7

Perspective: Biodiversity and the (data) beast

Holly M. Bik and W. Kelley Thomas

7.1 Introduction

The challenges faced in the analysis of high-throughput sequencing data are discussed so frequently that the issues have become palpable stereotypes. Phrases such as 'data deluge', 'hockey stick graph' and 'bioinformatics bottleneck' are ubiquitous to the point of spawning an internet bingo card of overused sound bytes for audiences to check off during seminars (http://bit.ly/wYNxrF). Yet for all the discussion of these challenges, the dialogue about potential solutions is ignored or wildly speculative. In the sequencing world, the game is changing and no one knows how to make the next play. Computational pipelines progress slowly compared to the pace of sequencing technology, with each new platform requiring updated iterations of code and new empirical tests of error rates and data formats.

In spite of the myriad challenges left to surmount, high-throughput sequencing has already transformed and accelerated the pace of biodiversity research. Our current bioinformatic capabilities have been hard-won: characterizing and grappling with fundamentally different sequencing chemistries and order-of-magnitude-increases in file size have required substantial initial investments. The infancy of high-throughput fields means that the current biological insights are rudimentary compared to the sophisticated, complex analyses that will become available over the next decade. Yet by simply investigating ecosystems from a new perspective (genome-scale and community-level exploration, versus the narrower genetic and

Next Generation Systematics, eds P. D. Olson, J. Hughes and J. A. Cotton. Published by Cambridge University Press.

taxonomic questions previously necessitated by lower throughput Sanger sequencing), we have instantly gained a transformative view of biodiversity and ecological processes. These fledgling insights are already unprecedented, and the steadily increasing breadth of computational tools continues to widen our capacity for integrative data analysis.

7.2 The birth and death of sequencing technologies

Researchers impact sequencing technology almost as much as sequencing technology drives research. The platform currently in vogue may quickly fall out of fashion when a better (and cheaper) option hits the market. Biomedical applications drive the market and design for sequencers, with many large-scale sequencing centres focusing their resources on clinical applications (BGI@UCDavis, the Broad Institute), or species of agricultural or economic importance (BGI's facilities in China). Although many 'megasequencing' projects focused on biodiversity are now underway (Table 7.1), more fundamental and blue-skies research questions are inherently at the mercy of the technology and protocols favoured across biomedical fields. The dominance of BGI and the falling cost of sequencing are also prompting a reshuffling of long-term visions for many core facilities. For example, the US Joint Genome Institute (JGI), a behemoth in terms of sequencing power, is slowly shifting its resources towards bioinformatics capabilities and analytical support for users (JGI 2011); large-scale sequencing facilities are now much more ubiquitous (and from JGI's perspective, less practical to maintain).

Another factor impacting biodiversity research is the perpetual gap between leading-edge computational research and basic fields of biology. While the genomics world can readily keep abreast of the latest developments, more traditional biologists/ecologists often rely on secondhand (or published) information that does not reflect the most current and forward-thinking opinions of computational biologists and core sequencing facilities. Many classically trained biologists continue to focus on 454 technology as the platform of choice, but this sentiment is not echoed in the genomics world; JGI has already phased out these machines in favour of higher-throughput and more in demand Illumina HiSeq platforms; Illumina's desktop MiSeq now represents the quicker, lower throughput option for small-scale questions or preliminary data generation. Although 454 error rates can be lower than traditional Sanger sequencing (Huse et al. 2007), Roche's platform has much higher cost (per base) for sequencing, and the specific nature of 454 sequencing error (inaccurate base calls for homopolymer runs) can require computationally expensive workarounds such as Denoising (Quince et al. 2011). At the time of press, Illumina stands as the golden child of the sequencing world.

Table 7.1 A list of ongoing and planned 'megasequencing' projects that are leveraging high-throughput sequencing for biodiversity research

Initiative	Ecosystem	Data collection	Reference
Human Microbiome Project	Human body	Metagenomes, metatranscriptomes, rRNA, reference genomes	http://www.hmpdacc.org/
Earth Microbiome Project	Terrestrial, aquatic, marine	Metagenomes, metatranscriptomes, rRNA	http://www.earthmicrobiome.org/
Your Wild Life	Insects, human body, buildings	Morphology, rRNA	http://www.yourwildlife.org
Home Microbiome Project	Residential buildings, human body	rRNA	http://homemicrobiome.com/
TARA Oceans	Marine	Metagenomes, metatranscriptomes, rRNA, reference genomes, morphology	http://oceans.taraexpeditions.org/
NEON	Terrestrial soil	Metagenomes, metatranscriptomes, rRNA	http://www.neoninc.org/
TerraGenome	Terrestrial soil	Metagenomes, metatranscriptomes, rRNA	http://www.terragenome.org/
Moorea BioCode	Terrestrial, marine	Morphology, rRNA	http://www.mooreabiocode.org/
Micro B3 Project	Marine	Metagenomes, metatranscriptomes, rRNA, metaproteomics (planned)	http://www.microb3.eu/
Genomic Observatories Network	Terrestrial, aquatic, marine	Metagenomes, metatranscriptomes, rRNA, reference genomes (planned)	http://www.genomicobservatories.org/
Microbiology of the Built Environment (microBEnet)	Buildings, indoor environments	rRNA, metagenomes, reference genomes	http://www.microbe.net/

