Prospects for the use of Next-Generation Sequencing Methods in Ornithology

Authors: Lerner, Heather R. L., and Fleischer, Robert C.

Source: The Auk, 127(1) : 4-15

Published By: American Ornithological Society

URL: https://doi.org/10.1525/auk.2010.127.1.4

BioOne Complete (complete.BioOne.org) is a full-text database of 200 subscribed and open-access titles in the biological, ecological, and environmental sciences published by nonprofit societies, associations, museums, institutions, and presses.

Your use of this PDF, the BioOne Complete website, and all posted and associated content indicates your acceptance of BioOne's Terms of Use, available at www.bioone.org/terms-of-use.

Usage of BioOne Complete content is strictly limited to personal, educational, and non-commercial use. Commercial inquiries or rights and permissions requests should be directed to the individual publisher as copyright holder.

BioOne sees sustainable scholarly publishing as an inherently collaborative enterprise connecting authors, nonprofit publishers, academic institutions, research libraries, and research funders in the common goal of maximizing access to critical research.
Nearly all aspects of avian biology, from behavior to conservation to systematics, have benefited from the application of molecular methods over the past two decades. New technologies of high-throughput DNA sequencing are sparking a revolution in the life sciences that is sure to further transform our understanding of avian systems. Next-generation (also called "massively parallel") sequencing methods were first introduced commercially just 5 years ago (Roche GS FLX; Margulies et al. 2005), yet their capacity to process millions of sequences in parallel, in contrast to the conventional 96-capillary capacity of Sanger sequencing, has rapidly placed them at the forefront of genetic research (Fig. 1). Although individual read lengths are currently limited (~500 base pairs [bp] for most platforms), the depth of coverage per base pair and advanced sequence-assembly software allow sequencing of 0.5–60 giga base pairs (Gbp). Thus, in a single run, depending on the platform (Table 1), it is possible to sequence anywhere from one half to 53 times the size of the chicken genome, or a minimum of 3,000 mitochondrial genomes. These methods extend far beyond simply genome sequencing and have already greatly benefited the field of biology, leading to advances in evolution (Shendure et al. 2005, Toth et al. 2007), epidemiology (Cox-Foster et al. 2007, Palacios et al. 2008), phylogenetics (Moore et al. 2006, 2007), comparative genomics (Romainov et al. 2009), microbial diversity (Edwards et al. 2006, Sogin et al. 2006, Roesch et al. 2007), DNA marker discovery, and studies of gene function and expression (Morozova and Marra 2008, Varshney et al. 2009).

These new tools make this an exciting time for ornithology because vast quantities of genetic data can be obtained affordably and with relative ease. With a little creativity, next-generation sequencing methods can be used to address new and long-standing questions previously inhibited by technological and financial limitations (e.g., obtaining enough sequence data to resolve phylogenies for adaptive radiations, study of avian major histocompatibility complex [MHC] while eliminating cloning bias, and whole-genome sequencing). At the very least, the huge increase in available DNA sequence will enable development of important markers for population genetic, systematic, behavioral, "evo-devo," and gene mapping studies.

The benefits of NGSMs in comparison to traditional capillary-based sequencing include (1) massive amounts of sequence data for a single or multiple individuals in a single run; (2) low cost per base; (3) reduction of the role of cloning and polymerase chain reaction (PCR) and, thus, reduced bias in resulting sequences; and (4) the ability to identify rare variant sequences rather than a single consensus sequence (see Fig. 1). There are three main drawbacks to NGSMs. The main obstacle at present, although this is beginning to change, is the relatively short read length for most platforms in comparison to traditional methods. For highly repetitive or complex genomes or genome regions, short reads (even at high coverage) may not suffice to generate an alignment. Thus, de novo sequence assembly is not always possible unless a reference or scaffold sequence of the same species or a very closely related species is available. Avian species, however,
Commercial next-generation sequencing methods can be distinguished by the role of PCR in library preparation. Four main platforms are amplification-based: Roche 454 GS FLX, Illumina Genome Analyzer IIx, ABI SOLiD 3 Plus System, and Polonator G.007 (Table 1). Two single-molecule sequencing methods (i.e., not PCR-based) are either very recently available or nearly available: Helicos Genetic Analysis System and Pacific Biosciences SMRT technology, respectively (Table 1).

Amplification-based Sequencing

Currently, the specifics of the chemistries and sample preparation for NGSMs are largely irrelevant because the machines are too expensive to buy or maintain for a single-researcher laboratory, with the exception of the Polonator G.007 and the recently announced 454 Junior, a next-generation bench-top pyrosequencer to be marketed in 2010 for small to medium labs. Sequencing facilities will perform much of the benchwork of sample preparation, which can be tricky depending on the platform, as part of the sequencing service for a modest fee. To choose the most appropriate platform for a project, some knowledge of the differences among methods is useful. Here, we briefly describe the methods and the unique benefits and drawbacks of each platform (see Table 1). There are several reviews that describe the technologies in greater detail, as well as animated demonstrations on the company websites (Hudson 2008, Mardis 2008, Shendure and Ji 2008; see Table 1 for company websites).

