphasize that "the limitation is not the ability to produce sequence data but the ability to store and analyze it in new revealing ways".

Platforms: which one is best?

All sequencing platforms have been advancing towards massive DNA sequencing of longer DNA fragments, and to the production of even larger data sets. To take the full advantage of the generated data, equally massive computational tools are required. However, these computational requirements vary from platform to platform, thus it is important to evaluate the available bioinformatic resources (e.g., hardware, software, human resources) before deciding to choose one technology over another. The bioinformatics challenge is of such relevance for the development of current and future sequencing technologies, that it has triggered a revolution in the information technologies and massive data storage (Mardis, 2008; Henson et al., 2012).

Error rates. All massive sequencing platforms have relatively high error rates (compared with traditional Sanger technology). In addition, each platform has specific types of error associated with the sequencing protocol. The type and rate of error have consequences in the informatic processing of the data, as they need to be identified and "cleaned" before data analyses *per se* (Mardis, 2008). Even after being "cleaned", these errors can have consequences during the assembly of fragments. As a result of the different types of error between platforms, the common suggestion is to simultaneously use 2 different platforms for the assembly of complete genomes (Aury et al., 2008; Dalloul et al., 2010). Also, in cases where the error rates are high or high diversity of sequences is expected, it is recommended to compensate with a high coverage per sequencing sample.

When to pick which? As we have reviewed in this paper, coverage, length of reads, error rates, and costs will vary depending on the sequencing platform, making some of

these more suitable than others for specific studies (Table 1). In the absence of reference genomes or in non-model organisms, genome sequencing (*de novo* sequencing) and assembly, as well as characterization of transcriptomes, can be a daunting task from a bioinformatic point of view. Thus, the assembly process benefits greatly from longer reads and higher coverage, both offered by platforms such as pyrosequencing and synthesis, respectively (Glenn, 2011; Martin and Wang, 2011; Cahais et al., 2012). When reference genomes are available, a good alternative when performing experiments and evaluating transcriptome results would be those technologies that, even with short reads, have a sufficiently high coverage to use a reference genome as a scaffold.

In terms of number of publications, sequencing by synthesis using Illumina is dominant (http://www.illumina.com/science/publications-list.ilmn). This platform has been used in studies focused on characterizing transcriptomes, re-sequencing entire genomes and in studies that look for single nucleotides polymorphisms (SNPs), because it is possible to verify that the mutations found are not a product of sequencing errors (Li et al., 2013). The steady increment in read length for Illumina platforms (Shokralla et al., 2012) has made it appropriate for *de novo* assembly of genomes (Dalloul et al., 2010; Li et al., 2010) or transcriptomes (Martin and Wang, 2011), particularly if used in parallel with other platforms that provide longer reads. Moreover, given the development of new software tools and pipelines (e.g. SHARE, Rodrigue et al., 2010), Illumina technologies are also being used in metagenomic analyses (Coetzee et al., 2010; Rodrigue et al., 2010).

Choosing sequencing approaches for biodiversity inventories or monitoring should be done considering the number of individuals and the portion of the genome to be sequenced. If macroscopic organisms are to be identified via barcoding sequences, it is usually more practical to follow a Sanger approach because specific sequences for only a few individuals are needed. In contrast, inventories for microorganisms have been performed following a metagenomic approach, where entire communities are sequenced for a conserved region that can be used to identify taxonomic groups. To date, pyrosequencing have been traditionally used in microbial ecology studies (Sogin et al., 2006). Given the vast diversity usually found in microbial communities, approaches that give more coverage are being used more recently (Degnan and Ochman, 2012). Nonetheless, given the number of publications, we observe that in metagenomic studies, pyrosequencing is the preferred choice above other approaches with shorter reads (http://454.com/publications/all-publications.asp). However, Roche recently announced the shutdown of pyrosequencing platforms (http://www.genomeweb.com/sequencing/roche-shutting-down-454-sequencing-business), which would probably trigger the use of Illumina or platforms such as Ion Torrent.

Although platforms such as Ion Torrent and SOLiD are in the market, publications using these approaches are notably scant in comparison to 454 or Illumina. In the case of Ion Torrent, this could partly be the result of their recent availability in the market. Ion Torrent equipment is being sold massively and it is possible that many publications will appear soon, probably in the field of metagenomics. In the case of SOLiD, its limited impact (in terms of the number of publications) may be the result of the complexity of data processing and assembly (Flicek and Birney, 2010). Nonetheless, SOLiD is recognized as a very reliable platform in the characterization of SNPs and transcriptomes, and in the quantification of mRNA (Glenn, 2011). Furthermore, this platform has been recently used for *de novo* sequencing of the pig genome (Rubin et al., 2012).