But Illumina is neither the epitome nor peak of sequencing technology (see Table 7.2 for an overview of high-throughput platforms). Short reads still pose a very real problem for computational data processing, and longer reads are fundamentally necessary for many applications and biological questions. Although single-molecule sequencers are prone to quality issues (~ 90% base calling accuracy, versus 99.5% for Illumina), long read platforms such as Pacific Biosystems are now seen as a natural complement for short read datasets. This is particularly true for *de novo* and/or metagenomic sequencing, where hybrid approaches can significantly improve the subsequent assemblies; long read technologies (including 454's new 1000 bp XL+ chemistry) can be harnessed to produce scaffolds, which are then combined with the deep coverage obtained from short read Illumina datasets. Such hybrid sequencing is increasingly supported by genome assembly pipelines (e.g. ALLPATHS-LG; Gnerre et al. 2010). Supplementary techniques such as optical mapping (Freeland et al. 2011) represent other increasingly common methods for improving the accuracy of contig assembly. However, hybrid sequencing approaches have not yet become widely popular (perhaps in part due to community hesitation regarding single-molecule error rates), and many genome sequencing projects continue to rely on short read data alone. Short read assemblies can be problematic, particularly when no closely related, high-quality reference genome index exists for the species being sequenced. The increasing prevalence of such unfinished 'draft genomes' has been cited as a critical threat impacting comparative genomic studies and our understanding of genome evolution (Alkan et al. 2011).

The playing field continues to shift. Whispers and press releases routinely circulate about newer and greater technologies. The most recent announcement from Oxford Nanopore has generated one such frenzied buzz: its single-molecule GridION platform has promised extremely long read lengths (sequence fragments in the thousands of kilobases), real-time data return at the rate of 1.4 Gigabases per hour and the ability to continuously generate sequence data for several days on end. In addition, the company announced the MinION, a single-use, affordable version of the larger nanopore machine (expected to retail for under USD $900) that is effectively the size of a USB stick. If this technology lives up to the hype, and the resulting data are not as error prone as other single-molecule machines, the dream of a field-ready sequencing technology could finally become reality (see this volume, Chapter 6). Researchers of any discipline could instantly plug the MinION into a laptop and generate data in the forest or on the beach. With Oxford Nanopore's announcement fresh in their minds, computational biologists are now looking to design more efficient algorithms for processing data – pipelines which can be quickly executed on an isolated computer, without needing to connect to a remote laboratory server or cloud-based service.

Table 7.2 Comparison of high-throughput sequencing platforms

Platform	Manufacturer	Read Length	Data output	Run time	Primary error type	Final error rate
HiSeq 2000	Illumina	2 × 100 bp	600 Gb	11 days	Substitution	~ 0.1%
HiSeq 2500[a]	Illumina	2 × 150 bp	120 Gb	27 hrs	Substitution	~ 0.1%
MiSeq[a]	Illumina	2 × 250 bp	6–7 Gb	35 hrs	Substitution	~ 0.1%
GS-FLX Titanium XL+	Roche	1000 bp	0.7 Gb	23 hrs	Indel	1%
GS-Junior	Roche	400 bp	0.03 Gb	10 hrs	Indel	1%
PacBio RS	Pacific Biosystems	860–1100 bp	0.01 Gb	0.5–2 hrs	CG Deletions	≤ 15%
Ion Proton – Proton II Chip[b]	Ion Torrent	200 bp	10 Gb	2 hrs	Indel	1%
Ion PGM – 318 Chip[b]	Ion Torrent	2 × 200 bp	1 Gb	1.5 hrs	Indel	1%
GridION[c]	Oxford Nanopore	100 000 bp	N/A	Flexible	Deletions	4%
MinION[c]	Oxford Nanopore	100 000 bp	N/A	Flexible	Deletions	4%
HeliScope[d]	Helicos	35 bp	28 Gb	28 hrs	Deletions	≤ 0.1%
SOLiD 550xl	ABI Biosystems	75 + 35 bp	155 Gb	8 days	AT Bias	≤ 0.1%

Sequencer output listed in Gigabases.
Platform-specific data obtained from Glenn (2011) and manufacturer websites.
[a] Illumina's projected performance estimates for planned instrument upgrades released in the third quarter of 2012.
[b] Ion Torrent specifications reflect planned upgrades released in the second quarter of 2012.
[c] Oxford Nanopore error rates, run time and read lengths as reported in *Nature* News (17 Feb 2012; doi:10.1038/nature.2012.10051); platforms were not yet commercially available at the time of press, and thus data reflect projected specifications based on press announcements.
[d] Helicos no longer sells reagents or instruments. HeliScope error rate as reported by the manufacturer, representing consensus error rate at 20× coverage.
N.B. Due to rupid advances in DNA sequencing, some of the above listed platforms will represent legacy technologies at the time of publication.