Clearly, the advantages of NGSMs will outweigh the few drawbacks for many avian studies and study systems. In fact, certain types of questions and large data sets may only be truly approachable using NGSMs. Given the relative simplicity of the laboratory techniques needed to generate next-generation sequencing libraries, we envision broad application of such methods in ornithology. Here, we provide a gateway for the avian biologist by comparing next-generation sequencing platforms, describing preparatory techniques to make such methods more broadly applicable, and discuss previous and potential applications in several subfields of avian biology.

Comparison of Next-Generation Platforms

Currently, the specifics of the chemistries and sample preparation for NGSMs are largely irrelevant because the machines are too expensive to buy or maintain for a single-researcher laboratory, with the exception of the Polonator G.007 and the recently announced 454 Junior, a next-generation bench-top pyrosequencer to be marketed in 2010 for small to medium labs. Sequencing facilities will perform much of the benchwork of sample preparation, which can be tricky depending on the platform, as part of the sequencing service for a modest fee. To choose the most appropriate platform for a project, some knowledge of the differences among methods is useful. Here, we briefly describe the methods and the unique benefits and drawbacks of each platform (see Table 1). There are several reviews that describe the technologies in greater detail, as well as animated demonstrations on the company websites (Hudson 2008, Mardis 2008, Shendure and Ji 2008; see Table 1 for company websites).

Commercial next-generation sequencing methods can be distinguished by the role of PCR in library preparation. Four main platforms are amplification-based: Roche 454 GS FLX, Illumina Genome Analyzer IIx, ABI SOLiD 3 Plus System, and Polonator G.007 (Table 1). Two single-molecule sequencing methods (i.e., not PCR-based) are either very recently available or nearly available: Helicos Genetic Analysis System and Pacific Biosciences SMRT technology, respectively (Table 1).

Amplification-based Sequencing

Currently, the specifics of the chemistries and sample preparation for NGSMs are largely irrelevant because the machines are too expensive to buy or maintain for a single-researcher laboratory, with the exception of the Polonator G.007 and the recently announced 454 Junior, a next-generation bench-top pyrosequencer to be marketed in 2010 for small to medium labs. Sequencing facilities will perform much of the benchwork of sample preparation, which can be tricky depending on the platform, as part of the sequencing service for a modest fee. To choose the most appropriate platform for a project, some knowledge of the differences among methods is useful. Here, we briefly describe the methods and the unique benefits and drawbacks of each platform (see Table 1). There are several reviews that describe the technologies in greater detail, as well as animated demonstrations on the company websites (Hudson 2008, Mardis 2008, Shendure and Ji 2008; see Table 1 for company websites).

Commercial next-generation sequencing methods can be distinguished by the role of PCR in library preparation. Four main platforms are amplification-based: Roche 454 GS FLX, Illumina Genome Analyzer IIx, ABI SOLiD 3 Plus System, and Polonator G.007 (Table 1). Two single-molecule sequencing methods (i.e., not PCR-based) are either very recently available or nearly available: Helicos Genetic Analysis System and Pacific Biosciences SMRT technology, respectively (Table 1).

Amplification-based Sequencing

Currently, the specifics of the chemistries and sample preparation for NGSMs are largely irrelevant because the machines are too expensive to buy or maintain for a single-researcher laboratory, with the exception of the Polonator G.007 and the recently announced 454 Junior, a next-generation bench-top pyrosequencer to be marketed in 2010 for small to medium labs. Sequencing facilities will perform much of the benchwork of sample preparation, which can be tricky depending on the platform, as part of the sequencing service for a modest fee. To choose the most appropriate platform for a project, some knowledge of the differences among methods is useful. Here, we briefly describe the methods and the unique benefits and drawbacks of each platform (see Table 1). There are several reviews that describe the technologies in greater detail, as well as animated demonstrations on the company websites (Hudson 2008, Mardis 2008, Shendure and Ji 2008; see Table 1 for company websites).

Commercial next-generation sequencing methods can be distinguished by the role of PCR in library preparation. Four main platforms are amplification-based: Roche 454 GS FLX, Illumina Genome Analyzer IIx, ABI SOLiD 3 Plus System, and Polonator G.007 (Table 1). Two single-molecule sequencing methods (i.e., not PCR-based) are either very recently available or nearly available: Helicos Genetic Analysis System and Pacific Biosciences SMRT technology, respectively (Table 1).

Amplification-based Sequencing

Currently, the specifics of the chemistries and sample preparation for NGSMs are largely irrelevant because the machines are too expensive to buy or maintain for a single-researcher laboratory, with the exception of the Polonator G.007 and the recently announced 454 Junior, a next-generation bench-top pyrosequencer to be marketed in 2010 for small to medium labs. Sequencing facilities will perform much of the benchwork of sample preparation, which can be tricky depending on the platform, as part of the sequencing service for a modest fee. To choose the most appropriate platform for a project, some knowledge of the differences among methods is useful. Here, we briefly describe the methods and the unique benefits and drawbacks of each platform (see Table 1). There are several reviews that describe the technologies in greater detail, as well as animated demonstrations on the company websites (Hudson 2008, Mardis 2008, Shendure and Ji 2008; see Table 1 for company websites).