Besides genome sequencing, other options exists where NGS can be used to study genetic variation of entire populations of non-model organisms, for which funding may be limited or research questions are not directed towards full genome sequencing. There are some strategies that take advantage of NGS for SNP detection without the need for assembly and that can be used to simultaneously screen genetic variation of entire populations. One of these strategies is known as RAD-tag sequencing, and it seems to be a productive and promising avenue in population genetics (Baird et al., 2008). Restriction-site Associated DNA (RAD) markers are short fragments of DNA adjacent to a particular restriction enzyme recognition site. This approach can be linked to the use of microarrays and hybridization to screen thousands of polymorphic markers. The use of NGS platforms, such as Illumina, now allows to screen, sequence and detect particular SNPs within the RAD-tag fragments generated, facilitating the rapid discovery of thousand SNPs and high throughput of many populations (Baird et al., 2008). Thus, RAD-seq combines molecular biology techniques with Illumina sequencing: enzymatic restriction of DNA (as for RFLPs and AFLPs) and the use of molecular identifiers (MID) to associate sequence reads to particular individuals (Davey et al., 2010). RAD-tag sequencing has been found to be particularly useful in studies of wild populations and non-model study species, promising to become a practical approach taken in ecological population genomics (Davey et al., 2010). The main reasons for the popularity of RAD-tag sequencing and similar methods are that: *i)* there is no variant discovery assay step; *ii)* it is a very versatile method, expected to work with any restriction enzyme or any species; *iii)* it can be applied to many research

problems, from identifying SNPs at large-scale population genotyping to create linkage maps of Mendelian or quantitalive trait loci (QTL).

NGS: opportunities and challenges for the study of biodiversity

The advent of high throughput sequencing technologies has made genomics not the end itself, but an ideal tool to advance in the study of biodiversity and related processes with relative ease and much greater detail and depth than ever before. The questions that can be addressed using NGS technologies range from classical population genetics to community ecology, passing through phylogeography (McCormack et al., 2013), historical demography (Pool et al., 2010; Li and Durbin, 2011), molecular systematics and phylogenetics (Parks et al., 2009). In particular, the NGS methods allow a more detailed and comprehensive description of biodiversity at any level of biological organization than any previous technology (Taberlet et al., 2012), information that can be used to design conservation and management strategies with a more systemic vision.

Studies on population genetics and molecular evolution have historically been limited by the amount of genetic variation that can be accessed with traditional gene-by-gene approaches (Eguiarte et al., 2013). In addition, the fragmented genomic information coming out from these approaches limits our capacity to determine key parameters, such as recombination, substitution rates and demographic changes (Hedrick, 2000). The acquisition of tens to hundreds of markers and genome data for non-model organisms through classical or Sanger sequencing becomes both costly and slow, while high throughput sequencing makes possible (budget and technically wise) to scan entire genomes of non-model organisms to perform comparative and population genomics studies (Hohenlohe et al., 2010). Such studies have been mainly applied to domesticated species of plants and animals (Groenen et al., 2012; Qin et al., 2014). Some studies, however, have dealt with wild organisms, resulting in the identification of the genetic basis of adaptation (for example see studies on the 3-spine stickle back fish, (Hohenlohe et al., 2010); a tropical lizard, (Freedman et al., 2010); and the wild maize (teosinte), (Eguiarte et al., 2013; Pyhäjärvi et al., 2013)).

A common challenge for these type of studies involves the mere size of the genomes of many organisms, specially those with relatively big, polyploid genomes (Niklas, 1997; Amborella Genome Project, 2013). Additionally, the genomes of many organisms are very dynamic, due to the high levels of recombination, the presence of repetitive non-coding regions and mobile elements and there can be wide variance in the genome sizes of a given species (Chia et al., 2012; Díez et al., 2013). In many cases, these difficulties can be tackled with a low coverage transcriptome or by reducing the number of studied genomic regions (*e.g.*, RAD-tag, see above). These 2 approaches are good starting options to search for genome-wide genetic variation in non-model organisms, and to estimate population genetics parameters, such as genetic structure, effective population sizes, gene flow, inbreeding depression and natural selection (Allendorf et al., 2010; Hohenlohe et al., 2010; Andrew et al., 2012; Ashrafi et al., 2012; Hill et al., 2013). Furthermore, transcriptome data can be used in functional studies to compare patterns of gene expression related to environmental conditions, sexes, life history stages and different structures of an organism (Ekblom and Galindo, 2011). The same genomic and transcriptomic approaches can be employed to identify regions suitable to study phylogenetic relationships, particularly between closely related species or intra-specific varieties, for instance the study of Eaton and Ree (Eaton and Ree, 2013) using RAD-seq (a variant of RAD-tag method described above) data in a small group of *Pedicularis* (Orobanchaceae) plant species from Tibet. Whole plastome sequencing in pines has been used to evaluate divergence among closely related species (Parks et al., 2009) and between small endemic populations of North American pines (Whitall et al., 2010).

