Harnessing genomics for evolutionary insights

Antonis Rokas and Patrick Abbot

Department of Biological Sciences, Vanderbilt University, VU Station B 35-1634, Nashville, TN 37235, USA

Next-generation DNA sequencing technologies can generate unprecedented amounts of genomic data, even for non-model organisms. Here we describe how these new technologies have facilitated recent key advances in ecology and evolutionary biology, and highlight several outstanding ecological and evolutionary questions that are distinctly suited to the innovations they provide. Importantly, using these technologies to their full potential requires careful experimental design and critical consideration of several caveats associated with them. Although several significant challenges remain to be resolved before the integration of next-generation sequencing technologies into single-investigator research programs, we argue that they will soon transform ecology and evolution by fundamentally changing the ranges and types of questions that can be addressed.

High-throughput sequencing for the masses

Investigation of the DNA record is fundamental to several lines of ecological and evolutionary research. However, the majority of studies in both fields have so far utilized only tiny fragments of the DNA record. This has been the case primarily because genome-scale sequencing has not only been prohibitively expensive but has also required significant investments in infrastructure, well beyond the reach of single-investigator laboratories [1–3]. Change is in the air, however. Several different next-generation sequencing technologies (NGST) have recently become commercially available that dramatically reduce the per-base sequencing cost, while at the same time drastically increasing the number of bases sequenced [4–12]. The potential uses and applications of NGST span the spectrum of ecological and evolutionary biology research, and have already been put to use in epidemiology [13–15], experimental evolution [16], social evolution [17], paleogenetics [11,18–21], population genetics [21,22], phylogenetics [23,24] and biodiversity [25–29] (Table 1).

The key technological advance shared by all NGSTs is their unique ability to sequence DNA in a massively parallel fashion, with typical runs yielding three to four orders of magnitude more sequence compared to capillary sequencing [10]. This is accomplished by combining innovations in sequencing chemistry and detection of strand synthesis via microscopic imaging in real time. Specifically, the incorporation of each nucleotide base is detected through a light or fluorescent signal. Several competing technologies are already available (454 technology, Solexa, SOLiD and HeliScope), each differing in how the strands are synthesized and how base detection occurs (several excellent reviews that discuss them are available [9–12]). All technologies physically separate and fix the position of a large number of single-stranded DNA fragments on a substrate, and via repeated cycles of nucleotide incorporation and detection permit the simultaneous tracking of thousands to millions of sequencing reactions, in a process akin to photographing the twinkling of city lights over the course of a night.

Many are convinced that NGSTs do not simply represent incremental advances in sequencing technology. But to what extent will they advance the study of ecology and evolution? There are both features of NGSTs and evidence of their application indicating that they have the potential to fundamentally change how a large number of ecological and evolutionary questions are addressed. Here we critically evaluate NGSTs and their likely impact on ecological and evolutionary studies, and provide a brief overview of methods of analysis and caveats associated with their use.

The impact of NGSTs: pathogen hunts and the genomic signatures of behavior and lifestyle

Even though NGSTs have been commercially available for only a few years, their potential to transform ecological and evolutionary studies is evident. For example, the sensitivity offered by NGSTs has been a key asset in searches for the causative agents of infectious outbreaks. A telling example was the discovery, using 454 technology [5], of an Old World arenavirus as the cause of death in a cluster of human fatalities associated with transplants from a single donor [14]; the arenavirus had gone undetected in a series of standard diagnostic assays, including culture, PCR and microarray analysis [14]. NGSTs proved to be likewise useful in the investigation of the well-publicized colony collapse disorder of American honey bees by providing a common platform within which the relative importance of several candidate pathogens could be quickly and efficiently evaluated [13]. Using an NGST-based metagenomic approach, researchers compared the microflora of affected and unaffected hives, and identified the Israeli acute paralysis virus of bees as the most likely candidate pathogen responsible for the disorder [13].

NGST is capable of accurately detecting low-frequency DNA polymorphisms present within a single sample. Consequently, one discipline that is being transformed by NGST is population genetics, for which detection of such polymorphisms is often critical. Not surprisingly, several population genetics studies have taken advantage of NGST

Corresponding author: Rokas, A. (antonis.rokas@vanderbilt.edu).

0169-5347/$ – see front matter © 2009 Elsevier Ltd. All rights reserved. doi:10.1016/j.tree.2008.11.004 Available online 7 February 2009

Table 1. Example questions in ecology and evolution that have been addressed with NGSTs

Discipline	Example questions	Refs
Epidemiology	What is the etiology of emerging, uncharacterized infectious diseases?	[13,14]
Epidemiology	What are the core genomes of different microbes or pathogens occupying different niches or hosts?	[15]
Epidemiology	What are the mutations separating drug-resistant from sensitive pathogen strains?	[82]
Social evolution	Are maternal and sibling care behaviors regulated by similar patterns of gene expression?	[17]
Biodiversity	What is the microbial composition of different soil types?	[25]
Biodiversity	What is the extent of undocumented microbial diversity in different ecosystems?	[26,27]
Biodiversity	What is the microbial flora of the human body?	[28,29]
Phylogenetics	Can we reconstruct the phylogeny of a clade using organelle genome data?	[23,24]
Population genetics	What is the mutational profile of a eukaryotic genome?	[83]
Experimental evolution	What is the genetic basis of phenotypes emerging during laboratory evolution?	[16]
Palaeontology	What are the evolutionary relationships of ancient organisms to extant taxa?	[19–21]
Palaeontology	What was the microbial palaeoenvironment associated with permafrost-preserved organisms?	[18]
Developmental evolution	What are all the *cis*-regulatory targets of a transcription factor in a genome?	[48,49]

in genome-wide polymorphism discovery, including single-nucleotide [21,22,30,31], insertion–deletion (indel) [31] and copy-number variation [31].

NGST has also been successfully employed in testing theories of social evolution. For example, a long-standing hypothesis in the field is that cooperative brood care in social insects evolved out of maternal care in an ancestor [32,33]. A direct prediction of this theory is that patterns of gene expression in sibling care behavior should be positively correlated with the expression patterns of genes involved in maternal care [32]. This prediction was tested by 454 sequencing [5] of the brain transcriptome of *Polistes metricus* wasps, a species of primitively eusocial hymenopterans [17]. Comparison of the wasp transcriptome with the well-annotated honey bee genome led to the identification of 32 wasp genes homologous to those known to participate in worker honey bee behavior. Further characterization of these genes' brain expression patterns in the wasps revealed a positive correlation between the gene sets involved in sibling care and maternal care behaviors, thus providing support for the hypothesis that the molecular basis of cooperative brood care was likely ancestrally derived from molecules participating in maternal care [33].