Commercial next-generation sequencing methods can be distinguished by the role of PCR in library preparation. Four main platforms are amplification-based: Roche 454 GS FLX, Illumina Genome Analyzer IIx, ABI SOLiD 3 Plus System, and Polonator G.007 (Table 1). Two single-molecule sequencing methods (i.e., not PCR-based) are either very recently available or nearly available: Helicos Genetic Analysis System and Pacific Biosciences SMRT technology, respectively (Table 1).
<table>
<thead>
<tr>
<th>Company</th>
<th>Roche</th>
<th>Applied Biosystems</th>
<th>Illumina</th>
<th>Dover Systems</th>
<th>Helicos</th>
<th>Pacific Biosciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform</td>
<td>Genome Sequencer FLX</td>
<td>SOLID 3 Plus System (Supported Oligonucleotide Detection)</td>
<td>Genome Analyzer IIX</td>
<td>Polonator G.007</td>
<td>Helicos Genetic Analysis System</td>
<td>Not yet announced</td>
</tr>
<tr>
<td>Method of sequencing</td>
<td>Emulsion PCR of bead-bound oligos</td>
<td>Emulsion PCR of bead-bound oligos</td>
<td>Isothermal bridge amplification on flowcell</td>
<td>Emulsion PCR of bead-bound oligos</td>
<td>NA (single-molecule sequencing)</td>
<td>NA (single-molecule sequencing)</td>
</tr>
<tr>
<td>feature generation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequencing chemistry</td>
<td>Pyrosequencing using polymerase</td>
<td>Ligation (octamers, dual-base encoding)</td>
<td>Reversible terminator using polymerase</td>
<td>Ligation (nonamers, single-base encoding)</td>
<td>Asynchronous extension using polymerase (single-molecule synthesis)</td>
<td>SMRT technology (single-molecule synthesis)</td>
</tr>
<tr>
<td>Machine cost</td>
<td>List price $500,000</td>
<td>$600,000</td>
<td>$600,000</td>
<td>$170,000</td>
<td>$999,000</td>
<td>$99,000</td>
</tr>
<tr>
<td>Reagents cost per run</td>
<td>−$5,000</td>
<td>$3,000</td>
<td>−$4,000−6,000</td>
<td>−$1,000</td>
<td>−$18,000</td>
<td>Projected: 72 Mbp, 80,000 reads</td>
</tr>
<tr>
<td>Capacity (base pairs)</td>
<td>0.5 Gbp</td>
<td>60+ Gbp, &gt;1 billion tags per run</td>
<td>27.5–33 Gbp</td>
<td>4–5 Gbp, 8–10 Gbp with enrichment; 150 million mapable reads</td>
<td>28 Gbp, 400 million reads</td>
<td>286 bp average; 2,805 bp longest read</td>
</tr>
<tr>
<td>per run</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read length</td>
<td>400–500 bp (titanium chemistry) or−250 bp (standard chemistry)</td>
<td>50 bp</td>
<td>100 bp</td>
<td>13 bp (26-bp paired-end reads)</td>
<td>25–50 bp, average 35 bp</td>
<td>10–15 min minimum</td>
</tr>
<tr>
<td>Run length</td>
<td>10 h; 2 h data processing</td>
<td>3.5 days</td>
<td>2.5–9.5 days</td>
<td>4 days</td>
<td>8 days</td>
<td></td>
</tr>
<tr>
<td>Cycle length</td>
<td>35 s</td>
<td>90 min</td>
<td>8</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>Run subdivision</td>
<td>16</td>
<td>8</td>
<td>8 flow cells</td>
<td>8 lanes per flow cell, 2 flow cells</td>
<td>Application specific, as low as 50 pg</td>
<td></td>
</tr>
<tr>
<td>Input DNA requirements</td>
<td>500 ng (higher for titanium chemistry)</td>
<td>10 ng (fragment library); 5 μg for mate-paired reads; input DNA of 200–10,000 bp in length</td>
<td>10 ng (fragment library);</td>
<td>8 flow cells</td>
<td>Application specific, as low as 50 pg</td>
<td></td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Company</th>
<th>Roche</th>
<th>Applied Biosystems</th>
<th>Illumina</th>
<th>Helicos</th>
<th>Pacific Biosciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of prep time</td>
<td>3–4 days</td>
<td>2–4.5 days</td>
<td>2 days</td>
<td>1 day</td>
<td>&lt;12 h</td>
</tr>
<tr>
<td>Advantages</td>
<td>Long read length</td>
<td>Increased accuracy due to dual base calls, homopolymer reads sequence-able, very high output (Gbp)</td>
<td>No emulsion PCR</td>
<td>No clonal amplification, low error rate with dual pass</td>
<td>No clonal amplification, low cost, speed of run, direct RNA sequencing</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>Unreliable determination of homopolymer regions and large repeats</td>
<td>Short read length</td>
<td>Medium read length, long run time</td>
<td>Cost of machine</td>
<td>Cost of machine</td>
</tr>
<tr>
<td>Future developments</td>
<td>2010: 1 kb read length, high throughput, no cDNA intermediate</td>
<td>100 GB and 1,200 GB per run anticipated in February 2010</td>
<td>120 bp reads in early 2010, 250 bp paired-end reads in 2010</td>
<td>Rolling circle colonies, &quot;kolonies&quot;</td>
<td>Up to 10× improvement in throughput, no hardware change necessary, 18 months, direct RNA sequencing no cDNA intermediate in the first half of 2010</td>
</tr>
<tr>
<td>Platform-specific barcoding</td>
<td>12 barcodes</td>
<td>16 barcodes with 8 plate divisions × 2 plates = 256 tags per run</td>
<td>12 per flow cell</td>
<td>12 per flow cell</td>
<td>12 barcodes</td>
</tr>
</tbody>
</table>