Other, alternative sequencing approaches are envisioned on a regular basis. Magnetic sequencing, another single-molecule approach, utilizes DNA hairpins anchored between a glass slide and a magnetic bead: DNA can be unzipped and manipulated simply by applying a magnetic field (Ding et al. 2012). Although such technologies are still in proof-of-concept stages, they present exciting new possibilities for the sequencing world and prospective solutions for technological limitations in existing platforms related to optics (sophisticated lenses needed to record fluorescent base incorporations) and phasing (sequencing errors stemming from sequence extension failure) (Linnarsson 2012). As the market becomes increasingly crowded with distinct and competing sequencing technologies, the phrase 'next generation sequencing' becomes rapidly outdated terminology. A semantic shift towards the label 'high-throughput sequencing' thus seems more appropriate as technology evolves beyond third (fourth, fifth. . .) generation sequencing platforms.

7.3 Next generation barcoding

Since its advent in 2003, DNA barcoding has established itself as a booming, well-funded sector. However, the utility and future prospects of these approaches are rapidly changing (Taylor and Harris 2012) as the scientific community embraces high-throughput sequencing technologies and simultaneously expands pipelines for data analysis. Traditional DNA barcodes still have a place – they can provide a quick solution or diagnostic assay for specific questions; confirming a taxonomic ID for a few specimens is still most easily carried out via PCR and Sanger sequencing, especially for laboratories that need these data quickly. However, these low-throughput approaches are increasingly relegated to well-studied or larger species that are easily sampled, for example within Barcode of Life consortia focusing on fish (FishBOL; http://www.fishbol.org/) or insects (Lepidoptera, bees, mosquitoes; http://www.barcodeoflife.org/content/community/projects). Larger scale studies, by necessity, must instead capitalize on the low per-base cost and multiplexing capacities enabled by high-throughput platforms. For poorly studied groups such as microbes, meiofauna or viruses, high-throughput sequencing represents the *only* methodology which can rapidly illuminate hidden patterns and novel diversity – traditional culturing methods or taxonomic identifications are inherently time-consuming and provide a much more limited view.

A recent study by Hajibabaei et al. (2011) laid the foundation for a transition to high-throughput DNA barcoding of known diagnostic loci (the mitochondrial *cox1* gene). But such approaches face another looming shift towards ecological genomics ('ecogenomics'), a fundamental expansion of locus-specific barcoding

approaches, where deep characterization of genomic-scale natural variation will be achieved through comparative sequencing (Tautz et al. 2010).

Draft genomes are quickly becoming the new barcode – for ecologically important species (abundant taxa, environmental indicator species, novel lineages), generating a draft genome is becoming almost as easy as the work previously required to sequence a single gene from a single specimen. Illumina sequencing is now deep enough to generate a draft genome within one lane of a HiSeq 2500 flow cell (45 Gb of 250 bp paired-end reads), or even multiplex and sequence several smaller genomes (e.g. bacteria, with typical genome sizes < 10 Mb). While 'finishing' a genome continues to represent a major expense (costs of additional sequencing and manpower needed for accurate assembly and annotation), draft genomes (gene-size contigs) can instead be generated for minimal costs (< $2000 USD and dropping). Comprehensive reference databases are critical for robustly assigning taxonomy to unknown sequences, and genome contigs from a known organism provide a robust reference for the sequence-based comparisons that are central for studies of environmental biodiversity. In addition, draft genomes will provide a long-term repository of information applicable to other biological questions; making these data freely and easily accessible (providing application program interface (API) access or a simple workflow for downloading raw data and assembled contigs) additionally provides a public service benefiting the wider research community. The genome-as-a-barcode model must also serve to supplant the continued overselling of every new published genome sequence – dubbed 'Yet Another Genome Syndrome' by science writer Carl Zimmer (http://bit.ly/IbKFCW). One genome will never contain all the answers, and we must move towards sophisticated analyses spanning the breadth of taxonomic diversity, taking into account within-species genomic variation and plasticity, in order to obtain the most powerful and transformative insights about biodiversity and evolution.

Re-envisioned barcoding approaches will also harness the power of high-throughput sequencing for population genomics – a new frontier for characterizing environmental genetic diversity. Biodiversity research has historically maintained a strong phylogenetic viewpoint, with this tree-based perspective also dominating the move towards high-throughput environmental sequencing. However, deep sequencing can provide the explicit potential to search for signatures of natural selection (Yang and Bielawski 2000; Kryazhimskiy and Plotkin 2008; Suzuki 2010) as a way of understanding shifting constraints during environmental change and local adaptation. The study of population-level haplotypes has traditionally been carried out using PCR-based approaches and microarrays for a small sample size of specimens (Freeland et al. 2011), often a prohibitively laborious task for microscopic taxa. Because 454 and Illumina-based technologies inherently sequence individual strands of DNA, high-throughput amplicon studies of single genes (e.g.

mtDNA, rRNA) not only recover species-specific barcodes, but can quantify the entire range of population-level variation across space and time.