Another area where NGST has already been applied is in the identification of the genomic signature of symbiosis. Most, if not all, eukaryotes are involved in some form of mutualistic or antagonistic symbiosis with bacteria [34]. Such symbioses have had a profound impact on eukaryote evolution, so much so that the genome architectures of symbiotic bacteria bear tell-tale signs of their adaptation to the symbiotic lifestyle. To test whether independent symbiosis events result in similar genomic signatures, 454 technology [5] was employed to sequence the genome of *Sulcia muelleri*, an obligate symbiont of sap-feeding insects and member of the Bacteroidetes clade [15]. Comparison of the *Sulcia* genome with those from the γ proteobacterial symbionts *Buchnera* and *Carsonella* revealed that symbiont genomes have independently evolved several distinct features, including small genome size, similar base composition and complementarity in metabolic capabilities, with that of their insect hosts and coresident symbionts [15]. This study illustrates how NGST can be used to better understand the genomic signature of diverse lifestyles and the evolutionary forces that shape genome architecture.

NGST for ecology and evolution; capturing diversity across genomes and clades

We are at the beginning of what many are calling a fundamental shift in biological research [4,9,35]. Several of the major questions in our field can be addressed with NGSTs, promising to provide key insights into long-standing questions (Table 1). Take, for example, the task of assembling the tree of life, a long-standing goal of evolutionary biologists. Less than 1% of species on the planet are included in any kind of phylogeny [36]. Considering the challenges associated with the design of primers that amplify DNA from distantly related organisms [37], and the labor-intensive nature of building the multigene matrices frequently required for robust inference [37,38], the abundance of data offered by NGST is likely to provide a major boost to the assembly of the tree of life. For example, use of 454 technology enabled Moore and colleagues to sequence the complete plastid genomes of two basal eudicot angiosperms within 2 weeks (although this estimate does not include the computational time required for data processing), allowing them to identify a deep evolutionary radiation at the base of the angiosperm tree [23,24]. The impact of NGST data on phylogenetic biology can extend beyond the accumulation of more data from more organisms, and can help to address long-standing questions in the field. For example, NGST data are well suited to evaluate the relative contribution of processes such as lineage sorting [39,40], hybridization [41] and horizontal gene transfer [42,43] to shaping the DNA record. Although evolutionists have long been studying these topics on a gene-by-gene and clade-by-clade basis, the opportunity to address them at the genome level and across life can for the first time be seized with NGSTs.

NGST can also help us further understand the evolutionary forces that shape genome architecture in eukaryotes, from the origins of new genes and gene clusters to the evolution of introns and repetitive elements [44]. For example, why are transposable elements or introns so unevenly distributed across life's genomes? Until now, the taxonomic distribution of sequenced genomes has been too sparse to enable precise testing of hypotheses about the origin and evolution of such key features of genomic complexity or to evaluate whether they arose through adaptive or nonadaptive means. Comparisons of the genomes of multiple independently evolved pairs of close relatives that differ in one or more such features using NGST are likely to

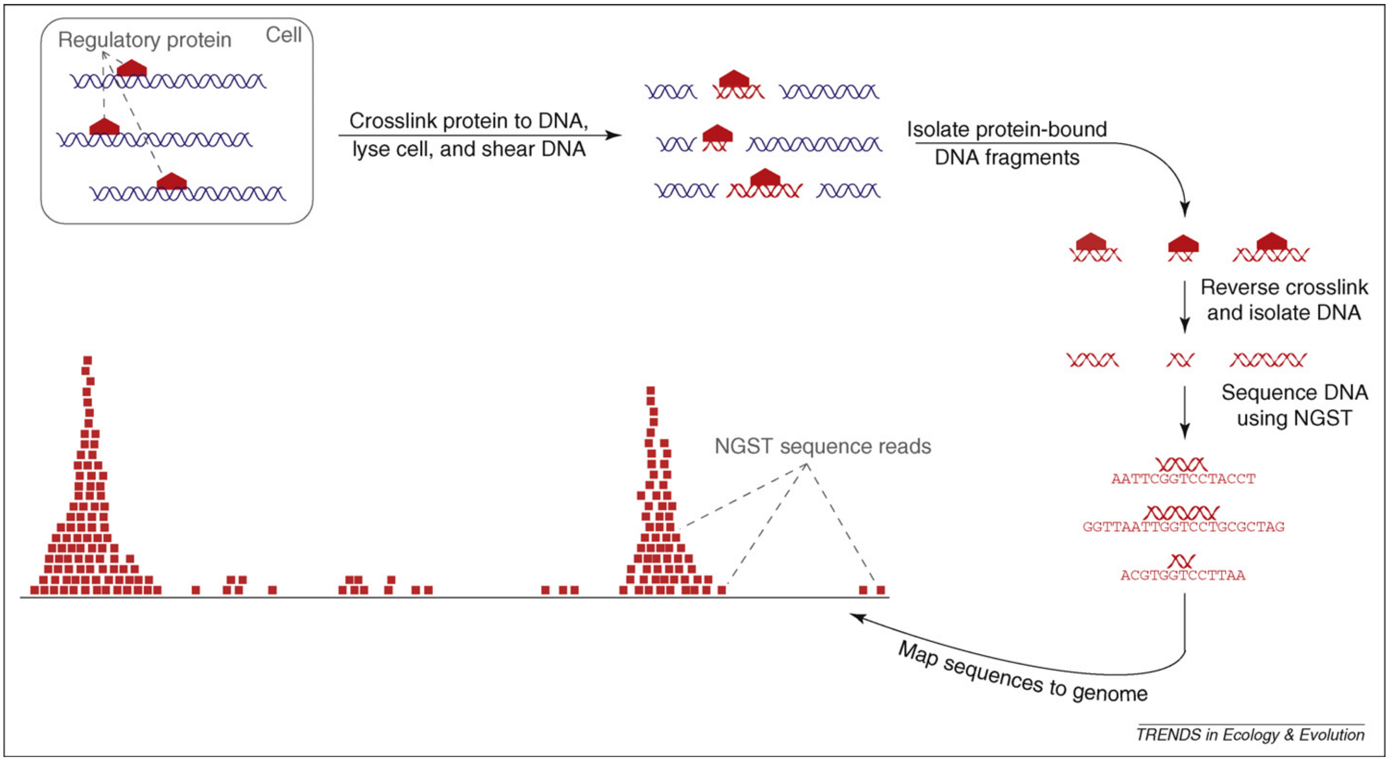
Regulatory protein
Cell
Crosslink protein to DNA, lyse cell, and shear DNA
Isolate protein-bound DNA fragments
Reverse crosslink and isolate DNA
Sequence DNA using NGST
AATTCGGTCCTACCT
GGTTAATTGGTCCTGCGCTAG
ACGTGGTCCTTAA
Map sequences to genome
NGST sequence reads
TRENDS in Ecology & Evolution