*Information in this table was personally communicated by agents of the companies or is from other publications where noted.

b See section of text on amplification-based sequencing for explanation.

c Millar et al. 2008.
a variant of traditional PCR using the adapter sequence as priming sites or hybridization targets (Fig. 2D–E). Emulsion- or bead-based PCR (Nakano et al. 2003) is used in 454 pyrosequencing (Roche), SOLiD sequencing, and Polonator G.007 sequencing, whereas PCR employing primers covalently bonded to a flow cell (i.e., bridge PCR; Adams and Kron 1997) is used for the Illumina Genome Analyzer IIx. Following amplification, separate but parallel sequencing of each of the millions of single clonally amplified targets is performed on a substrate (Fig. 2F), for example within a micrometer-sized well on a microtiter plate (454 pyrosequencing) or directly on the tile of a flow cell (Illumina Genome Analyzer IIx).

At this point, the chemistries diverge dramatically, resulting in either large numbers of short reads (<50 bp) from Illumina, SOLiD, and the Polonator or smaller amounts of longer reads (250–500 bp) from 454 pyrosequencing. Both Illumina and Helicos use reversible dye-terminators in which a single nucleotide bound to a terminator is added by DNA polymerase and detected in real time by fluorescence, followed by removal of the terminator group (Ju et al. 2006, Mitchelson 2007). This method of "sequencing by synthesis" is continued by the addition of a different nucleotide (with terminator), and so on for a predetermined number of cycles. Pyrosequencing also uses a sequencing-by-synthesis method, but instead of reversible dye-terminators, pyrophosphate is released during nucleotide incorporation, fueling a downstream series of reactions that results in luciferase light emission (Ronaghi et al. 1998, Margulies et al. 2005). The strength of the luciferase signal is proportional to the number of nucleotides incorporated in a template so that regions of single nucleotide repeats (homopolymers) may be read with a single light pulse. The drawback to the pyrosequencing method is that the light signal reaches an asymptote with increasing length of the single nucleotide repeat region. Thus, the ability to determine the length of a homopolymer drops off the longer the chain gets, such that homopolymers of 8 bp or longer are not reliably determined and homopolymers of as few as 4 bp can be questionable in our experience.

Alternatively, sequencing by ligation is used in SOLiD sequencing (Brenner et al. 2000, Shendure et al. 2005), where oligos of all possible di-mers (bound as degenerate 8-mers) are ligated to the single-stranded template DNA and read by fluorescence corresponding to their unique di-mers. Terminal nucleotides and the fluorescent group are cleaved, which is followed by successive rounds of ligation, detection, and cleavage up to either 25 bp or 35 bp of template DNA (depending on run specifications). The unique dual-base interrogation method of SOLiD sequencing provides very high sequencing accuracy at lower cost than many of the other methods, making it a particularly good platform for SNP discovery. The higher output (6 Gbp) from SOLiD at a lower cost per mega base pair (Mbp) also makes this an attractive sequencing method for projects that can suffice with short reads.
The Polonator G.007 also uses ligation for sequencing, but a single base rather than two bases is interrogated per ligation attempt (Dressman et al. 2003, Shendure et al. 2005). The main benefit of Polony sequencing by ligation is that the cost of the machine and recurrent reagent costs are low. Furthermore, all associated protocols and software are open-source, which maximizes flexibility.