Soon, our capacity for population genomics will play a critical role in identifying and tracking the biological repercussions of environmental disturbance. In addition to decreasing species diversity, environmental stressors can reduce genetic diversity within populations of a species by causing a strong reduction in population size and thereby increasing the importance of genetic drift (Selkoe et al. 2010). Intraspecific genetic diversity provides critical ecosystem resilience, enabling species to persist in drastically altered habitats (Gienapp et al. 2008); species with high genetic diversity may be more persistent than species with low genetic diversity. Thus, elucidating changes in species' population genetic structure is a key step in assessing initial impacts, likely resilience and recovery of an environment in response to disturbance.

The rapid and inexpensive nature of sequencing-based methods (versus specialist labor required for morphology-based taxonomic surveys) will also likely translate into applied approaches for habitat management and monitoring. Using deep sequencing technologies, it is now possible to utilize well-characterized population-level markers (e.g. mitochondrial genes) and metagenomics to thoroughly investigate gene flow, genetic drift and signatures of selection in microbial species – taxonomic groups which are historically understudied yet play critical roles in ecosystem function (Snelgrove et al. 1997). The most informative approaches will characterize populations across multiple scales and dimensions, including time (short-term versus annual fluctuations) and space (e.g. changes in genetic structure across pollution or salinity gradients).

7.4 Moving beyond pie charts

Although computational bottlenecks are still significant, it is becoming routine (even *easy*) to produce a suite of sophisticated outputs from high-throughput sequence datasets. The development of well-documented, complete software toolkits such as QIIME (http://qiime.org; Caporaso et al. 2010) now enables budding genomicists to manipulate their own personal high-throughput datasets with minimal bioinformatic training. Extensive video and web tutorials, and the packaging of QIIME into Amazon AWS Cloud servers, removes the need for a computer science degree or dedicated local hardware to process and analyse millions of raw sequence reads.

However, the easiest outputs to obtain often confer the most limited biological insights. Taxonomic summaries, rarefaction curves and community-level cladograms provide useful but unsurprising pictures (most environments are diverse, contain lots of taxa and habitat factors drive community assemblages),

often producing more questions than answers. Sifting through data evokes questions such as 'now what?' and 'tell me more'. In addition, packaged software and graphical interfaces can perpetuate a 'black box' mentality since users are confronted with an additional degree of separation from the underlying computational mechanisms. Because high-throughput approaches are still in their infancy, researchers applying computational pipelines to new ecosystems or poorly described taxa (microscopic eukaryotes, viruses) may fail to appreciate how the nuances of data analysis can affect the biological conclusions which can be drawn from pipeline outputs. For example, the presence of intragenomic rRNA variants in eukaryotic genomes severely complicates the interpretation of operational taxonomic units (OTUs) clustered under arbitrary sequence identity cutoffs (e.g. 97%). Singleton or low-copy OTUs containing few sequence reads may represent sequencing errors, chimeric hybrid sequences formed during PCR, minor variant rRNA copies that persist within a genome and fall under the radar of selection, or dominant rRNA gene copies from biological species representing the 'rare biosphere' (Sogin et al. 2006; Huse et al. 2010). In some ways, the shift to Illumina sequencing may help to alleviate some of these problems associated with rRNA genes: OTU signatures (and abundances) of rare taxa may be increased to a degree such that they are readily distinguished from sequencing error OTUs. Evidence from microbial communities suggests that more (but shorter) reads can bring the rare biosphere out of the twilight zone (where error OTUs and rare taxa are indistinguishable, e.g. in lower throughput 454 datasets) and up to detectable abundances that are not discarded during data filtering steps (Caporaso et al. 2011). The move to explicitly phylogenetic approaches – interpreting environmental sequence data within a tree topology – may provide additional relief, particularly if it becomes possible to use head–tail patterns (Porazinska et al. 2010) to link minor rRNA variants with dominant rRNA OTUs (reference sequences) for any given species.

An open question – regardless of locus – is the effect that computational pipelines can have on the overarching biological conclusions researchers may draw from a dataset. Currently, we do not fully understand the differential effects of OTU picking algorithms (differences in calculating pairwise sequence identity, how OTU membership varies across approaches) and the resulting effects for processing raw sequence data. The use of multiple pipelines is a common substitute, where researchers effectively pummel datasets with a wide range of algorithms, parameters and data filtering (chimaera checks, statistical filtering based on read counts per OTU) to assess the overall robustness of biological patterns.

Ribosomal data will always be useful – purely for the historical knowledge base, the many reference sequences in public databases and the ease with which rRNA can be amplified via conserved PCR primers. However, the questions we can ask about biodiversity and evolution are expanding as high-throughput fields evolve.

Ribosomal loci alone cannot necessarily answer the more pressing questions we are asking about ecosystem function. What taxa are active versus inactive (differential gene expression)? What species are resident versus transient? The most widely used ribosomal genes (16S/18S) are too conserved for exploring population-level differences; closely related biological species may only differ by 1 bp over a diagnostic rRNA gene sequence (e.g. 18S in nematodes; Porazinska et al. 2010), and thus differentiating closely related sister species may be impossible with short read datasets. In addition, genome plasticity in bacteria means that the same rRNA sequence can be associated with drastically different genome architectures, and thus, a single reference sequence may span multiple, distinct ecological niches. For example, *E. coli* and *Shigella* strains possess identical 16S sequences, yet share a core genome comprising only 6% of common gene families across strains (Lukjancenko et al. 2010); bacterial species thus exhibit a pan-genome continuum rather than having distinct species boundaries. Similar analysis has not been carried out for eukaryotic species, although the specific life history and reproductive modes of nucleated species are unlikely to yield the dramatic plasticity observed in bacterial and archaeal species. Nonetheless, the prevailing focus on ribosomal genes for eukaryotic species precludes population-level inferences and the associated analysis of their potential ecological relevance.