NGS technologies can also be used to describe microbial species or groups of species present in complex biological samples [e.g., tissue (Kemler et al., 2013), soil (López-Lozano et al., 2013), or water samples (Zaremba-Niedzwiedzka et al., 2013)]. These studies are generally known as metagenomic, and they can look at specific genomic signatures or entire genomes of communities without the need to culture or isolate microbial species in the laboratory. The culture independent metagenomic approach using NGS opens a wide new window for microbiology, allowing us to go way beyond the less than 5% of the estimated microbial diversity that can be successfully characterized using traditional culturing methods (Torsvik et al., 1990). A simple metagenome would involve the NGS of 16S ribosomal RNA, or other highly conserved gene, that can serve to build a coarse inventory of diversity of a given sample (Sogin et al., 2006). Given the sequencing capacity that NGS platforms offer, meta-genomes can be easily extended to look for multiple genes or entire genomes (Bonilla-Rosso et al., 2008; for recent studies in Cuatro Ciénegas in Mexico see Bonilla-Rosso et al., 2012 and Peimbert et al., 2012). Experimental designs looking at different samples and treatments, either across space or time, promise to "unlock the potential of metagenomics" (Knight et al., 2012), which can greatly advance our understanding of the processes occurring at

the microbial level (Kahvejian et al., 2008). These NGS technologies have already revolutionized our understanding of the microbial diversity of the planet (Sogin et al., 2006; Lauber et al., 2009; Martiny et al., 2011) and even that occurring within our bodies (Costello et al., 2012; Human Microbiome Consortium, 2012). Nonetheless, challenges remain, such as taxon-biased DNA extraction, intrinsic sequencing errors, PCR primers biases (in the case of 16S rDNA approaches), overestimation of taxon abundance or distinguishing meta-genomic signatures of uncultured taxa from artifacts (chimeras).

Conclusions

The main revolution and challenge with the availability of massively parallel sequencing technologies is the production of enormous amounts of data in relatively short times. The methodological differences for each platform bring along different advantages and limitations that impact their use in different studies. Thus, a key step is to identify the features that each platform has in order to make a better choice when designing research projects. Finally, given the amount of information generated by these technologies we must stress the importance of the need for computational resources and efficient pipelines for data processing, because these can be the major limitation in the success of a research project based on massive sequencing technologies.

Acknowledgements

This review is a result from the workshop "Avances y perspectivas del uso de métodos de secuenciación de próxima generación en el studio de la genética de poblaciones y su empleo en problemas ambientales", held in México City at Instituto de Ecología in January 2013, and funded by Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. AEE wants to acknowledge financial support from AMC-Loreal-UNESCO scholarship (2012), UC-Mexus collaborative grant (2013) and PAPIIT-UNAM IA200814 (2014). LEE also wants to thank the support of the grants Conacyt CB2011/167826, Sep-Conacyt-ANUIES-ECOS Francia M12-A03, PAPIIT UNAM IN202712 and Conabio KE 004 for NGS related recent research.

Literature cited

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y. H. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor, J. F. Abril, A. Agbayani, H. J. An, C. Andrews-Pfannkoch, D. Baldwin, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M. H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G. H. Karpen, Z. Ke, J. A. Kennison, K. A. Ketchum, B. E. Kimmel, C. D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. C. McIntosh, M. P. McLeod, D. McPherson, G. Merkulov, N. V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S. M. Mount, M. Moy, B. Murphy, L. Murphy, D. M. Muzny, D. L. Nelson, D. R. Nelson, K. A. Nelson, K. Nixon, D. R. Nusskern, J. M. Pacleb, M. Palazzolo, G. S. Pittman, S. Pan, J. Pollard, V. Puri, M. G. Reese, K. Reinert, K. Remington, R. D. Saunders, F. Scheeler, H. Shen, I. Sidén-Kiamos, M. Simpson, M. P. Skupski, T. Smith, E. Spier, A. C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A. H. Wang, X. Wang, Z. Y. Wang, D. A. Wassarman, G. M. Weinstock, J. Weissenbach, S. M. Williams, T. Woodage, K. C. Worley, D. Wu, S. Yang, Q. A. Yao, J. Ye, R. F. Yeh, J. S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. H. Zheng, F. N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. O. Smith, R. A. Gibbs, E. W. Myers, G. M. Rubin and J. C. Venter. 2000. The Genome Sequence of *Drosophila melanogaster*. Science 287:2185-2195.

Allendorf, F. W., P. A. Hohenlohe and G. Luikart. 2010. Genomics and the future of conservation genetics. Nature Reviews Genetics 11:697-709.

Amborella Genome Project. 2013. The *Amborella* genome and the evolution of flowering plants. Science 342:12410891-124108910.

Ammerpohl, O., J. I. Martín-Subero, J. Richter, I. Vater and R. Siebert. 2009. Hunting for the 5th base: techniques for analyzing DNA methylation. Biochimica et Biophysica Acta 1790:847-862.

Anderson, S. 1981. Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Research 9:3015-3027.

Andrew, R. L., K. L. Ostevik, D. P. Ebert and L. H. Rieseberg. 2012. Adaptation with gene flow across the landscape in a dune sunflower. Molecular Ecology 21:2078-2091.