Single-molecule Sequencing

One single-molecule sequencing system is currently available commercially: the Helicos Genetic Analysis System, which uses technology developed by Braslavsky et al. (2003). Sequencing features are prepared for the Helicos library by the simple addition of poly-A tails to DNA templates, with no amplification of the template necessary. The template DNA is then hybridized to poly-T oligos anchored to a slide. Using a sequence-by-synthesis method (as described above) with reversible fluorescent dNTP terminators, the sequence of single-molecule templates is determined. By ligation of both a poly-A tail at the 3’ end and an adaptor to the 5’ end of the template, a template can be read twice, in this manner: after sequencing from the poly-A tail is completed, one strand can be removed by denaturation and the DNA can subsequently be sequenced from the 5’ end using the adaptor sequence as a priming site for initiation of synthesis. This bidirectional sequencing greatly improves accuracy (from 2–7% deletion error rate with one pass to 0.2–1% with two passes, and a raw substitution rate of ~0.001% for two passes). The simplicity of library preparation and complete independence from PCR or cloning (and related errors and biases; Kanagawa 2003, Arinas et al. 2005) make this a highly attractive option (see section on ancient DNA below). Single-molecule sequencing and faster methods of massively parallel sequencing are an active area of research and development (e.g., Oxford Nanopore and Pacific Biosciences).

The speed of the five previously described NGS Ms is limited because each addition of a different nucleotide species (G, T, A, or C) is necessarily separated in time by several steps, particularly nucleotide detection, washing away the previous nucleotide species, and introducing a new species for incorporation. For instance, a single cycle (i.e., steps from one nucleotide incorporation to the next) requires 1–3 h, depending on the platform (with the exception of Roche 454 at 35 s cycle−1). A single molecule–sequencing method under development by Pacific Biosciences (Eid et al. 2009), SMRT observes the real-time sequencing of a single template by using a chip of thousands of nanometer-scale chambers (zero-mode waveguides; Levene et al. 2003), each with a single anchored polymerase molecule. When the DNA polymerase incorporates nucleotides with phosphate-linked fluorescent labels, a window at the bottom of the zero-mode waveguide is used to detect the fluorescence signal. As polymerase incorporates the nucleotide, it naturally cleaves the phosphate group (including the phosphate-linked fluorescent label), and through random diffusion the next nucleotide is available to the polymerase. Because all four nucleotide species are present at any given time, this method is limited only by the polymerase’s rate of incorporation and the machine’s speed of detection. By using circularized templates and a strand-displacing enzyme, multiple independent reads can be obtained from each DNA template to improve the sequencing accuracy to the desired level. Read lengths using the SMRT technology will also not be as limited as in other methods, currently averaging 586 bp and reaching 2,506 bp in some cases. Such long read lengths make this a particularly promising method for de novo sequencing projects. Test platforms that employ the SMRT technology will be available for preselected test laboratories in January 2010, with commercial availability dependent on their success.

APPLICATIONS

Multispecies, Multilocus Studies (Multiplexing)

Next-generation sequencing methods were, for the most part, originally designed with funding from genome-sequencing initiatives. Thus, runs on most platforms can be subdivided only to a limited degree (8–50 divisions; Table 1), which allows only a few individuals (or otherwise unique samples) to be sequenced in a single run. Most ecological and evolutionary studies require homologous sequence data, sometimes from multiple genomic regions and often for many individuals. High levels of sequence divergence among samples could make it possible to separate samples simply on the basis of the resulting sequence alignments (Pollock et al. 2000). For individuals of the same or a closely related species, however, samples are better distinguished by attaching a unique sequence-based tag to each sample’s template before preparing the sequencing library. Both PCR-based and ligation-based techniques of tag attachment have been developed by independent researchers (Binladen et al. 2007; Meyer et al. 2007, 2008), and some additional platform-specific barcodes exist (e.g., 16-bar codes for the SOLiD 3 Plus System; Table 1). Tags can be attached either to short pieces of genomic DNA or to PCR amplicons and then pooled in equimolar ratios for preparation of the sequencing library. The pooling of tagged templates, called “multiplexing,” can be done at several levels in a single library. For instance, multiple PCR amplicons of different genomic regions from the same individual can be labeled with the same tag, pooled in an equimolar ratio, and then pooled again with similar libraries that bear other unique tags. Combining tags with subdivided runs can vastly increase the number of unique samples processed in a single run to hundreds or thousands.

Targeted resequencing, also called “genome partitioning” or “DNA capture,” is a revolutionary way to isolate large amounts of homologous sequence data from genomic DNA for downstream sequencing applications (Hodges et al. 2009, Summerer et al. 2009). Two primary methods exist, both based on hybridization to sequence-specific probes either in solution (Agilent and Invitrogen) or on a microarray. Such methods are particularly appropriate for studies using ancient DNA and fecal material, in which background or contaminating DNA would compete with target DNA in next-generation sequencing libraries, greatly reducing the number of useful sequences obtained (e.g., Briggs et al. 2009). At present, these methods are applicable only to single-species studies and have been applied predominantly to human studies, but development of multispecies probe applications is underway in several laboratories. Combining DNA capture with multiplexing followed by next-generation sequencing is likely to transform genetic research in the future.
Phylogenetics

The avian tree of life at all levels stands to benefit dramatically from the orders-of-magnitude more sequence data available from NGSMS. Many studies have shown that phylogenetic resolution and accuracy are improved by increasing the numbers of loci and taxa sampled (e.g., Zwickl and Hillis 2002, Rokas et al. 2003, Prasad et al. 2008). Previously unresolved avian relationships (e.g., placement of Cathartidae and Otididae; Hackett et al. 2008), particularly those that involve rapid evolutionary radiations (e.g., Hawaiian honeycreepers; Fleischer et al. 1998), may be more readily resolved with the application of NGSMS to obtain massive amounts of sequence. Such “phylogenomic” analyses of rapid radiations are likely to provide insight into processes of lineage sorting, whereas studying adaptive radiations with large data sets can also provide information about functional or adaptive variation. Phylogenomics can also be applied to selected groups (e.g., ducks, warblers, and hummingbirds) to study processes of hybridization.

