Introduction: studying diversity in an era of ubiquitous genomics

JAMES A. COTTON AND PETER D. OLSON

Just as scientific discoveries enable the development of new technology, novel technologies can drive scientific progress. Similar to the adoption of PCR as a mainstream laboratory technique in the 1990s, the ability to readily sequence whole genomes today has opened up new areas of biology and fundamentally changed the way people work in existing fields.

The most obvious feature of so-called 'next generation' sequencing (NGS) technologies (a misnomer that includes a wide array of platforms developed over the past decade) is the enormous increase in throughput of generating sequence data, resulting in an unprecedented reduction in cost. A single sequencing 'run' of a high-end platform can generate up to 5 billion reads and determine the sequence of 1500 billion bp of DNA–the equivalent of 500 human genomes–in three to four days. The US National Human Genome Research Institute has tracked the changing price of DNA sequencing they fund from about $5000 per Mb to 5 cents per Mb over the last 15 years: a 100 000-fold drop (see Fig 1). At the time of writing (2015) the sequencing equipment market is dominated by Illumina, and a relative lack of competition and the maturity of the current technology has at least temporarily slowed the fall in price. However, the development of newer sequencing platforms is expected to soon spark another era of rapidly declining prices and rising throughput.

Fig 1 The effect of next generation sequencing technologies on cost.

This enormous technological progress has been a boon for many areas of biology, but the change in technology has also required researchers to change the way they do science and has led to changes in the types of questions people can ask. Traditional sequencing required massive amplification of specific, targeted DNA sequences by PCR prior to sequencing. While it is possible to sequence PCR amplicons with high-throughput sequencing technology (and PCR is used in the sequencing process), the enormous throughput and short sequencing reads typically mean that the most cost-effective way of collecting even limited amounts of data of interest is by shotgun sequencing of the input DNA.

This untargeted nature opens up new kinds of science. For example, it is now possible–indeed, often technically easier–to assay entire genomes rather than investigate candidate genes. Similarly, the ability to sequence mixtures of DNA from multiple species has enabled the ability to use sequencing to investigate the genetic diversity of entire communities of organisms. By converting RNA to cDNA, NGS is also ideally suited to transcriptomic profiling (e.g. RNA-seq), giving high-coverage, quantitative, genome-wide estimates of gene expression that can be used to more easily and fully characterize the phenotype-genotype map by understanding how different genes are transcribed in space and time.

The origins of DNA sequencing are in the development of recombinant DNA techniques in the 1970s. Fred Sanger shared the Nobel prize in 1989 for inventing a practical chemical approach to DNA sequencing based on the use of 'terminator' dideoxy nucleotides that could be added to growing DNA molecules to block further extension (Sanger et al. 1977). The lengths of chains arrested at different nucleotides can then be sorted by gel electrophoresis to determine their sequence. An alternative approach, based on chemical modification of DNA to allow cleavage of the nucleotide chain at specific bases (Maxam and Gilbert 1977) was initially more widely used, but Sanger's method became the method of choice, partly as this chemistry became the basis for automated sequencing using fluorescent, rather than radioactive, terminators (Smith et al. 1986). Automated methods powered the sequencing of the first complete genomes of cellular organism to be published, from the first bacterium (Fleischmann et al. 1995) to the human, mouse and rice genomes a few years later.

The first commercially successful 'next generation' sequencing technologies (produced by companies such as illumina, 454 Life Sciences, Applied Biosystems, and Ion Torrent) are based on massively parallel sequencing of short DNA 'reads' from fragmented input DNA, and have now evolved to generate gigabases of sequence data with very low error rates. The details of the technology underlying the different platforms varies significantly (see Shendure and Ji 2008), but they all rely on producing a sequencing 'library' by randomly fragmenting input DNA followed by ligation of adaptors of known sequence that allow some kind of PCR-based amplification of single molecules. This amplification produces millions of immobilized clonal groups of molecules that can then be sequenced in parallel by recording the incorporation of bases (or pairs of bases) during the synthesis of one strand by a variety of different means (e.g. pyrosequencing). Upcoming technologies are based on sequencing single molecules directly, avoiding the need for ligation and PCR to prepare libraries, and promising much longer 'reads' of contiguous data. Current commercial technologies from Helicos, Pacific Biosciences and Oxford Nanopore achieve this at the cost of much lower throughput and accuracy—but this area is developing very quickly (Thompson and Milos 2011).