As we shift towards pure metagenomic approaches, there are many more challenges. Environmental studies must go beyond the one gene model (particularly for eukaryotes, the current focus remains on rRNA marker genes), and steadily aim to reduce biases that cloud data interpretation (minimizing or avoiding PCR, investing significant time and effort into devising robust sampling strategies, and developing increasingly sophisticated bioinformatics tools), all while moving towards an explicitly phylogenetic framework for data analysis. One lingering barrier is the paucity of informative data for comparative analyses; for example, while it is possible to generate shotgun metagenome datasets for microbial eukaryote communities (meiofauna, protists, fungi, etc.), the overall lack of molecular data for these groups precludes gene annotation, taxonomic assignments and a coherent understanding of ecological function for species in these assemblages. Routinely generating disparate data types as part of high-throughput studies – even if such data confer limited insights in the beginning – is key to removing the 'crutch' of rRNA loci. A combination of metagenome and metatranscriptome data (supplementary to rRNA) provides a built-in method for annotating genes in a novel environmental sample: expressed mRNA transcripts can be used to validate predicted gene contigs from shotgun genomic data (Moran 2009). Data from single-cell sequencing can also be used to link environmental gene contigs to a taxonomic ID, even for sequences with 'hypothetical protein' annotations (Thrash et al. 2014). Although precise functional characterization of such genes may remain difficult, identification and annotation can

itself provide a solid foundation for sequence-based comparative analyses across space, time and habitats.

As we build environmental gene catalogues, molecular data must also be linked to species names. Interpreting genomic data in the context of biology and morphology will be required to construct a complete understanding of ecosystem function. Microfluidic technologies offer substantial promise towards this goal, as these approaches can be applied as a practical way to reduce the complexity of metagenomic samples. Furthermore, single-cell approaches can elucidate metabolic pathways, ecological roles and organismal interactions for uncultured species (Swan et al. 2011; Martinez-Garcia et al. 2012a; 2012b). For example, a recent study of marine picobiliphytes (a group of protists previously assumed to be photosynthetic) revealed no evidence for light-harvesting plastid genes and additionally provided insight into single-cell interactions with bacteria and viruses (Yoon et al. 2011). A number of centres now offer research services for single-cell sorting through to genome sequencing, such as the Single Cell Genomics Center at the Bigelow Laboratory for Ocean Sciences and a newly launched centre at the Broad Institute (the latter in partnership with Fluidigm). The less complex a metagenomic sample, the more powerful the bioinformatic approaches to separate and bin taxa according to sequence properties. Metrics such as tetranucleotide frequencies, GC content and read abundance information can be analysed in combination to build emergent self-organizing maps (ESOMs) – a visual display of 'genome signatures' generated from assembled contigs, constructed in the absence of reference genomes or public sequence databases (Dick et al. 2009).

For study design and data products, the NEON (National Ecological Observatory Network) project (http://www.neoninc.org) is poised to set a shining example for future ecological studies. NEON represents a long-term initiative to characterize terrestrial ecosystems across ecological observatories in the US (a network of established NEON observatories and Long-Term Ecological Research Network sites). Funding from the National Science Foundation is enabling the construction of 62 strategic sample sites over the next 5 years, and will allow for consistent collections of instrument and field data over decadal timescales. Overarching project goals will characterize the ecological repercussions of climate change and human impacts on terrestrial habitats (land use changes, invasive species), using a suite of biological and physical measurements collected over the next 30 years. Sample collection has been designed to provide the deepest ecological insight possible, with network sites being contextualized across many spatial scales (within-site to continental-scale questions). NEON has emphasized the collection of genomic data as the most informative approach for deeply characterizing microbial diversity and function in terrestrial habitats. Parallel high-throughput sequencing approaches will be carried out to assess community structure (16S rRNA assays), as well as both potential

(metagenomics) and realized (metatranscriptomics) community function. The incorporation of established long-term ecological research (LTER) sites will further enable the project to leverage decades of historical metadata (such as temperature and soil pH), and incorporate this information into bioinformatic data analysis and predictive ecological models. Most importantly, there will be no time lag on data release: all NEON data will be immediately uploaded onto a public web portal. This commitment to open access will not only provide unprecedented data accessibility and efficiency for the research community, but will also ultimately benefit educators, policymakers and the general public. Projects like NEON are setting new standards for high-throughput biodiversity research; the carefully planned, integrative nature of data collection, coupled with a strong emphasis on open access, represents a fundamental shift in the design and execution of such large-scale research projects.