Figure 1. NGST can be very powerful in the genome-wide identification of the DNA binding sites of regulatory proteins. During chromatin immunoprecipitation (ChIP), all the DNA fragments in a genome on which a regulatory protein of interest binds are identified and isolated. This is achieved by crosslinking the DNA sites on which the regulatory protein binds with formaldehyde. The treated cells are then lysed, and all DNA (including the sites crosslinked to the regulatory protein) is sheared into small (200–300 bp) fragments. The next step is the isolation of only those DNA fragments that are protein bound, which is typically achieved by using a protein-specific antibody. Once the protein-bound DNA fragments have been isolated the formaldehyde crosslinks are reversed, the regulatory protein is removed and the DNA fragments are isolated. Sequencing of these DNA fragments using NGST (a technique known as ChIP-Seq) identifies the genomic locations that this specific regulatory protein binds to. For example, the final panel depicts two clusters of NGST-generated sequence reads that map to two distinct genomic locations and identify two protein binding sites (the peak heights of the clusters are a direct measure of how often the protein was bound to each genomic location). Comparison of the binding sites of a regulatory protein across a clade can identify species-specific binding differences and generate testable hypotheses as to their downstream phenotypic consequences.

become a powerful discriminator between alternative hypotheses of genome evolution and complexity [45].

Genomic resources are often missing for ecologically well-studied organisms. For example, the ecology and population biology of the Glanville fritillary butterfly *Melitaea cinxia* is well understood [35,46]; however, much of the ecology of *M. cinxia* has been studied in the absence of genomic resources, preventing more mechanistic approaches to understanding adaptation and phenotypic variability in this species. In a recent study, 454 technology [5] was used to rapidly sequence the transcriptome of *M. cinxia*, providing a genomic dimension to the work on *M. cinxia* and potentially removing a key obstacle toward a fully integrative understanding of functional traits.

NGSTs have also given rise to new experimental assays, pushing the application of NGSTs beyond simply sequencing genomes. One exciting use of NGSTs with much potential is in furthering our understanding of how changes in gene regulation can give rise to phenotypically novel forms. Chromatin immunoprecipitation (ChIP) is a powerful assay that identifies the actual DNA sequences bound by transcription factors [47]. Combining ChIP methodology with NGST (giving rise to what is affectionately becoming known as ChIP-Seq) has allowed the genome-wide identification of binding sites of transcription factors [48–50] (Figure 1). By comparing the genome-wide binding sites of a regulatory protein across a clade, differences in binding sites can be linked to their downstream phenotypic consequences and phenotypes. There is tremendous potential offered by ChIP-Seq in elucidating the role and contribution of regulatory elements to differences in form, or in explaining how the same genotype in different environments can yield manifestly different phenotypes at the whole-genome level.

To anyone wondering whether NGST has the potential to transform the way ecology and evolution studies are performed, these few examples should be tantalizing. In the end, it might well prove to be the imagination of evolutionary and ecological researchers which limits how rapidly NGST becomes integrated in the evolution and ecology research program. Genomics data remain largely underutilized in evolutionary and ecological studies, but the tremendous sequencing power offered by NGST can potentially lead to the addition of a genomic dimension to the study of the ecology and evolution of non-model organisms.

Design of NGST-based experiments

How then should evolutionary biologists and ecologists whose research could benefit by NGST get started? The four commercially available NGSTs (454 technology, Solexa, SOLiD and HeliScope [9–12]) differ with respect to several characteristics, but most importantly in the length of nucleotide read lengths they produce as output. For

Box 1. Powerful tools for NGST data processing and analysis

A major obstacle to the integration of NGSTs into the mainstream of ecology and evolution studies is that 'traditional' bioinformatics software is not always suitable to the requirements and specifications of the new data [22,58,68–73]. This is typically so because sequence reads are much shorter for NGSTs than capillary sequencing, making their assembly into larger contigs more challenging (for a review of the topic see Ref. [68]). However, several novel algorithms have been developed (see Table I; lists of NGST-relevant programs can also be found at http://seqanswers.com/forums/showthread.php?t=43 and at http://www.sanger.ac.uk/Users/lh3/seq-nt.html). Although their precision has not been extensively studied, the available data indicate that several short-read algorithms can be very accurate in assembling viral, bacterial and small-sized eukaryotic genomes [69–71]. For example, analysis of 2 700 000 Solexa sequence reads (each 36 bp long) from a 2 Mb sized bacterial genome with the Velvet algorithm resulted in an assembly that covered 97% of the genome with 99.98% accuracy [69]. Use of such algorithms on short sequence reads from larger genomes also generates accurate, but highly fragmented, assemblies. A solution to the problem of fragmentation might be the capability of all NGSTs to generate paired-end sequence reads [12], in which both ends of a sequence fragment of approximately known size are sequenced [68]; simulation studies have shown that paired-end sequence reads can yield highly accurate *de novo* assemblies of small eukaryotic genomes [70].

The performance of short-read assemblers can be further improved by combining data from technologies that produce short reads (e.g. Solexa, SOLiD or HeliScope) with those that produce longer reads (e.g. 454 technology, capillary sequencing) [70]. But *de novo* assembly is not always desirable; in cases where an assembled draft genome of a close relative is already available, programs have been developed that enable mapping of the sequence reads to the reference genome [72]. One such program is rmap [72], which allows the user to specify cutoffs for different mismatch and base quality scores.

Although the majority of NGST-related software has so far centered on issues of sequence assembly and alignment to annotated genomes, a variety of software tailored to the design and analysis of ecological and evolutionary data are also becoming available (Table I). For example, the MetaSim program allows the development and evaluation of candidate NGST-based experimental designs for questions employing genomic and metagenomic data [74]. Once NGST data have been obtained, researchers can identify single-nucleotide and indel polymorphisms using pyrobayes [58], or taxonomically annotate large-scale metagenomic data sets using MEGAN [73].