To date, only a few studies (all non-avian; e.g., Moore et al. 2007, Willerslev et al. 2009) have employed NGSMS for phylogenetic analysis, and none has utilized the tagging methods described above. In a study that addressed the close relationships among extinct and extant rhinoceroses, Willerslev et al. (2009) used next-generation sequencing of whole mitochondrial genomes from the black, woolly, Javan, and Sumatran rhinoceroses. This data set included >16 kilobases (kbp) of mitochondrial data and provided strong support for sister relationships within the rhinoceros clade, something that studies based on smaller data sets had been unable to provide. However, higher-level relationships were not resolved with this data set, which suggests a hard polytomy in this rapid radiation that should be further investigated with data from nuclear loci.

As tagging methods become routine, next-generation sequence data of whole mitochondrial genomes and many nuclear loci are likely to become the standard in phylogenetic studies, so we may be perched on the verge of a phylogenomics era in ornithology.

Molecular Evolution and Comparative Genomics

Patterns and processes of sequence evolution—including evolutionary rates of nucleotide and amino acid change, nucleotide and codon usage bias, and insertion-deletion patterns—can be investigated to a deeper level with the massive data sets made possible by NGSMS. Next-generation sequencing methods also provide a way to investigate the unique patterns of evolution in avian-specific microchromosomes and macrochromosomes on a large scale. With the availability of multiple complete avian genomes, we anticipate new insights into the distribution of genes and recombination between the two chromosome types as well as evolution or generation of novel avian microchromosomes. Difficulties inherent to previous genome-sequencing technologies contributed to the incomplete status of the Red Jungle Fowl (i.e., chicken, Gallus gallus) genome. The missing sequence, predominantly on microchromosomes (5%), is being targeted with NGSMS at the Genome Center at Washington University.

Comparative avian genomics, facilitated by the ease and reduced cost of genome sequencing with NGSMS, has made great strides in recent years, leading to genomic resources for the chicken, Wild Turkey (Meleagris gallopavo), Zebra Finch (Taeniopygia guttata), California Condor (Gymnogyps californianus), and White-throated Sparrow (Zonotrichia albicollis) (Romanov et al. 2009). A comparative genomic approach is extraordinarily useful for identifying functional loci related to morphological, behavioral, and physiological variation and thus enables us to better understand the process of avian evolution. For instance, sequencing multiple genomes of diverse taxa from an adaptive radiation such as the Galapagos finches or Hawaiian honeycreepers may identify the genes responsible for particular bill morphologies and other phenotypic traits. Furthermore, genome sequences for threatened taxa can be useful in developing comprehensive conservation plans that increase genetic resistance to known threats. For example, Romanov et al. (2009) performed 454 cDNA sequencing of a fibroblast cell line in California Condors and found that 78% of the reads were homologous with chicken genes, mapping to nearly all chicken chromosomes. Further analysis of this data set and additional transcriptomes is expected to identify the mutation(s) responsible for a heritable embryonic lethal condition (chondrodystrophy). Genomic data appear to be useful for managing genetic diversity in the California Condor and the goal of establishing a viable, self-sustaining population. Indeed, it is hard to overestimate the potential benefits of comparative avian genomics to species conservation and the study of avian evolution. With the advent of NGSMS, avian comparative genomics will likely crescendo to the forefront of studies of avian evolution.

Conservation and Population Genetics

Next-generation sequencing methods can contribute broadly to conservation and population genetic studies of birds (Romanov et al. 2009). Although we believe that whole-genome sequencing will be used sparingly for such studies in the near future (mainly because such studies usually require large sample sizes of individuals but do not usually have the large budgets associated with medical or agricultural applications), NGSMS can still be very useful for developing variable markers such as microsatellites (Allentoft et al. 2009) and single-nucleotide polymorphisms (SNPs; e.g., Noevas et al. 2008, Vera et al. 2008) for non-model organisms. These variable markers can then be used for subsequent high-resolution analyses of population samples at lower cost.