A corollary of the changes in sequencing technology, and the changing data they generate has been a change in the kinds of skills needed. Molecular systematists have long had to be able to use computers to interrogate data – for example, in constructing alignments and phylogenetic trees – but shorter reads and much larger datasets require much more sophistication in informatics. User-friendly 'packaged' software is starting to catch up with the needs of evolutionary biologists to analyse next-generation sequence data, but researchers need to have some understanding of the assumptions of the algorithms involved, and the ability to tailor analyses to the particularities of individual datasets. Bioinformatics expertise and even the ability to write simple computer programs are becoming key skills for contemporary molecular systematists.

Next generation phylogenetics

Modern molecular phylogenetics can trace its origins to the convergence of contemporary systematics, which had a new analytical rigor through the adoption of cladistics and tree-centric thinking, with the advent of PCR and Sanger sequencing, which were making molecular data accessible. The parallel development of explicit numerical methods for phylogenetic inference in the genetics literature and character-based parsimony and compatibility approaches by systematists (see Felsenstein, 2004) spawned a minor (and largely adversarial) industry in the development of tree building methods—while at the same time the generation of molecular data was becoming commonplace. Ultimately, the promise of sequence data to provide a record of evolutionary history independent of morphology gave a renewed interest in (and funding for) understanding organismal evolution in the 1990s, and modern statistical phylogenetic methodology using maximum likelihood or Bayesian inference became widely adopted. The field of molecular phylogenetics (e.g. Hillis, Moritz and Mable, 1994) was born and its impact was such that several dedicated journals soon followed, all of which remain active today.

The following two decades were dominated by phylogenetic estimates based on single or few genes (principally rDNA, rbcl and CO1) alongside considerable development of analytical methods, especially models of nucleotide substitution and automated methods of alignment. These influential early molecular systematic studies provided both new concepts of interrelationships (e.g. Aguinaldo et al 19??), as well as confirmation of existing ideas, and many remain our best estimates to date despite subsequent investigation. How the introduction of whole genome data—the ultimate source of heritable characters in the form of nucleotide sequences and the genetic elements they encode—will affect our current understanding of the tree of life is not yet clear as relatively few organisms have been fully sequenced and many methodological challenges remain. Nevertheless, as section 1 of the book shows, it is clear that NGS data are becoming integral to phylogenetics.

The major attraction of high-throughput sequencing technology for phylogenetics is the abundance of data this can generate to resolve previously intractable phylogenetic questions. As the perspective piece by Sanderson (Chapter 1) discusses, the amount of data needed to resolve a phylogeny increases with the number of taxa included, so large-scale phylogenetics will require large-scale sequencing efforts so that data from many loci are available for many species. As Sanderson discusses, scaling-up molecular phylogenetic datasets in this way introduces a number of methodological challenges (see Chapters 12 and 15). These methodological challenges also bring opportunities for learning about molecular evolution, and analysis of multi-locus datasets–spurred by the increasing capacity of DNA sequencing–is the major theme in contemporary phylogenetic theory and methods development.

The next two chapters present case studies, reviewing how NGS technology is driving progress in understanding relationships in two of the most diverse animal groups – the insects, reviewed by Hughes and Longhorn (Chapter 2) and nematodes, reviewed by Blaxter and colleagues (Chapter 3). In both of these groups, the representation of genomic data is highly taxonomically biased, partly by the presence of important model organisms (e.g. *Caenorhabditis* and *Drosophila*) and by the biomedical importance of vectors and parasites. There are differences between the groups – while multi-locus data has become widely used in the insects, the current phylogeny of nematodes is largely based on a single (ribosomal RNA) locus, and NGS is also becoming important in studying nematode ecology, as morphological identification is particularly challenging. As Hughes and Longhorn discuss, current genomic data covers only 10% of insect phylogenetic diversity, while well over half could have been covered if an optimal choice of sequencing targets had been made over the last few years. Both chapters discuss how co-ordination of future genome sequencing targets within these research communities is desirable, for example via formal consortia or more informally online.