As environmental sequencing approaches expand and become increasingly elegant, there are still many significant hurdles to overcome. How do software workflows differ (e.g. methods for OTU picking), and how do these differences affect the biological interpretation of sequence data? How do we ensure accurate taxonomy assignments? For example, when mining orthologous marker genes, how might gene trees be reconciled and combined to accurately depict environmental species assemblages? How do we visualize and interpret increasingly complex data structures such as evolutionary models and measures of statistical confidence (Fox and Hendler 2011)? Although such questions continue to linger, the high-throughput community is now aiming to define and target some of these critical unknowns as high priorities for research.

7.5 The critical role of reference genomes

For investigators looking to intensely study an unknown ecosystem, reference sequences represent the limiting factor for accurately quantifying species assemblages and firmly linking sequences with taxonomy (and thus, inference of ecological function). Reference databases are repeatedly cited as the critical bottleneck preventing meaningful interpretation of environmental datasets, and barriers that can effectively reduce the accuracy of otherwise robust computational tools.

For eukaryotes, patchy database coverage is an important consideration. First, it severely limits the capacity for accurate taxonomic assignments from high-throughput sequencing datasets, even for historically well-represented sequences such as rRNA (as of release 108, the SILVA SSU rRNA database contained only 62 587 eukaryote sequences, compared to 530 197 for bacteria and 25 658 for the much lower diversity archaea). Given the current sparsity of taxonomic sampling

across the Tree of Life, BLAST-based approaches routinely return a significant proportion of sequences as 'no match' (e.g. no close relatives in the database exhibiting pairwise sequence identities > 90%) or 'unclassified environmental sequences' (uninformative taxonomic assignments resulting from poor annotations in GenBank). In shotgun metagenomic datasets, genes from uncultured environmental lineages are also unlikely to show high homology with proteins known from model organisms, and novel proteins may show no resemblance at all to anything present in the database. This scenario is particularly true for habitats with diverse, divergent fauna such as the deep sea, where a scarcity of reference sequences severely hinders the application of BLAST-assigned taxonomy (Bik et al. 2011). Model-based approaches such as the RDP naïve Bayesian classifier (Wang et al. 2007) rely on reference collections as training sets, where better optimization (better reference sets) will improve the accuracy of taxonomic assignments for unknown environmental sequences (Werner et al. 2011). In tree-based methods such as pplacer (Matsen et al. 2010) or the Evolutionary Placement Algorithm (Berger and Stamatakis 2011), unknown sequences are placed onto guide tree topologies via posterior probabilities and likelihood scores. Sparse taxon sampling within reference trees can increase uncertainty and lower confidence values (e.g. posterior probabilities in pplacer) for the phylogenetic placement of short environmental sequence reads. High-throughput studies must work towards identifying taxa at species level or below, since species and populations (e.g. bacterial strains) can be functionally distinct and fill disparate ecological roles. However, current bioinformatic pipelines such as MEGAN (Huson et al. 2011) continue to deliver BLAST-based summaries of species assemblages; the inherent reliance on reference databases means that ecologically distinct units are lumped together and interpreted as a single entity (e.g. sequence abundances reported at the family, order or phylum level). Taxonomic assignments at family-level or higher, while somewhat informative, are not the ultimate goal. In complex biological systems, understanding the intricacies of species interactions (particularly between closely related sister taxa) is critical for expanding our knowledge of, and gaining the capacity to predict, biodiversity.

Secondly, existing database resources have sparse functional annotations for non-model organisms. Such database annotations are key, since the vast majority of organisms in any given habitat have no published genome sequence and functional inferences inherently rely on identifying homology with known proteins and protein families. In metagenomic and trascriptomic datasets, sample comparisons may reveal significant shifts in gene expression (e.g. across space, time, or following environmental disturbance), but many gene functions may remain uncharacterized because of uninformative 'hypothetical protein' annotations. While an expansion of sequencing efforts can help to overcome sparse taxon sampling in public databases, more data will not solve the annotation problem. A substantial

investment in infrastructure and manual database curation will instead be required to provide more useful functional annotations for environmental sequencing studies.

The lack of reference genomes is a common lament across high-throughput fields, and this challenge encapsulated a central theme at the recent 13th Workshop of the Genomic Standards held at BGI in Shenzhen, China (March 5–7, 2012; Gilbert et al. 2012). Some progress is being made, albeit slowly and targeted towards specific taxonomic groups. The Genomic Encyclopedia of Bacteria and Archaea (Wu et al. 2009) is one such initiative that has significantly expanded the number of available reference genomes for non-nucleated taxa, specifically targeting phylogenetic gaps in existing reference collections. Other reference genomes are being generated as part of discrete hypothesis-driven projects, such as Tara Oceans (single-cell genomics focusing on planktonic eukaryote species http://oceans.taraexpeditions.org) and the Human Microbiome project (plans to sequence > 1000 genomes from key human-associated microbial species http://commonfund.nih.gov/hmp).