Ashrafi, H., T. Hil, K. Stoffel, A. Kozik, J. Yao, S. R. Chin-Wo and A. Van Deynze. 2012. De novo assembly of the

pepper transcriptome (Capsicum annuum): a benchmark for in silico discovery of SNPs, SSRs and candidate genes. BMC Genomics 13:1471-2164.

Aury, J. M., C. Cruaud, V. Barbe, O. Rogier, S. Mangenot, G. Samson, J. Poulain, V. Anthouard, C. Scarpelli, F. Artiguenave and P. Wincker. 2008. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC Genomics 9:603.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko and E. A. Johnson. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS One 3:e3376.

Barbazuk, W. B., S. J. Emrich, H. D. Chen, L. Li and P. S. Schnable. 2007. SNP discovery via 454 transcritome sequencing. The Plant Journal 51:910-918.

Bonilla-Rosso, G., M. Peimbert, L. D. Alcaraz, I. Hernández, L. E. Eguiarte, G. Olmedo-Álvarez and V. Souza. 2012. Comparative metagenomics of two microbial mats at Cuatro Ciénegas Basin II: community structure and composition in oligotrophic environments. Astrobiology 12:659-673.

Bonilla-Rosso, G., V. Souza and L. E. Eguiarte. 2008. Metagenómica, genómica y ecología molecular: la nueva ecología en el bicentenario de Darwin. TIP Revista Especializado en Ciencias Químico-Biológicas 11:41-51.

Cahais, V., P. Gayral, G. Tsagkogeorga, J. Melo-Ferreira, M. Ballenghien, L. Weinert, Y. Chiari, K. Belkhir, V. Ranwez and N. Galtier. 2012. Reference-Free transcriptome assembly in non-model animals from next-generation sequencing data. Molecular Ecology Resources 12:834-845.

Cárdenas, E. and J. M. Tiedje. 2008. New tools for discovering and characterizing microbial diversity. Current Opinion in Biotechnology 19:544-549.

Carneiro, M. O., C. Russ, M. G. Ross, S. B. Gabriel, C. Nusbaum and M. A. DePristo. 2012. Pacific Biosciences Sequencing Technology for genotyping and variation discovery in human data. BMC Genomics 13:375.

Church, G. M. and W. Gilbert. 1984. Genomic sequencing. Proceedings of the National Academy of Sciences of the United States of America 81:1991-1995.

Coetzee, B., M. J. Freeborough, H. Maree, J. M. Celton, D. J. G. Rees and J. T. Burguer. 2010. Deep sequencing analyses of viruses infecting grapevines; virome of a vineyard. Virology 400:157-163.

Dalloul, R. A., J. A. Long, A. V. Zimin, L. Aslam, K. Beal, L. A. Blomberg, P. Bouffard, D. W. Burt, O. Crasta, R. P. M. A. Crooijmans, K. Cooper, R. A. Coulombe, S. De, M. E. Delany, J. B. Dodgson, J. J. Dong, C. Evans, K. M. Frederickson, P. Flicek, L. Florea, O. Folkerts, M. A. M. Groenen, T. T. Harkins, J. Herrero, S. Hoffmann, H. -J. Megens, A. Jiang, P. de Jong, P. Kaiser, H. Kim, K.-W. Kim, S. Kim, D. Langenberger, M.-K. Lee, T. Lee, S. Mane, G. Marcais, M. Marz, A. P. McElroy, T. Modise, M. Nefedov, C. Notredame, I. R. Paton, W. S. Payne, G. Pertea, D. Prickett, D. Puiu, D. Qioa, E. Raineri, M. Ruffier, S. L. Salzberg, M. C. Schatz, C. Scheuring, C. J. Schmidt, S. Schroeder, S. M. J. Searle, E. J. Smith, J. Smith, T. S. Sonstegard, P. F. Stadler, H. Tafer, Z. J. Tu, C. P.Van Tassell, A. J. Vilella, K. P. Williams, J. A. Yorke, L. Zhang, H.-B. Zhang, X. Zhang, Y. Zhang and K. M. Reed 2010. Multi-Platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. PLoS Biology 8:e1000475.

Davey, J. W., J. L. Davey, M. L. Blaxter and M. W. Blaxter. 2010. RADSeq: next-generation population genetics. Briefings in Functional Genomics 9:416-423.

Degnan, P. H. and H. Ochman. 2012. Illumina-based analysis of microbial community diversity. The ISME Journal 6:183-194.

Delsuc, F., H. Brinkmann and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Reviews. Genetics 6:361-375.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivanchenko, K. Cibulskis, S. B. Gabriel, D. Altshuler and M. J. Daly. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43:491-498.

Consortium, H. M. P. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207-214.

Costello, E. K., K. Stagaman, L. Dethlefsen, B. J. M. Bohannan and D. A. Relman. 2012. The application of ecological theory toward an understanding of the human microbiome. Science 336:1255-1262.