A significant issue in adopting next-generation sequence data for phylogenetic use is how best to leverage this technology to generate smaller, more targeted datasets for many samples: key to making economic use of the technology. The conventional approach would be to create sequencing libraries for each sample independently, including oligonucleotide index tags to identify reads derived from each library. This approach has drawbacks: as sequencing throughput rises, the costs of the molecular biology steps involved in creating libraries becomes a significant component of the overall costs of a sequencing experiment. Building on decades of interest in using abundant, universal mitochondrial markers for animal phylogenetics, Foster et al. (Chapter 4) describe their experiments with one attractive approach, mixing together the DNA of the samples before library construction and sequencing, and then reconstructing assemblies for each input sequence *in silico*. They argue that, as sequencing data becomes cheaper, using this approach to sequence entire mitochondrial genomes will provide an 'ultimate barcode' for unambiguous species identification that will also allow accurate phylogenetic reconstruction (see also this volume, Chapter 8).

The first completed genomes of cellular organisms where for bacterial pathogens, and the small genome size of many prokaryotes has kept bacteriology at the forefront of genome biology. Finally in this section, Bryant and Harris (Chapter 5) review how NGS and parallel developments in informatics are enabling the study of bacterial pathogen evolution in unprecedented detail. The cost of bacterial genome sequencing is now low enough that genomics can be used as a clinical tool, for example using phylogenetic methods to identify transmission chains or methods to detect recombination that identify events leading to vaccine escape. Their work shows how high-throughput sequence data can dramatically improve the resolution of phylogenies, allowing novel, population-level inferences about evolutionary processes such as recombination, mutation and migration. As sequencing costs continue to fall, similar techniques will become widespread for organisms with much larger genomes.

Next generation biodiversity science

The ability to sequence complex mixtures of DNA in an untargeted ‘shotgun’ way has opened up a large field of research using sequencing to investigate biodiversity. Broadly speaking, two approaches have emerged – one is to isolate some specific conserved marker to use as a ‘barcode’ to identify the species present in an environment by comparing against a set of sequences from the same locus from some known species. This allows a rapid assessment of the species present, and the amplification or capture step involved in this approach also allows the targeting of particular taxa (Fonseca et al. 2010). Metagenomics aims to sequence all of the DNA present in a particular environment, in principal allowing both characterization of the organisms present (by comparison to reference genomes) and some insight into the biological processes that might be occurring in an environment.

We start this section with a perspective piece taking the long view – of how developments in sequencing technology promise to allow DNA sequencing in much smaller packages – the first ‘pocket sequencer’ is on the verge of commercial availability. Bateman (Chapter 6) argues that the ability to sequence DNA in the field will provoke a much-needed renaissance in field biology by strengthening links between professional taxonomists, molecular systematists and practitioners such as ecologists and para-taxonomists who actually perform most species identifications in the field. Bateman's vision is that combining digital imaging, GPS co-ordinates and sequence data for field specimens will allow practitioners to identify samples with unprecedented accuracy, while the data they gather in the process will feed into more accurate species description and circumscription. Something like this vision will surely, eventually, come to pass.

Bik and Thomas (Chapter 7) review the contemporary scene in environmental sequencing, focusing in particular on the analysis challenges introduced by the large amounts of data being generated in biodiversity sequencing projects, particularly as both bioinformatics approaches struggle to keep up with both the rapid changes in sequencing technology and the changing questions biologists want to ask of their data. For example, ecologists are increasingly attempting to investigate ecosystem function as well as community composition using sequence data. This is complemented by the chapter by Hajibabaei and King (Chapter 8), who review both the practicalities of environmental sampling, sequencing and analysis and applications of NGS in both ecology research and in more applied settings. They emphasise that collaboration is key to making the most of the data being collected—both between disciplines so that informaticians and molecular biologists develop tools appropriate to the questions ecologists want to answer, and bringing together workers on ecologically and geographically diverse habitats to ensure data is comparable between sites. Both chapters emphasise the need to carefully adapt existing barcoding approaches to the short read lengths and higher error rate of current NGS platforms.

Finally, Bass and Bell (Chapter 9) describe how these approaches have impacted on our understanding of the biology of protists – the understudied paraphyletic assemblage that represents the vast majority of eukaryote diversity. The difficulty of culturing many protist lineages has meant that even obtaining 'specimens' for traditional morphological or molecular investigation is difficult and taxon delimitation is challenging. They present a case study illustrating how combining phenotypic and molecular datasets from all culturable lineages in a group with environmental sequence data from environmental samples provides a powerful platform for studying both the ecology and evolution of challenging microbial groups, empowering both the discovery and description of extensive, previously hidden, diversity.