7.6 Open access, community standards and data sharing

The problems outlined above are looming, persistent challenges that cannot (and should not) be overcome by individual researchers. As we enter the high-throughput era of biology, the best long-term solutions are likely to cross many disciplinary boundaries, with such solutions inherently requiring community-wide collaboration and data sharing. Some communities have fostered data-sharing cultures from the start, such as the Bermuda Accord established at the start of the Human Genome Project in 1996 (Contreras 2011); this initiative required immediate public release of DNA sequences, assemblies and gene annotations. Other fields have remained more insular, stratified according to study organism or focal habitat, without established frameworks promoting data accessibility. Addressing long-standing, overarching questions in biodiversity will require globally coordinated efforts to connect researchers, share data, and produce community-driven database resources. Given the flood of data (and the Herculean effort it can take a lone researcher to analyse their own dataset), it does not make sense *not* to share. Data repositories must consequently emphasize easily accessible user interfaces (including public APIs for data mining and complex queries) and standard metadata submission (e.g. MIxS (minimum information about any (x) sequence) standards; Yilmaz et al. 2011).

The increasingly rapid pace of high-throughput fields means that most research developments (including genome sequences and metagenome datasets)

remain unpublished. A scientific culture that promotes open science and strong interdisciplinary research networks will be imperative for effectively disseminating new developments, and fostering transformative research. Researchers must aim to make increasingly complex connections (moving beyond species lists and isolated data types), since scientists within any given discipline cannot be expected to maintain deeply multifaceted skill sets. Intimate, interdisciplinary collaborations will be key.

Persistent 'elephants in the room' are issues related to data storage. There are immense computational costs that come with increasing data volumes, which the research community is not currently prepared to face. Sequencing outputs have not yet reached the stage where long-term storage is unfeasible, although gigabyte-sized files certainly pose challenges for manipulating and analysing sequence data (even cloud computing can necessitate mailing physical hard drives to load data onto servers (Baker 2010)). Currently, the Illumina HiSeq 2000 platform typically generates 50–60 Gb of data per flow cell run, and purchasing a 2TB external hard drive for data storage costs around $100. A two order of magnitude increase in sequencing output would require three such hard drives ($300 in storage costs each time a run generates 5 Tb of data), while an additional order of magnitude increase on top of that would incur storage costs of $2500 per sequencing run (25 hard drives to store 50 Tb of data)!

The research community must address these looming data storage issues or entirely give up storing raw data. The sheer quantity of raw data (particularly in biomedical and clinical fields) is already reducing the likelihood that such data will be reused or referenced after initial processing and analysis. Although scientists are extremely reluctant to part with raw data, exponentially escalating storage costs for large projects such as the 1000 Genomes Project are prompting researchers to retain only the processed sequence data or comparative results (Baker 2010). This might be feasible in clinical applications searching for a specific SNP or gene variant, but not within biodiversity studies asking fundamentally distinct sets of questions. In addition, for historical or archaeological samples (where DNA content is low and samples are precious), data storage is of the utmost importance if the source material is not available for resequencing.

Environmental sequencing approaches are used as a method to circumvent traditional bias, based on the underlying rationale that we lack a firm understanding of species assemblages and ecological interactions in most ecosystems. In the absence of prior knowledge, it becomes impossible to search through datasets for specific taxa or biological observations: patterns may only become apparent on an extremely large scale (e.g. globally comparative datasets to investigate patterns such as latitudinal gradients, or very deep sequencing efforts undertaken to characterize the 'rare biosphere'). In addition, the biodiversity community does not currently possess the necessary computational tools for teasing out these complex

patterns from environmental datasets. Data must be retained indefinitely because its utility has not (and cannot) be maximized with our existing cyberinfrastructure. Because of these considerations, the biodiversity community has an ingrained need to foster a dialogue on data storage and access. One potential solution may lie in a partnership with industry leaders such as Google or Amazon, who have the capacity to assess and address data storage problems for researchers. Google's recent investment in the bioinformatics startup DNAnexus suggests that such links between the research community and software infrastructure may soon come to fruition.

Future biodiversity analyses will need to harness all available (raw) datasets from disparate environments in order to build a global view of species distributions and environmental parameters that drive ecological interactions.

7.7 Moving towards ecosystem function

As we move beyond pie charts and single-gene barcodes, high-throughput sequencing approaches will soon allow us to monitor the ecological pulse that drives life within any given habitat. The development of new modelling approaches is set to kick start another wave of key advances in high-throughput fields. Recent studies have given us a tempting glimpse of the future analytical possibilities when robust sampling plans, environmental metadata and mathematical models are combined within a study. Over a 6-year time period, 72 molecular samples were collected from surface waters in the western English Channel in conjunction with physical, chemical and biological measurements. In addition to deep characterization of microbial communities using rRNA profiles and functional expression (Gilbert et al. 2008; 2010) and seasonal dynamics indicating oscillation and persistence across different microbial taxa (Gilbert et al. 2009; 2010), the design of the study has further enabled temporal and spatial predictions of microbial assemblages using an artificial neural network approach (Larsen et al. 2012).

Besides modelling, another trend is to combine data types that have historically represented disparate fields of study. Parallel sequencing of ribosomal genes (PCR amplified), expressed mRNA transcripts, and shotgun fragments of total genomic DNA is an insightful and increasingly common approach (Gilbert et al. 2010). But even more unconventional couplings can provide even more powerful insights, such as complementary genomic and chemical profiling: combined metagenomic and metaproteomic approaches have been successfully applied to detect genomic recombination (Lo et al. 2007), adaptation (Denef et al. 2009), ecological differentiation (Denef et al. 2010) and interspecific interactions (Mueller et al. 2010) in uncultured bacteria inhabiting low-diversity acid mine drainage ecosystems.