Table I. Software for processing and analyzing NGST data

Software	Description	Refs	Webpage
Velvet	*De novo* assembler that works with any technology; can be used to analyze paired-end reads or combinations of short and long reads generated by different technologies	[69]	www.ebi.ac.uk/~zerbino/velvet
MIRA2	*De novo* assemblers that work with any technology or combinations of short and long reads generated by different technologies		www.chevreux.org/projects_mira.html
euler-sr		[71]	euler-assembler.ucsd.edu
ALLPATHS	*De novo* assembler that works with any technology; produces assemblies that retain intrinsic ambiguities stemming from assay errors or polymorphism in diploid genomes	[70]	
rmap	Maps sequence reads to a reference sequence; allows specification of cutoffs for different mismatch and base quality scores	[72]	rulai.cshl.edu/rmap
maq	Maps sequence reads to a reference sequence and predicts the genotypes (for a diploid sample); works only for Solexa or SOLiD	[78]	maq.sourceforge.net/index.shtml
pyrobayes	Allows single-nucleotide polymorphism discovery from 454 data	[58]	bioinformatics.bc.edu/marthlab/PyroBayes
pbshort	Allows single-nucleotide and small-indel polymorphism discovery		bioinformatics.bc.edu/marthlab/PbShort
MEGAN	Allows evaluation and exploration of the taxonomic content of a metagenomic data set	[73]	www-ab.informatik.uni-tuebingen.de/software/megan
rast	Allows high-throughput analysis of genomic (rast) and metagenomic (mg-rast) data and evaluation of their phylogenetic and functional content	[79]	rast.nmpdr.org
mg-rast		[80]	metagenomics.nmpdr.org
MetaSim	Simulates genomic and metagenomic data; user can specify sequencing technology-specific error models, taxon composition and diversity	[74]	www-ab.informatik.uni-tuebingen.de/software/metasim
findpeaks	Identify enriched regions and the genomic locations of transcription factor binding sites from ChIP-Seq data	[81]	vancouvershortr.sourceforge.net
chipseq peak finder		[48]	woldlab.caltech.edu/html/chipseq_peak_finder

example, Solexa, SOLiD and HeliScope all produce large amounts of short reads (30–50 nucleotides), whereas 454 technology produces smaller amounts of longer reads (200–400 nucleotides). Because the reads produced by any NGST are all shorter than the reads generated by capillary sequencing (650–1000 nucleotides), the *de novo* assembly of large sequence contigs is computationally challenging (Box 1). Thus, deciding which NGST is most suitable is crucial and requires considering the question at hand, the size and part of the genome being sequenced and the availability of a draft genome of a close relative.

Experiments that require the analysis of large numbers of relatively small sequence fragments (e.g. the size of the protein-bound DNA fragments in a ChIP-Seq experiment is typically 200–700 base pairs) are likely better suited to technologies producing small sequence reads, because their assembly is much easier than that of a whole genome. But short read lengths can be utilized for the sequencing of larger genomic fragments, especially when a draft genome assembly of a close relative is available. For example, both simulation analyses [51] and real data [22] suggest that reads of 25–30 bp can be uniquely mapped to 60–80% of a typical large eukaryotic genome, like that of humans or nematodes, although distinguishing genuine polymorphisms from erroneous base calls can be challenging [52]. With draft genome assemblies of more and more model organisms becoming available, the phylogenetic proximity to study organisms favored by ecologists and evolutionists is increasing, thus enabling the use of short read length data in large-scale sequencing applications. By contrast, experiments that require *de novo* assembly of large genomic regions, such as sequencing a species transcriptome or genome, are better suited to 454 sequencing [46], currently the only NGST with the ability to generate sequence read lengths of a few hundred base pairs.

The design of an NGST-based experiment does not end once the best-fit technology has been chosen, however. Identifying the best study organisms for NGST-based experiments is a key step to maximizing their usefulness. Algorithms are available that identify suitable candidate genomes for sequencing which can maximize the potential of identifying evolutionarily conserved genomic regions [53] or optimal taxon sampling in phylogenetic experiments [54].

Finally, many ecological and evolutionary experiments require the processing of hundreds or thousands of samples. Both theoretical and experimental work suggest that tools exist that could harness the ability of NGST to generate massive quantities of sequence data for pooled samples (multiplexing), provided that they are sequenced at a sufficient coverage depth (see next section). If the samples to be analyzed are sufficiently divergent, theoretical work indicates that they can be analyzed by NGST and reassembled without any prior tagging [55]. Alternatively, experimental protocols that add forward and reverse oligonucleotide barcodes to individual samples for multiplexing are also available [56,57], and certain NGSTs (e.g. 454 technology) allow the sequencing run substrate to be subdivided into a small number of discrete regions (the two strategies, multiplexing and subdividing, can also be combined). This potential of NGSTs to simultaneously sequence thousands of samples in a single experiment harbors enormous potential for any ecology and evolution areas in which processing of a large number of samples might be critical for study accuracy (e.g. conservation genetics, phylogeography and phylogenetics) (Box 2).

Caveats of NGST data

Novel technologies bring novel challenges. A key parameter in the potential applicability of NGST to address questions in ecology and evolution is the accuracy of NGST-produced sequencing reads and associated quality values. This is a topic that, owing to the fledgling nature of the new technologies, has only recently come under scrutiny [10,22,58]. Traditionally, the accuracy of sequence reads has been assessed by 'quality scores,' values associated with the probability that the base called at a particular position in a sequence read is correct or not [59,60].

Accurate knowledge of quality scores is critical for the correct assembly of sequence reads as well as for interpretation of analytical results downstream. For reads produced by capillary sequencing, years of work have led to a firm understanding of error levels and means of accounting for them. Consequently, it has become possible to assess the quality of sequenced genomes and to estimate, for example, that the error rate in the finished human genome is less than 1 in 10 000 bp [61]. Thus, knowing that the distribution of single-nucleotide variants in the average pairwise comparison of the genomes of two humans is 1 for every 1000 bp, one is provided with a robust statistical framework within which to evaluate whether, say from a polymorphism detection experiment, the identified differences are genuine or the result of sequencing error.

Because the sequencing chemistry underlying each NGST is different, the error profiles for each NGST also differ, and quality scores need to be derived anew for each platform [62]. For example, base calling in 454 sequencing is done by reading the cumulative light signal emanating from the newly synthesized strand when the plate is inundated with one of four fluorescent nucleotides [5]. Hence, when the template strand has runs of single bases,

Box 2. NGSTs in practice: solving the mitochondrial DNA recombination riddle

The degree to which mitochondrial DNA (mtDNA) is maternally inherited is still a matter of debate [75,76]. Biparental mtDNA inheritance can result in heteroplasmic offspring (i.e. offspring that contain two or more distinct mtDNA genomes), which in turn can lead to the generation of mtDNA recombinants. Understanding the frequency of mtDNA heteroplasmy and recombination is critical for molecular evolution analyses and our understanding of human mtDNA diseases [75]. Progress on both has been hampered by a lack of methodology that permits efficient and accurate estimation of the degree of paternal leakage (i.e. paternal mtDNA being passed to the offspring) and recombination, especially when such events are rare. NGST methodology is very powerful in detecting low-frequency polymorphisms and thus is capable of definitively addressing key questions on mitochondrial inheritance (see Figure I).