Methods to obtain microsatellite sequences can include direct shotgun libraries or, preferably, enrichment procedures to increase the representation of microsatellite sequences within the pool of DNA fragments to be sequenced (Santana et al. 2009). The enrichment procedures followed by Santana et al. (2009) produced 873 microsatellite loci for three species (a wasp, a nematode, and a fungus) from a total of only 4.26 Mbp of sequence (14–28% of the contigs >100 bp contained a usable microsatellite locus). The amount of sequence they obtained is a bit more than can be sequenced on one 16th of a plate on a single 454 GS-FLX run or about one third of one 16th of a plate on a 454 Titanium run (at a cost of less than $1,500 at most 454 core facilities). Allentoft et al. (2009) were able to obtain microsatellite sequences from shotgun (unenriched) 454 sequence derived from DNA extracted from bone of a late Holocene moa, Pachyornis elephantopus. They used the sequences to design primers for a microsatellite locus that was then successfully amplified from DNA extracted from 52 bones representing three moa genera.
Discovery and screening of SNPs generally requires that one obtain sufficient coverage (at least 5–10 times), ideally from multiple individuals (often via resequencing), to identify polymorphisms. This can be a problem for taxa with large genomes (i.e., most eukaryotes) because of the cost required to obtain full genomic sequences at that level of coverage. Thus, discovery of SNPs can be assisted by reducing the pool of homologous DNA sequences, thereby increasing the coverage level. For instance, the pool can be limited by using only transcribed sequences (the transcriptome or expressed sequence tags [ESTs]), or by restriction enzyme digestion of multiple individuals and size selection (van Tassell et al. 2008). Vera et al. (2008) used the EST method by pooling cDNA from 80 individuals from eight families of fritillary butterflies. This method produced >600,000 EST sequences of ~110 bp in average length on the 454 GS-FLX platform. They generated 355 contigs totaling ~60,000 bp with at least 6× coverage from the ESTs and were able to identify 751 SNPs, 149 of which involved nonsynonymous polymorphisms. Another recent study (Novaes et al. 2008) generated 148 Mbp of EST sequence from Eucalyptus grandis cDNA using 454 GS runs and was able to identify 23,742 high-confidence SNPs from 71,384 contigs averaging ~250 bp in length. Single-nucleotide polymorphism microarrays are available for chicken (Burt and White 2007, Schmidt et al. 2008) and Zebra Finch (Naurin et al. 2009). With >3 million SNPs identified from genomic sequences, microarray analyses with a 3-kbp SNP array from G. gallus or T. guttata can provide good coverage of the genome for association studies of diverse avian species. Given the increased length per read on the 454 with current (and near future) technology, even greater recovery of SNP sequences is expected, and we believe that application of these NGSMs can greatly enhance studies in population and conservation genetics of birds.

Disease Diagnosis and Analysis

Avian disease is of broad interest because of its potential impact on human health, its importance for conservation of rare species, and its potential for insight into the coevolution of hosts and pathogens. Because research on infectious diseases is typically well funded and involves relatively small genomes, NGSMs have been applied extensively to sequencing genomes of a range of pathogens. These include viruses such as avian influenza (Höper et al. 2009) or avian bornavirus (Gancz et al. 2009), bacteria such as MRSA Staphylococcus (Highlander et al. 2007) and avian Mycobacterium (Pausterian et al. 2008), and protozoans such as Toxoplasma gondii (Bontell et al. 2009). These methods have also been used in genome-sequencing or genetic-association studies to identify mutations responsible for genetic disorders or diseases (ten Bosch and Grody 2008, Vasta et al. 2009). In addition, researchers are beginning to apply NGSMs to diagnosis or identification of pathogenic organisms by random screening (Nakamura et al. 2008, 2009; Adams et al. 2009; Jones et al. 2009) and by targeted methods using PCR amplicons (Jordan et al. 2009). Targeted, tagged pyrosequencing methods have also been used to assess variation in host immune-system genes, such as those of the MHC (Babik et al. 2009, Bentley et al. 2009, Wiseman et al. 2009) and immunoglobulins (Glänville et al. 2009). Thus far, very few NGSM studies have addressed pathogen–host relationships within birds, with the exception of a few studies on avian influenza (e.g., Höper et al. 2009) or diseases of domestic fowl (e.g., Paustian et al. 2008, Spatz and Rue 2008, Gancz et al. 2009).

Gene Expression and Transcriptome Analyses

One of the most powerful new applications of NGSMs is in measuring gene expression (Nielsen et al. 2006, Torres et al. 2007, Morozova et al. 2009). Most NGSMs produce high average coverage (i.e., number of sequence copies) per nucleotide site, providing an estimation of the frequency of any particular DNA molecule in an overall DNA pool. In most expression analyses, mRNA is isolated from developing organs or from tissues that have undergone differential treatment (e.g., pathogen-infected vs. non-infected individuals), and the RNA is reverse transcribed in vitro into cDNA. With appropriate experimental controls and an assumption that the reverse-transcription process does not alter the relative frequencies of the various mRNA transcripts, the relative depth of coverage of the sequence should be proportional to the expression of the particular gene. Some early analyses (e.g., Torres et al. 2007) suggested that there were biases in the representation of certain, usually smaller, transcript sequences. But later analyses showed ways to avoid (e.g., size standardization by nebulization) or correct for these biases and revealed NGSMs as a powerful tool for gene expression and evo-devo studies (Weber et al. 2007, Morozova et al. 2009). Direct RNA sequencing (i.e., without reverse transcription to cDNA) is a promising new application of the Helicos platform (Table 1) that will surely advance gene expression and transcriptome analyses.