Chia, J. M., C. Song, P. J. Bradbury, D. Costich, N. de Leon, J. Doebley, R. J. Elshire, B. Gaut, L. Geller, J. C. Glaubitz, M. Gore, K. E. Guill, J. Holland, M. B. Hufford, J. Lai, M. Li, X. Liu, Y. Lu, R. McCombie, R. Nelson, J. Poland, B. M. Prasanna, T. Pyhäjärvi, T. Rong, R. S. Sekhon, Q. Sun, M. I. Tenaillon, F. Tian, J. Wang, X. Xu, Z. Zhang, S. M. Kaeppler, J. Ross-Ibarra, M. D. McMullen, E. S. Buckler, G. Zhang, Y. Xu and D. Ware. 2012. Maize HapMap2 identifies extant variation from a genome in flux. Nature Genetics 44:803-807.

Díez, C. M., B. S. Gaut, E. Meca, E. Scheinvar, S. Montes-Hernández, L. E. Eguiarte and M. I. Tenaillon. 2013. Genome size variation in wild and cultivated maize along altitudinal gradients. New Phytologist 199:264-276.

Diguistini, S., N. Y. Liao, D. Platt, G. Robertson, M. Seidel, S. K. Chan, T. R. Docking, I. Birol, R. A. Holt, M. Hirst, E. Mardis, M. A. Marra, R. C. Hamelin, J. Bohlman, C. Breuil and S. J. Jones. 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina Sequence Data. Genome Biology 10:R94.

Eaton, D. A. R. and R. H. Ree. 2013. Inferring phylogeny and introgression using RADseq Data: An example from flowering plants (Pedicularis: Orobanchaceae). Systematic Biology 62:689-706.

Eguiarte, L. E., J. A. Aguirre-Liguori, L. Jardón-Barbolla, E. Aguirre-Planter and V. Souza. 2013. Genómica de poblaciones: nada en Evolución va a tener sentido si no es a la luz de la genómica, y nada en genómica tendrá sentido si no es a la luz de la evolución. TIP Revista Especializada En

Ciencias Químico-Biológicas 16:42-56.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach and S. Turner. 2009. Real-time DNA sequencing from Single Polymerase Molecules. Science 323:133-138.

Ekblom, R. and J. Galindo. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107:1-15.

Escalante, A. E. 2008. Ecología molecular en el estudio de comunidades bacterianas. *In* Ecologia molecular, X. Aguirre, V. Souza and L. E. Eguiarte (eds.). Conabio, INE. Ciudad de México. p. 393-424.

Farrer, R. A., E. Kemen, J. D. G. Jones and D. J. Studholme. 2009. De novo assembly of the *Pseudomonas syringae pv. syringae* B728a genome using Illumina/Solexa short sequence reads. FEMS Microbiology Letters 291:103-111.

Flicek, P. and E. Birney. 2010. Sense from sequence reads: methods for alignment and assembly. Nature Methods 6: S6-S12.

Freedman, A. H., H. A. Thomassen, W. Buermann and T. B. Smith. 2010. Genomic signals of diversification along ecological gradients in a tropical lizard. Molecular Ecology 19:3773-3788.

Glenn, T. C. 2011. Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11:759-769.

Groenen, M. A. M., A. Archibald, H. Uenishi, C. K. Tuggle, Y. Takeuchi, M. F. Rothschild, C. Rogei-Gaillard and E. Al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393-398.

Hedrick, P. 2000. Genetics of Populations. Jones and Bartlett Publishers, Subdury Massachussetts. 553 p.

Henson, J., G. Tischler and Z. Ning. 2012. Next-generation sequencing and large genome assemblies. Pharmacogenomics 13:901-915.

Hill, T. A., H. Ashrafi, S. Reyes-Chin Wo, J. Yao, K. Stoffel, M. J. Truco, A. Kozik, R. W. Michelmore and A. Van Deynze. 2013. Characterization of *Capsicum annuum* genetic diversity and population structure based on parallel polymorphism discovery with a 30K unigene Pepper GeneChip. PLoS One 8:e56200.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson and W. A. Cresko. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet 6:e1000862.

Holt, K. E, J. Parkhill, C. J. Mazzoni, P. Roumagnac, F.-X. Weill, I. Goodhead, R. Rance, S. Baker, D. J. Maskell, J. Wain, C. Dolecek, M. Achtman and G. Dougan. 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella typhi*. Nature Genetics 40:987-993.

Ibarra-Laclette, E., E. Lyons, G. Hernández-Guzmán, C. A. Pérez-Torres, L. Carretero-Paulet, T.H. Chang, T. Lan, A. J. Welch, M. J. Juarez, J. Simpson, A. Fernandez-Cortes, M. Arteaga-Vazquez, E. Gongora-Castillo, G. Acevedo-Hernandez, S. C. Schuster, H. Himmelbauer, A. E. Minoche, S. Xu, M. Lynch, A. Oropeza-Aburto, S. A. Cervantes-Perez, M. de Jesus Ortega-Estrada, J. I. Cervantes-Luevano, T. P. Michael, T. Mockler, D. Bryant, A. Herrera-Estrella, V. A. Albert and L. Herrera-Estrella. 2013. Architecture and evolution of a minute plant genome. Nature 498:94-98.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-892.