Estimating paternal leakage using NGST

Two alternative scenarios of paternal transmission of mtDNA need to be considered: paternal genomes in low frequency are present in each individual offspring or, alternatively, very few offspring exhibit high frequencies of paternal genomes (see Figure I). Testing the first scenario requires sequencing large numbers of mtDNA genomes from one or very few individuals. A single run of 454 technology or alternatively a single lane of Solexa (both produce ~100 Mb of sequence) would allow the sequencing of ~6000 mtDNA genomes (assuming that each genome is the length of a typical 16 Kb animal mtDNA genome), enabling the detection of paternal genomes even when present at very low frequency within an individual. If even greater accuracy is needed, amplification or isolation of a specific 1 Kb fragment of the mtDNA genome followed by NGST sequencing would allow the sequencing of ~100 000 mtDNA sequence fragments. Testing the second scenario requires sequencing mtDNA genomes from a very large number of individuals [76]. Here the best approach would likely be to pool all mtDNA genome samples from all individuals (ensuring that each individual's sample is in approximately equal concentration) into a single sample, which is then NGST sequenced at high coverage depth. Note that virtually the same approach could be employed to assess the degree of mtDNA heteroplasmy (i.e. the presence of two or more distinct mtDNA genomes in an individual) across different animal tissues.

Assessing mitochondrial recombination using NGST

Capturing mitochondrial DNA recombinants using NGST is technically more challenging, because methods for the assembly of a set of admixed, closely related genomes are not yet available. However, the need for assembly can be circumvented, as in the following scenario (see Figure I). Imagine a male with a known mtDNA genotype is crossed with a female of a different, but also known, genotype. The question then is, what is the frequency of recombinant mtDNA genomes in the offspring? Amplification or isolation of any (or all) 200–250 bp regions that differ at more than two nucleotide sites between the parental mtDNA genomes from a single offspring, followed by sequencing with 454 technology, would result in 400 000 sequence reads that, because each is 200–250 bp long, fully cover these regions. After accounting for sequencing error, the frequency of sequence reads that contain variant sites from both parents in a single read should directly correspond to the frequency of recombinants in the sample. Alternatively, at the cost of decreasing detection sensitivity, the mtDNA samples from several different offspring could be tagged and sequenced in parallel.

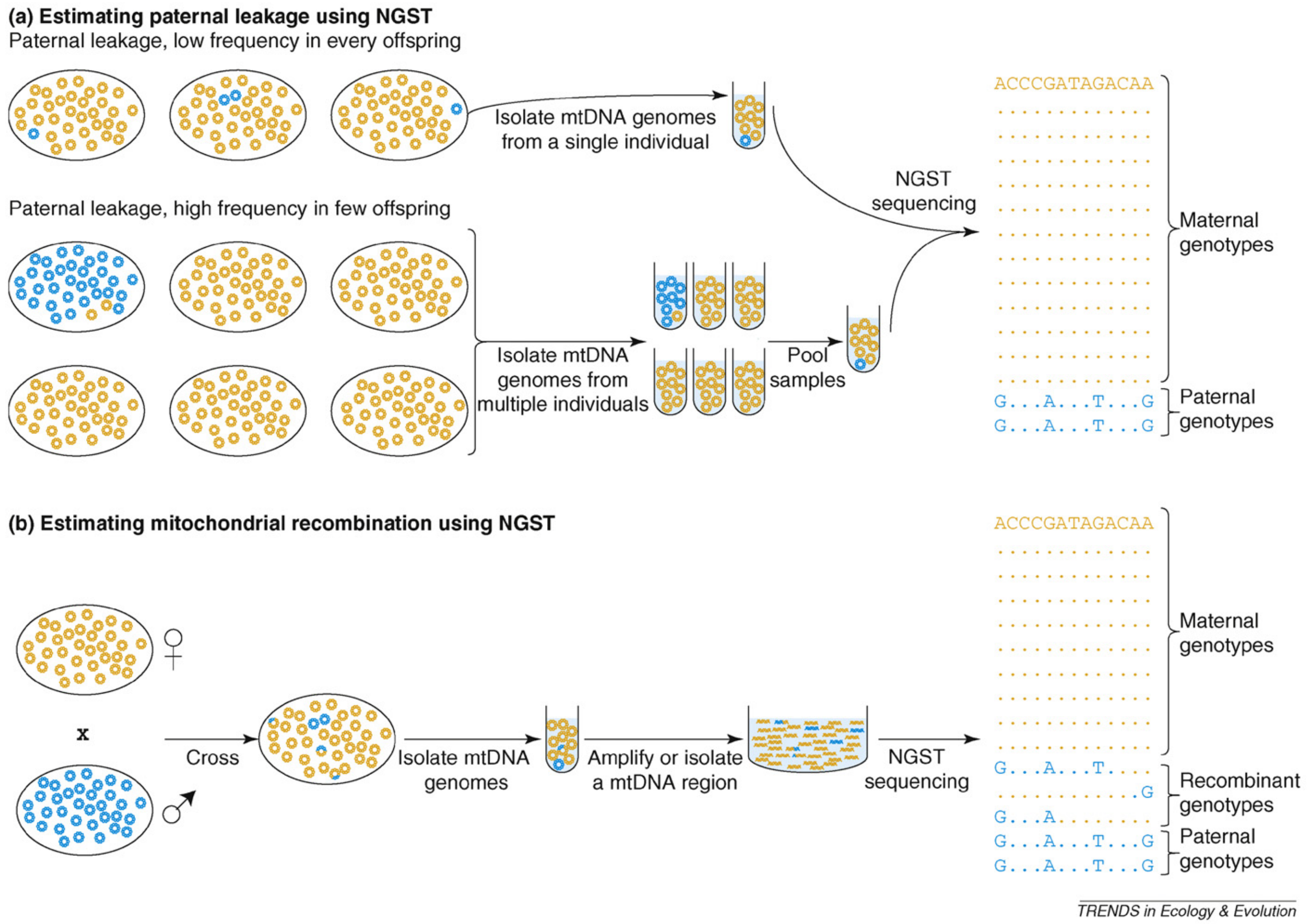
(a) Estimating paternal leakage using NGST
Paternal leakage, low frequency in every offspring
Isolate mtDNA genomes from a single individual
NGST sequencing
Paternal leakage, high frequency in few offspring
Isolate mtDNA genomes from multiple individuals
Pool samples
ACCCGATAGACAA
Maternal genotypes
G...A...T...G
G...A...T...G
Paternal genotypes
(b) Estimating mitochondrial recombination using NGST
ACCCGATAGACAA
Maternal genotypes
♀
x
♂
Cross
Isolate mtDNA genomes
Amplify or isolate a mtDNA region
NGST sequencing
G...A...T....
...........G
G...A.......
Recombinant genotypes
G...A...T...G
G...A...T...G
Paternal genotypes
TRENDS in Ecology & Evolution