Next-generation sequencing methods are well positioned to replace the use of more standard methods for measuring gene expression (Northern blots, real-time RT-PCR, or microarray analysis). We anticipate their use in studies of avian development such as those recently conducted by Abzhanov et al. (2004, 2006). These studies revealed, via comparative evaluations of expression levels of candidate genes in developing beak tissues and microarray analyses of expression, that the gene Bmp4 is up-regulated in Darwin's finches with larger, thicker bills and that the calmodulin gene is up-regulated in Darwin's finches with longer, narrower bills. We also expect to see their use in studies that evaluate changes in gene expression in hosts or vectors in response to infection by avian pathogens, such as avian malaria or avian influenza, much as has been done for pathogens of agricultural or medical importance (e.g., chestnut blight; Barakat et al. 2009).

Ancient DNA

Some of the earliest applications of next-generation sequencing involved sequencing of ancient materials such as subfossil bones, hair from mammals, and ancient soil samples (Poinar et al. 2006, Gilbert et al. 2007). Initially, these studies involved random, shotgun sequencing, but soon targeted amplification was applied to obtain full mitochondrial genomes from ancient samples (Briggs et al. 2009, Stiller et al. 2009). Most of the applications in vertebrates, thus far, have involved mammals (e.g., Noonan et al. 2005, 2006; Poinar et al. 2006; Willerslev et al. 2009), the only published use with ancient avian materials being a study on moa microsatellite development (see above; Allenoff et al. 2009).

A recent phylogenetic study (Gilbert et al. 2008) showed that sequencing of full mtDNA genomes with 454 pyrosequencing...
was able to resolve two divergent clades of wooly mammoth, 1.1–1.7 million years apart, much like the pattern found for Asian elephants (Fleischer et al. 2001). Interestingly, PCR and Sanger sequencing of mtDNA control regions of 160 mammoths from Eurasia and North America suggest that this mixed pattern resulted from late Pleistocene recolonization of Eurasia by mammoth lineages that originated in North America (Debruyne et al. 2008). Two of the main obstacles in ancient DNA studies are separating the desired DNA from a high level of background DNA and removing PCR inhibitors. As described above, DNA capture methods will be useful for the former, and single-molecule sequencing could obviate the need for PCR entirely.

Other Creative Applications

Several interesting applications of NGSMs have been developed, in addition to those discussed above. One useful method involves assessment of prey items from (noninvasive) fecal material. An exciting recent paper by Deagle et al. (2009) used PCR with redundant primers to amplify fish and cephalopod 16S (mitochondrial) and 28S (nuclear) rDNA sequences from genomic DNA isolated from Australian Fur Seal (Arctocephalus pusillus) feces but separating blocking primers to keep the Fur Seal DNA from being amplified. The products were pyrosequenced using a 454 GS-FLX platform, and the aligned contigs were compared with databases of fish and cephalopod sequences to identify 54 species of bony fish, 4 species of cartilaginous fish, and 4 cephalopod species from 105 fecal samples. The coverage, assuming no bias in amplification occurred, should indicate the relative frequency of the prey items in the Australian Fur Seal’s diet. This study identified more diet species of Australian Fur Seals in a single analysis than were found over multiple years using traditional, extensive hard-parts analysis of samples. Comparing individuals from three distant study sites showed site-specific variation in both type and frequency of prey.

Another recent study (Soininen et al. 2009) compared the traditional DNA barcoding method and direct pyrosequencing of plant remains from stomach samples of two rodent species but found that the barcoding via PCR outperformed the pyrosequencing in identifying a range of plant taxa. We feel that a tagged, targeted pyrosequencing; gene capture; and assessments of gene expression through cDNA coverage—are currently being developed. And the first whole genomes are being produced by NGSMs, with many more proposed for a huge suite of non-model organisms, including a large number and wide sampling of birds (Genome 10K Community of Scientists 2009). We predict that there will be a major acceleration in the near future in the application of NGSMs to ornithology and that many important findings in avian biology will arise from such studies. It is indeed an exciting time for ornithology.

ACKNOWLEDGMENTS

We thank M. Meyer, M. Hofreiter, F. Hailer, D. Locke, S. Schuster, and K. Helgen for elucidation and discussion of NGSMs and the National Science Foundation (DEB-0643291 for funding). O. Ryder and A. R. Hoelzel provided valuable comments on an earlier version of the manuscript. Online resources for analysis of next-generation sequencing data are available at the following websites: seqanswers.com/forums/showthread.php?t=43, and http://www.sanger.ac.uk/resources/software/.

LITERATURE CITED


Received 30 November 2009, accepted 12 December 2009.