Isola, N. R., S. L. Allman, V. V. Golovlov and C. H. Chen. 1999. Chemical cleavage sequencing of DNA using matrix-assisted laser desorptio/ionization time-of-flight mass spectometry. Analytical Chemistry 71:2266-2269.

Kahvejian, A., J. Quackenbush and J. F. Thompson. 2008. What would you do if you could sequence everything? Nature Biotechnology 26:1125-1133.

Kemler, M., J. Garnas, M. J. Wingfield, M. Gryzenhout, K. A. Pillay and B. Slippers. 2013. Ion Torrent PGM as tool for fungal community analysis: a case study of endophytes in *Eucalyptus grandis* reveals high taxonomic diversity. PLoS One 8:e81718.

Knight, R., J. Jansson, D. Field, N. Fierer, N. Desai, J. A. Fuhrman, P. Hugenholtz, D. van der Lelie, F. Meyer, R. Stevens, M. J. Bailey, J. I. Gordon, G. A. Kowalchuk and J. A. Gilbert. 2012. Unlocking the potential of metagenomics through replicated experimental design. Nature Biotechnology 30:513-520.

Lander, E. S. and S. Waterman. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 239:231-239.

Landergren, U., R. Kaiser, J. Sanders and L. Hood. 1988. A ligase-mediated gene detection technique. Science 241:1077-1080.

Lauber, C. L., M. Hamady, R. Knight and N. Fierer. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Applied and Environmental Microbiology 75:5111-5120.

Li, H. and R. Durbin. 2011. Inference of human population history from individual whole-genome sequences. Nature 475:493-495.

Li, R., Y. Li, X. Fang, H. Yang, J. Wang and K. Kristiansen. 2013. SNP detection for massively parallel whole-genome resequencing. Genome Research 19:1124-1132.

Li, R., W. Fan, G. Tian, H. Zhu, L. He, J. Cai, Q. Huang, Q. Cai, B. Li, Y. Bai, Z. Zhang, Y. Zhang, W. Wang, J. Li, F. Wei, H. Li, M. Jian, J. Li, Z. Zhang, R. Nielsen, D. Li, W. Gu, Z. Yang, Z. Xuan, O. a Ryder, F. C.-C. Leung, Y. Zhou, J. Cao, X. Sun, Y. Fu, X. Fang, X. Guo, B. Wang, R. Hou, F. Shen, B. Mu, P. Ni, R. Lin, W. Qian, G. Wang, C. Yu, W. Nie, J. Wang, Z. Wu, H. Liang, J. Min, Q. Wu, S. Cheng, J. Ruan, M. Wang, Z. Shi, M. Wen, B. Liu, X. Ren, H. Zheng,

D. Dong, K. Cook, G. Shan, H. Zhang, C. Kosiol, X. Xie, Z. Lu, H. Zheng, Y. Li, C. C. Steiner, T. T.-Y. Lam, S. Lin, Q. Zhang, G. Li, J. Tian, T. Gong, H. Liu, D. Zhang, L. Fang, C. Ye, J. Zhang, W. Hu, A. Xu, Y. Ren, G. Zhang, M. W. Bruford, Q. Li, L. Ma, Y. Guo, N. An, Y. Hu, Y. Zheng, Y. Shi, Z. Li, Q. Liu, Y. Chen, J. Zhao, N. Qu, S. Zhao, F. Tian, X. Wang, H. Wang, L. Xu, X. Liu, T. Vinar, Y. Wang, T.-W. Lam, S.-M. Yiu, S. Liu, H. Zhang, D. Li, Y. Huang, X. Wang, G. Yang, Z. Jiang, J. Wang, N. Qin, L. Li, J. Li, L. Bolund, K. Kristiansen, G. K.-S. Wong, M. Olson, X. Zhang, S. Li, H. Yang, J. Wang and J. Wang. 2010. The sequence and de novo assembly of the giant panda genome. Nature 463:311-317.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu and M. Law. 2012. Comparison of next-generation sequencing systems. Journal of Biomedicine and Biotechnology 2012:251364.

López-Lozano, N. E., K. B. Heidelberg, W. C. Nelson, F. García-Oliva, L. E. Eguiarte, and V. Souza. 2013. Microbial secondary succession in soil microcosms of a desert oasis in the Cuatro Cienegas Basin, Mexico. Peer J 1:e47.

Mardis, E. R. 2007. ChIP-Seq: welcome to the new frontier. Nature Methods 4:613-614.

Mardis, E. R. 2008. Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9:387-402.

Martin, J. A. and Z. Wang. 2011. Next-generation transcriptome assembly. Nature Reviews Genetics 12:671-682.

Martiny, J. B. H., J. A. Eisen, K. Penn, S. D. Allison and M. C. Horner-Devine. 2011. Drivers of bacterial beta-diversity depend on spatial scale. Procedings of the National Academy of Sciences of the United States of America 108:7850-47854.