Figure I. Elucidating mtDNA inheritance using NGST. Cartoon experimental design for estimation of levels of paternal leakage and mitochondrial recombination using NGST. **(a)** When paternal genomes are present at low frequency in each individual offspring, NGST can sample a very large number of mtDNA genomes from a single individual. Alternatively, when very few offspring exhibit high frequencies of paternal genomes, detection of paternal leakage requires pooling of mtDNA samples from multiple individuals prior to NGST sequencing. **(b)** NGST can also be employed to assess the frequency of mitochondrial recombination from a cross between individuals with distinct genotypes.

the most frequent errors associated with 454 sequencing are not miscalls of bases but stochastic artifacts caused by misalignment of sequences that result in undercalls (deletions) or overcalls (insertions) of bases [5,15,23,62,63]. This shortcoming, if left unaccounted for, can have a significant impact in evolutionary and ecological analyses. Importantly, until very recently, the quality scores produced by 454 were not incorporating errors introduced by deletions [5], and at least two independent groups developed quality score predictors specific to 454 technology that significantly increase the technology's accuracy [58,62].

In contrast to 454 sequencing, the main source of sequencing errors using Solexa sequencing is the actual miscalling of bases [62]. Solexa sequencing employs multiple cycles of differentially colored nucleotide terminators (a different color is attached to terminator nucleotides for each of the four bases) and image scanning [9,10]. Because Solexa sequencing relies on terminator chemistry, undercalls or overcalls of single-base runs are unlikely. Instead, the technology's main limitation is the reduced detection ability of the fluorescent bases that begins occurring after the completion of a large number of cycles [9]. The decreasing base-calling accuracy of Solexa sequencing as read lengths increase is well documented, evidenced by the negative correlation between the quality score of a particular base and the base's position in the sequence read, particularly for bases in positions 25 or more [22], and is precisely the reason why this technology generates sequence reads of relatively short length.

A final set of caveats applicable to all NGST data stems from the fact that NGSTs directly sequence the actual template. Direct sequencing of templates can be an advantage, in that it allows an unbiased measure of the starting DNA population used for sequencing (Figure 2). For example, the wide variation in levels of expression of genes in any organism will result in sequence reads whose coverage (coverage refers to the average number of sequence reads that contain any given nucleotide base of the DNA sample) mirrors this variation in expression levels. Specifically, sequence reads from highly expressed genes will be overrepresented, whereas reads from lowly expressed genes will be underrepresented or absent [64,65]. By contrast, sequencing of a genome will likely result in sequence reads that cover the genome sequence fairly evenly [22].

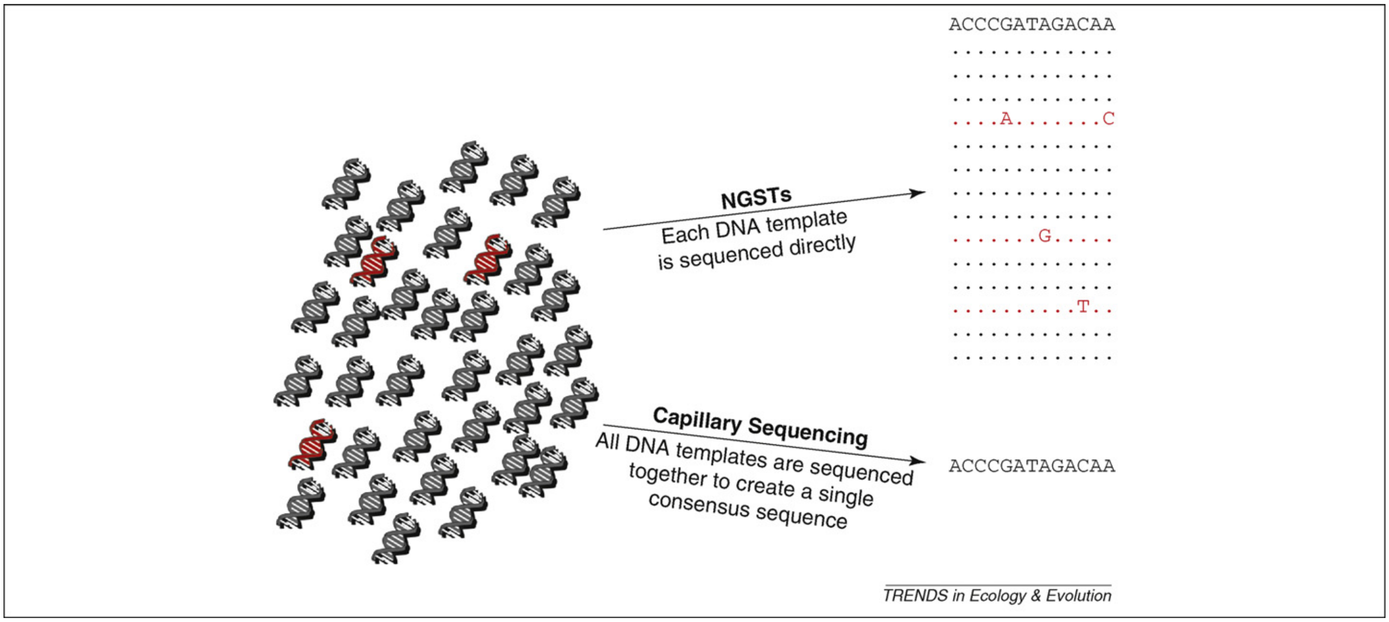

Figure 2. The quantitative nature of NGST can unmask biological variation as well as assay error. A sample composed of two kinds of DNA molecules (black and red) that has been generated by an assay is shown in the left panel. If the red molecules correspond to low-frequency biological variation, then the sequence variation observed in the sequence output of NGSTs is genuinely informative because each molecule is individually sequenced. However, if the red molecules correspond to low-frequency error associated with the assay that generated them (e.g. error due to Taq polymerase infidelity during PCR amplification [77]), then the sequence variation observed in the NGST sequence output simply reflects the error rate of the assay, a quantity that is generally not of interest. By contrast, both kinds of low-frequency variation (biological and due to error) are masked by capillary sequencing owing to the fact that all molecules are sequenced together.