Next generation challenges and questions

Section 3 broadens the scope of the book beyond the core disciplines of systematics to show how NGS data are being used to address questions on development, gene evolution and using ancient DNA. The section begins with a perspective by Rokas (Chapter 10) on the ways in which NGS has changed the types of questions that comparative biologists can ask, and thus the practice of systematics, arguing that 'tree thinking' is more important than ever in our attempts to put NGS data to use across a myriad of purposes. Drawing on examples from his own research on fungal evolution and referencing chapters in this volume and elsewhere, he provides a concise introduction to the diversity of questions being addressed in the NGS era.

Another perspective by Sommer (Chapter 11) describes an evo devo approach that centres on elucidating the genotype-phenotype map by integrating evolutionary history with population and developmental genetics. In the age of NGS, such integration is readily possible for most any organism, making our reliance on major model systems something that we can begin to move away from. By example he briefly introduces the *Pristionchis pacificus* model system that provides the only significant comparator to *C. elegans*, showing how similar phenotypes can be derived from evolutionarily independent means. Later, Walker et al. (Chapter 13) introduces botanical evo devo, discussing the evolution of petal spots and the different ways that NGS data can be used to identify trait-related markers: finding a needle in the haystack has never been easier. More than described in previous sections, evo devo makes use of expressed sequence (mRNA et al.) analysis (transcriptomics) for understanding the on/off states of genes in time and space—and NGS is perfectly suited to this task.

Nelson and Buggs (Chapter 12) introduces one of the most fundamental unanswered questions raised by the 'deluge' of new genomic data: the ever increasing presence of 'orphan', or taxonomically restricted, genes (TRG). Every genome studied to date has revealed an unexpected percentage of unique genes, and increased sampling of 'gene space' has resulted in this space expanding, rather than contracting as would be expected if genes only appeared to be orphans due to a lack of sampling. Moreover, there is strong evidence that many or most such genes are in fact functional, making them far more than genetic baggage. The authors provide a thorough dissection of the subject, explaining how relatedness is essential to defining TRG and how such loci can be in turn of value to systematics. The case for their further study is convincingly made.

In Chapter 14 Smith et al show how historical botanical specimens can provide empirical snapshots of genetic diversity in the past—something that must be typically inferred from contemporary data—as well as reveal the genetic identities of associated pests and pathgogens in a historical context. The highly fragmented and sometimes altered nature of genetic material in historical specimens has always been a major problem for generating accurate, representative sequences, but this is no great impediment for NGS. Moreover, the authors explain how RNA, despite its ephemeral role in the cell, is often recovered in as high yields as gDNA (at least in seeds), making possible transcriptomic analyses from historical samples. Thanks to NGS, the genetic record preserved in herbaria, seed banks and other botanical collections can be read for the first time, and with single-molecule sequencing, even more of that text is available for analysis.

In the final chapter 15, Cotton gives a comprehensive account of the analytical steps involved in dealing with NGS data in phylogenomics, aiding the reader in understanding the differences between 'traditional' (few loci) molecular phylogenetic analysis and NGS-driven, multi locus analyses. He highlights that, beyond purely technical challenges due to the size and complexity of NGS datasets, the advent of massively multi-locus data has driven new perspectives on how phylogenetic inference can be carried out, and opened up new questions that can be addressed with comparative molecular sequence data.

Acknowledgements

We acknowledge the support of the Systematics Association and Megan Waddington of Cambridge University Press. We wish to thank a large number of peer reviewers for their time. Most of all we are grateful to our contributors for providing their expertise

and perspectives on ‘next generation systematics’ and for their patience with the process of constructing the volume.

chapter-references

References

Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., et al. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, **387**, 489–93.

Comment [INT1]: Unlinked

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, MA, Sinauer Press.

Fleischmann, R. D., Adams, M. D., White, O., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus* Influenzae Rd. *Science*, **269**, 496–512.

Fonseca V. G., Carvalho, G. R., Sung, W. et al. (2010). Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications* **1**, 98.

Hillis, D. M., Moritz, C. and Mable, B. K. (1996). *Molecular Systematics*. Sunderland, MA, Sinauer Press.

Comment [INT2]: Unlinked

Maxam, A. M., and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 560–64.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–67.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology* **26**, 1135–45.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., et al. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–79.

Thompson, J. F., and Milos, P. M. (2011). The properties and applications of single-molecule DNA sequencing. *Genome Biology*, **12**, 217.