Mathee, K., G. Narasimhan, C. Valdes, X. Qiu, J. M. Matewish, M. Koehrsen, A. Rokas, C. N. Yandava, R. Engels, E. Zeng, R. Olavarietta, M. Doud, R. S. Smith, P. Montgomery, J. R. White, P. a Godfrey, C. Kodira, B. Birren, J. E. Galagan and S. Lory. 2008. Dynamics of *Pseudomonas aeruginosa* genome evolution. Proceedings of the National Academy of Sciences of the United States of America 105:3100-3105.

Maxam, A. M. and W. Gilbert. 1977. A new method for sequencing DNA. Proceedings of the National Academy of Sciences of the United States of America 74:560-564.

McCormack, J. E., S. M. Hird, A. J. Zellmer, B. C. Carstens and R. T. Brumfield. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution 66:526-538.

Metzker, M. L. 2010. Sequencing technologies - the next generation. Nature Reviews. Genetics 11:31-46.

Meyer, M., M. Kircher, M. T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso and S. Pääbo. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222-226.

Michel, A. P., S. Sim, T. H. Q. Powell, M. S. Taylor, P. Nosil and J. L. Feder. 2010. Widespread genomic divergence during sympatric speciation. Proceedings of the National Academy of Sciences of the United States of America 107:9724-9729.

Mullis, K. and F. Faloona. 1987. Specific synthesis of DNA in vitro via a polymerase catalyzed chain reaction. Methods in Enzymology 155:335-350.

Neale, D. B. and A. Kremer. 2011. Forest tree genomics: growing resources and applications. Nature Reviews. Genetics:111-122.

Niedringhaus, T., D. Milankova, M. Kerby, M. P. Snyder and A. E. Barron. 2011. Lanscape of next-generation sequencing technologies. Analytical Chemistry 83:4327-4341.

Niklas, K. J. 1997. The Evolutionary Biology of Plants. Chicago University Press, Chicago.

Parks, M., R. Cronn and A. Liston. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biology 7:84.

Peimbert, M., L. D. Alcaraz, I. Hernandez, G. Olmedo, F. García-Oliva, L. Segovia, G. Bonilla, L. E. Eguiarte and V. Souza. 2012. Comparative metagenomics of two microbial mats at Cuatro Cienegas Basin I: Ancient lesson on how to cope in an environment under severe environmental stress. Astrobiology 12:648-658.

Pool, J. E., I. Hellmann, J. D. Jensen and R. Nielsen. 2010. Population genetic inference from genomic sequence variation. Genome Research 20:291-300.

Pop, M., A. Phillippy, A. L. Delcher and S. L. Salzberg. 2004. Comparative genome assembly. Briefings in Bioinformatics 5:237-248.

Pyhäjärvi, T., M. B. Hufford, S. Mezmouk and J. Ross-Ibarra. 2013. Complex patterns of local adaptation in teosinte. Genome Biology and Evolution 5:1594-1609.

Qin, C., C. Yu, Y. Shen, X. Fang, L. Chen, J. Min, J. Cheng, S. Zhao, M. Xu, Y. Luo, Y. Yang, Z. Wu, L. Mao, H. Wu, C. Ling-Hu, H. Zhou, H. Lin, S. González-Morales, D. L. Trejo-Saavedra, H. Tian, X. Tang, M. Zhao, Z. Huang, A. Zhou, X. Yao, J. Cui, W. Li, Z. Chen, Y. Feng, Y. Niu, S. Bi, X. Yang, W. Li, H. Cai, X. Luo, S. Montes-Hernández, M. a Leyva-González, Z. Xiong, X. He, L. Bai, S. Tan, X. Tang, D. Liu, J. Liu, S. Zhang, M. Chen, L. Zhang, L. Zhang, Y. Zhang, W. Liao, Y. Zhang, M. Wang, X. Lv, B. Wen, H. Liu, H. Luan, Y. Zhang, S. Yang, X. Wang, J. Xu, X. Li, S. Li, J. Wang, A. Palloix, P. W. Bosland, Y. Li, A. Krogh, R. F. Rivera-Bustamante, L. Herrera-Estrella, Y. Yin, J. Yu, K. Hu and Z. Zhang. 2014. Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. Proceedings of the National Academy of Sciences of the United States of America 111:5135-5140.

Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow and Y. Gu. 2012. A tale of three next generation sequencing platforms:

comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341.

Roach, J. C., C. Boysen, K. Wang and L. Hood. 1995. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. Genomics 26:345-353.

Rodrigue, S., A. C. Materna, S. C. Timberlake, M. C. Blackburn, R. R. Malmstorm, E. J. Alm and S. W. Chisholm. 2010. Unlocking short read sequencing for metagenomics. PLoS One 5:e11840.