However, direct sequencing of templates can also be a disadvantage that needs to be accounted for in certain experimental designs. Because any NGST will generate reads that are an unbiased measure of the DNA population being sequenced, it is expected that rare variants will rarely be sequenced. Thus, capturing and accurate determination of the frequency of rare variants require sequencing the DNA population at a very deep level of coverage. This was recently demonstrated in an experiment measuring the sensitivity of detection of heterozygous single-nucleotide polymorphisms from a diploid genome; the results showed that sensitivity of polymorphism detection varied dramatically with coverage [62]. Furthermore, if the actual DNA population is polymorphic because of assay error (e.g. polymorphisms introduced in the template to be sequenced due to a prior PCR amplification step or DNA contamination), then the sequence output of any NGST will be a measure of the error rate of the template [62] (Figure 2), requiring filters that eliminate the impact of the error-laden sequences [63].

What does the future hold?

More than 20 years ago, the inventions of PCR, molecular fingerprinting, and capillary sequencing put simple, user-

Box 3. Outstanding questions

Determining the distribution of thousands of single-nucleotide polymorphisms across multiple populations, the percentage of species that are the result of hybridization or all the regulatory elements bound by a regulatory protein across a clade are all questions that can in principle be addressed with NGST data. But several practical and theoretical issues remain, the answers to which will determine the speed with which NGST becomes integrated into the mainstream of ecological and evolutionary research. We discuss the two most important below.

The practical utility of NGST data for non-model organisms

In the short term, studies are needed that address the practical utility of NGST data from non-model organisms (for example, which are the genomic regions most likely to be recovered by *de novo* NGST sequencing, and what are their characteristics? What are the functional or evolutionary properties of transcripts sequenced in a transcriptome-profiling NGST experiment?). These need to be coupled with theoretical and simulation studies that explore and evaluate alternative NGST-driven experimental designs (e.g. what is the optimal coverage required per NGST for sampling a genomic region of a given size and which contains a given level of polymorphism?). In the long term, key unknowns that might prove decisive in the implementation of NGST in ecology and evolution studies will be whether NGST can generate *de novo* genome assemblies that match the completeness and accuracy achieved by capillary sequencing and how quickly new theory that integrates NGST data within the ecological and evolutionary theoretical framework becomes available.

Access to informatics infrastructure required for NGST data processing and analysis

Perhaps the greatest challenge associated with NGST is that their analysis requires specialized bioinformatics facilities and knowledge. For example, the sizes of the sequence data files generated from a single NGST experiment can range from hundreds of megabytes to tens of gigabytes (irrespective of which technology is used). The memory and speed requirements for downstream bioinformatic analyses of NGST data can also be very high, typically requiring supercomputing centers or cluster facilities. Even if these limitations were removed (and there is no indication that this will happen in the near future), it remains to be determined how quickly and to what extent the required information technology support and the bioinformatic training become widely available in mainstream ecological and evolutionary research.

friendly molecular technologies in the hands of those with the right questions, thus giving rise to entirely new disciplines of ecological and evolutionary inquiry (e.g. phylogeography, molecular systematics) [66]. Today, many are convinced that the integration of NGSTs into ecology and evolution studies will usher us into a similar period of radical change, enabling researchers for the first time to address questions the undertaking of which only a few years ago was, for all practical purposes, simply unimaginable.

Of course, genomics has never been short on promises, and several outstanding questions remain with regard to the speed and ease of assimilation of NGSTs into the mainstream of ecology and evolution, as well as their ultimate impact (Box 3). But like it or not, the genomes of non-model organisms are the new frontiers [67]. But why go there? A mundane reason is that it might soon be, if it is not already, more productive to combine or replace traditional methodologies with NGSTs. But a less prosaic reason might be the intrinsic value of exploring the vast empty genomic spaces of the tapestry of life. If so, then a key question becomes what role the community of evolutionary biologists and ecologists will play. Making sense of biological variation and complexity, and understanding the evolution of form and function, are, after all, the natural jurisdiction of ecology and evolutionary biology. NGST has ushered us into a new era of exploration, in which the complexity of organisms and their myriad traits has become that much more explicable, now that we can have the data we always wished we had, but thought we never would.

Acknowledgements

We thank Mark Borowsky and Ulrich Mueller for insightful comments on the manuscript. Research in A.R.'s laboratory is supported by the Searle Scholars Program, and the Genome Technology Core at Vanderbilt University Medical Center. Research in P.A.'s laboratory is supported by the National Science Foundation.

References

1 King, N. *et al.* (2008) The genome of the choanoflagellate *Monosiga brevicollis* and the origins of metazoan multicellularity. *Nature* 451, 783–788

2 Cuomo, C.A. *et al.* (2007) The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science* 317, 1400–1402

3 Merchant, S.S. *et al.* (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* 318, 245–250

4 Schuster, S.C. (2008) Next-generation sequencing transforms today's biology. *Nat. Methods* 5, 16–18

5 Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380

6 Bentley, D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* 16, 545–552

7 Lundquist, P.M. *et al.* (2008) Parallel confocal detection of single molecules in real time. *Opt. Lett.* 33, 1026–1028

8 Harris, T.D. *et al.* (2008) Single-molecule DNA sequencing of a viral genome. *Science* 320, 106–109

9 Hudson, M.E. (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Mol. Ecol. Resour.* 8, 3–17

10 Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141

11 Millar, C.D. *et al.* (2008) New developments in ancient genomics. *Trends Ecol. Evol.* 23, 386–393

12 Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135–1145

13 Cox-Foster, D.L. *et al.* (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. *Science* 318, 283–287

14 Palacios, G. *et al.* (2008) A new arenavirus in a cluster of fatal transplant-associated diseases. *N. Engl. J. Med.* 358, 991–998

15 McCutcheon, J.P. and Moran, N.A. (2007) Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19392–19397

16 Shendure, J. *et al.* (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309, 1728–1732

17 Toth, A.L. *et al.* (2007) Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science* 318, 441–444

18 Poinar, H.N. *et al.* (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* 311, 392–394

19 Noonan, J.P. *et al.* (2006) Sequencing and analysis of Neanderthal genomic DNA. *Science* 314, 1113–1118

20 Green, R.E. *et al.* (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature* 444, 330–336

21 Gilbert, M.T. *et al.* (2007) Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science* 317, 1927–1930

22 Hillier, L.W. *et al.* (2008) Whole-genome sequencing and variant discovery in *C. elegans*. *Nat. Methods* 5, 183–188

23 Moore, M.J. *et al.* (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol.* 6, 17

24 Moore, M.J. *et al.* (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19363–19368

25 Roesch, L.F. *et al.* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J.* 1, 283–290

26 Sogin, M.L. *et al.* (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc. Natl. Acad. Sci. U. S. A.* 103, 12115–12120