Rothberg, J. M., W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, J. Hoon, J. F. Simons, D. Marran, J. W. Myers, J. F. Davidson, A. Branting, J. R. Nobile, B. P. Puc, D. Light, T. A. Clark, M. Huber, J. T. Branciforte, I. B. Stoner, S. E. Cawley, M. Lyons, Y. Fu, N. Homer, M. Sedova, X. Miao, B. Reed, J. Sabina, E. Feierstein, M. Schorn, M. Alanjary, E. Dimalanta, D. Dressman1, R. Kasinskas, T. Sokolsky, J. A. Fidanza, E. Namsaraev, K. J. McKernan, A. Williams, G. T. Roth and J. Bustillo. 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348-352.

Rubin, C. J., H. J. Megens, A. Martínez-Barrio, K. Maqbool, S. Sayyab, D. Schwochow, C. Wang, Ö. Carlborg, P. Jern, C. B. Jorgensen, A. Archibald, M. Fredholm, M. A. M. Groenen and L. Andersson. 2012. Strong signatures of selection in the domestic pig genome. Proceedings of the National Academy of Sciences of the United States of America 109:19529-19536.

Sanger, F., S. Nicklen and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74:5463-5467.

Schatz, M. C., A. L. Delcher and S. L. Salzberg. 2010. Assembly of large genomes using second-generation sequencing. Genome Research 20:1165-1173.

Schuster, S. C. 2008. Next-generation sequencing transforms today 's biology. Nature Methods 5:16-18.

Shokralla, S., J. L. Spall, J. F. Gibson and M. Hajibabaei. 2012. Next-generation sequencing technologies for environmental DNA research. Molecular Ecology 21:1794-1805.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connel, C. Heiner, S. B. H. Ken and L. E. Hood. 1986. Fluorescence detection in automated DNA sequence analysis. Nature 321:674-679.

Sogin, M. L., H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta and G. J. Herndl. 2006. Microbial diversity in the deep sea and the underexplored 'rare biosphere'. Proceedings of the National Academy of Sciences of the United States of America 103:12115-12120.

Staden, R. 1979. A strategy of DNA sequencing employing computer programs. Nucleic Acids Research 6: 2601-2610.

Taberlet, P., S. M. Prud'Homme, E. Campione, J. Roy, C. Miquel, W. Shehzad, L. Gielly, D. Rioux, P. Choler, J.-C. Clément, C. Melodelima, F. Pompanon and E. Coissac. 2012. Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Molecular Ecology 21:1816-1820.

Thomas, T., J. Gilbert and F. Meyer. 2012. Metagenomics - a guide from sampling to data analysis. Microbial Informatics and Experimentation 2:3.

Thompson, J. F. and P. Milos. 2011. The properties and applications of single-molecule DNA sequencing. Genome Biology 12:217-226.

Torsvik, V., J. Goksøyr and F. L. Daae. 1990. High diversity in DNA of soil bacteria. Applied and Environmental Microbiology 56:782-787.

Turner, T. L. and M. W. Hahn. 2007. Locus- and population-specific selection and differentiation between incipient species of *Anopheles gambiae*. Molecular Biology and Evolution 24:2132-2138.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. a Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V Merkulov, N. Milshina, H. M. Moore, K. Naik, V. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J.

Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu. 2001. The sequence of the human genome. Science 291:1304-1351.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. Rogers and H. O. Smith. 2004. Sequencing of the Sargasso Sea. Science 304:66-74.

Vera, J. C., C. W. Wheat, H. W. Fescemeyer, M. J. Frilander, D. L. Crawford, I. Hanski and J. H. Marden. 2008. Rapid transcriptome characterization for a non-model organism using 454 pyrosequencing. Molecular Ecology 17:1636-1647.

Whitall, J. B., J. Syring, M. Parks, J. Buenrostro, C. Dick, A. Liston and R. Cronn. 2010. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. Molecular Ecology 19:100-114.

Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs and J. M. Rotherberg. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452:472-476.

Yi, X., Y. Liang, E. Huerta-Sánchez, X. Jin, Z. Xi, P. Cuo, J. E. Pool, J. Xu, X. Liu, T. Jiang, R. Wu, G. Zhou, M. Tang, J. Qin, W. Ouyang, X. Ren, H. Liang, H. Zheng, Y. Huang, J. Li, L. Bolund, K. Kristiansen, Y. Li, Y. Zhang, X. Zhang, R. Li, S. Li, H. Yang, R. Nielsen, J. Wang and J. Wang. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329:75-78.

Yoccoz, N. G. 2012. The future of environmental DNA in ecology. Molecular Ecology 21:2031-2038.

Zaremba-Niedzwiedzka, K., J. Viklund, W. Zhao, J. Ast, A. Sczyrba, T. Woyke, K. McMahon, S. Bertilsson, R. Stepanauskas and S. G. E. Andersson. 2013. Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. Genome Biology 4:R130.

Zhang, J., R. Chiodini, A. Badr and G. A. Zhang. 2011. The impact of next-generation sequencing on genomics. Journal of Genetics and Genomics Yi Chuan Xue Bao 38:95-109.