27 Edwards, R.A. *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7, 57

28 Marcy, Y. *et al.* (2007) Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl. Acad. Sci. U. S. A.* 104, 11889–11894

29 Gill, S.R. *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355–1359

30 Van Tassell, C.P. *et al.* (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* 5, 247–252

31 Wheeler, D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872–876

32 Linksvayer, T.A. and Wade, M.J. (2005) The evolutionary origin and elaboration of sociality in the aculeate hymenoptera: maternal effects, sib-social effects, and heterochrony. *Q. Rev. Biol.* 80, 317–336

33 Wheeler, W.M. (1928) *The Social Insects, Their Origin and Evolution.* Harcourt, Brace

34 Moran, N.A. (2006) Symbiosis. *Curr. Biol* 16, R866–R871

35 Ellegren, H. (2008) Sequencing goes 454 and takes large-scale genomics into the wild. *Mol. Ecol.* 17, 1629–1631

36 Cracraft, J. (2002) The seven great questions of systematic biology: an essential foundation for conservation and the sustainable use of biodiversity. *Ann. Mo. Bot. Gard.* 89, 127–144

37 Rokas, A. *et al.* (2005) Animal evolution and the molecular signature of radiations compressed in time. *Science* 310, 1933–1938

38 Rokas, A. *et al.* (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804

39 Ane, C. *et al.* (2007) Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.* 24, 412–426

40 Edwards, S.V. *et al.* (2007) High-resolution species trees without concatenation. *Proc. Natl. Acad. Sci. U. S. A.* 104, 5936–5941

41 Arnold, M.L. (1997) *Natural Hybridization and Evolution.* Oxford University Press

42 Ochman, H. *et al.* (2005) Examining bacterial species under the specter of gene transfer and exchange. *Proc. Natl. Acad. Sci. U. S. A.* 102 (Suppl. 1), 6595–6599

43 Doolittle, W.F. and Bapteste, E. (2007) Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* 104, 2043–2049

44 Lynch, M. (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl. Acad. Sci. U. S. A.* 104 (Suppl. 1), 8597–8604

45 Charlesworth, B. (2008) The origin of genomes—not by natural selection? *Curr. Biol.* 18, R140–R141

46 Vera, J.C. *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol.* 17, 1636–1647

47 Solomon, M.J. *et al.* (1988) Mapping protein DNA interactions *in vivo* with formaldehyde—evidence that histone-H4 is retained on a highly transcribed gene. *Cell* 53, 937–947

48 Johnson, D.S. *et al.* (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* 316, 1497–1502

49 Robertson, G. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* 4, 651–657

50 Albert, I. *et al.* (2007) Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* 446, 572–576

51 Whiteford, N. *et al.* (2005) An analysis of the feasibility of short read sequencing. *Nucleic Acids Res.* 33, e171

52 Johnson, P.L. and Slatkin, M. (2008) Accounting for bias from sequencing error in population genetic estimates. *Mol. Biol. Evol.* 25, 199–206

53 Eddy, S.R. (2005) A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* 3, e10

54 Geuten, K. *et al.* (2007) Experimental design criteria in phylogenetics: where to add taxa. *Syst. Biol.* 56, 609–622

55 Pollock, D.D. *et al.* (2000) A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. *Mol. Biol. Evol.* 17, 1776–1788

56 Parameswaran, P. *et al.* (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res.* 35, e130

57 Meyer, M. *et al.* (2008) Parallel tagged sequencing on the 454 platform. *Nat. Protocols* 3, 267–278

58 Quinlan, A.R. *et al.* (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat. Methods* 5, 179–181

59 Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8, 186–194

60 Ewing, B. *et al.* (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8, 175–185

61 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

62 Brockman, W. *et al.* (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.* 18, 763–770

63 Huse, S.M. *et al.* (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8, R143

64 Wilhelm, B.T. *et al.* (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453, 1239–1243

65 Nagalakshmi, U. *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–1349

66 Avise, J.C. (1994) *Molecular Markers, Natural History and Evolution.* Chapman and Hall

67 Collins, F.S. *et al.* (2003) A vision for the future of genomics research. *Nature* 422, 835–847

68 Pop, M. and Salzberg, S.L. (2008) Bioinformatics challenges of new sequencing technology. *Trends Genet.* 24, 142–149

69 Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829

70 Butler, J. *et al.* (2008) ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.* 18, 810–820

71 Chaisson, M.J. and Pevzner, P.A. (2008) Short read fragment assembly of bacterial genomes. *Genome Res.* 18, 324–330

72 Smith, A.D. *et al.* (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* 9, 128

73 Huson, D.H. *et al.* (2007) MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386

74 Richter, D.C. *et al.* (2008) MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS ONE* 3, e3373

75 Rokas, A. *et al.* (2003) Animal mitochondrial DNA recombination revisited. *Trends Ecol. Evol.* 18, 411–417

76 Birky, C.W., Jr (2008) Uniparental inheritance of organelle genes. *Curr. Biol.* 18, R692–R695

77 Cline, J. *et al.* (1996) PCR fidelity of *Pfu* DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.* 24, 3546–3551

78 Li, H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18, 1851–1858

79 Aziz, R.K. *et al.* (2008) The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9, 75

80 Meyer, F. *et al.* (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9, 386

81 Fejes, A.P. *et al.* (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* 24, 1729–1730

82 Koenig, R. (2007) Tuberculosis. Few mutations divide some drug-resistant TB strains. *Science* 318, 901–902

83 Lynch, M. *et al.* (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 105, 9272–9277

Forthcoming Conferences

Are you organizing a conference, workshop or meeting that would be of interest to *TREE* readers? If so, please e-mail the details to us at TREE@elsevier.com and we will feature it in our Forthcoming Conference filler.

27 May–1 June 2009
74th Cold Spring Harbor Symposium on Quantitative Biology, Cold Spring Harbor, NY, USA
http://meetings.cshl.edu/meetings/symp09.shtml

5–10 July 2009
International Congress of Systematic and Evolutionary Biology (ICSEB VII), Veracruz, Mexico
http://www.botanik.univie.ac.at/ICSEB7/

11–16 July 2009
International Congress for Conservation Biology; 23rd Annual Meeting of the Society of Conservation Biology 2009, Beijing, China
http://scb2009.ioz.ac.cn/index.asp?CFID=10602810&CFTOKEN=24167527
http://www.conbio.org/Activities/Meetings/

28–31 July 2009
First International Entomophagous Insects Conference, Minneapolis, MN, USA
http://www.cce.umn.edu/conferences/entomophagous/index.html