As with any high-throughput approach, incorporating new cross-discipline methodology will require collection of reference data as well as an emphasis on robust empirical tests of computational pipelines (using simulations, control communities and real environmental datasets). Environmental baselines – intense characterization of both spatial and temporal patterns – are also needed to train models and track long-term impacts stemming from climate change and anthropogenic disturbance.

7.8 New paradigms for data analysis

The future landscape for DNA-based studies is poised to be far different from what has prevailed in the past. Single-organism genome studies are giving way to 'megasequencing projects', undertaking comparative analyses on tens of thousands of organisms or species (with China's BGI ambitiously aiming to sequence 1 million human genomes). Already, the flood of available sequence data makes it possible to easily conduct more generalized, exploratory data analysis, often as a precursor for hypothesis-driven research with defined questions to assess. Asking broad questions within the context of large, poorly characterized datasets (rRNA sequences from thousands of uncharacterized species, or newly generated genome sequences from divergent, uncultured lineages) can serve as a means to identify surprising trends or novel patterns. For example, deep sequencing of microbial eukaryotes from intertidal marine sediments revealed a surprising dominance of Platyhelminthes (Fonseca et al. 2010) – a taxonomic group where traditional taxonomic approaches had historically failed to adequately characterize this diversity; or the recent sequencing of the Little Skate genome, which appears to lack an otherwise ubiquitous cluster of developmental Hox genes (King et al. 2011).

Our approach towards data is also shifting in regard to the tools researchers will use to tease out biological patterns. No longer will we be using clunky, limited software designed by biologists-turned-programmers. Increasing data volumes are now necessitating well-designed, flexible software tools that can conduct sophisticated analyses within intuitive user interfaces. We are moving towards 'workflow' type pipelines with data provenance and effective tools for visualizing complex data. Software such as VisTrails (http://www.vistrails.org/) and Paraview (http://www.paraview.org/) has emerged for conducting large-scale data analysis across a multitude of scientific disciplines. Although high-throughput fields have witnessed exponential progress in a few short years, tackling the root of existing computational bottlenecks will conversely require slow and steady progress. Software design is a formidable task, particularly given the need for close interaction between product engineers (computer scientists who often have little knowledge of biology) and researchers (who have many ideas for software products but

do not understand the time and lengthy development process that is required to produce lasting, effective user interfaces).

7.9 Summary

Threats to global biodiversity are increasingly urgent, yet we are witnessing a parallel, exponential growth in our capacity to characterize and track biological patterns. The explosion of high-throughput sequencing technology and the evolution of computational pipelines will continue to set new standards for biodiversity research in the coming decades, with transformative consequences for myriad scientific disciplines.

References

Alkan, C., Sajjadian, S. and Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods*, **8**, 61–5.

Baker, M. (2010). Next-generation sequencing, adjusting to data overload. *Nature Methods*, **7**, 495–9.

Berger, S. A. and Stamatakis, A. (2011). Aligning short reads to reference alignments and trees. *Bioinformatics*, **27**, 2068–75.

Bik, H. M., Sung, W., De Ley, P., et al. (2011). Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Molecular Ecology*, doi, 10.1111/j.1365–1294X.2011.05297.x.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–6.

Caporaso, J. G., Lauber, C. L., Costello, E. K., et al. (2011). Moving pictures of the human microbiome. *Genome Biology*, **12**, R50.

Contreras, J. L. (2011). Bermuda's legacy, policy, patents and the design of the genome commons. *Minnesota Journal of Law, Science and Technology*, **12**, 61.

Denef, V. J., Kalnejais, L. H., Mueller, R. S., et al. (2010). Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 2383–90.

Denef, V. J., VerBerkmoes, N. C., Shah, M. B., et al. (2009). Proteomics-inferred genome typing (PIGT) demonstrates inter-population recombination as a strategy for environmental adaptation. *Environmental Microbiology*, **11**, 313–25.

Dick, G. J., Andersson, A. F., Baker, B. J., et al. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biology*, **10**, R85.

Ding, F., Manosas, M., Spiering, M. M., et al. (2012). Single-molecule mechanical identification and sequencing. *Nature Methods*, **9**, 367–72.

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

Fox, P. and Hendler, J. (2011). Changing the equation on scientific data visualization. *Science*, **331**, 705–8.

Freeland, J. R., Petersen, S. D. and H. Kirk, eds. (2011). *Molecular Ecology*, 2nd edn. Chichester, Wiley Blackwell.

Gienapp, P., Teplitsky, C., Alho, J. S., Mills, J. A. and Merila, J. (2008). Climate change and evolution, dientangling environmental and genetic responses. *Molecular Ecology*, **17**, 167–78.

Gilbert, J., Bao, Y., Wang, H., et al. (2012). Report of the 13th Genomic Standards Consortium Meeting, Shenzhen, China, March 4–7, 2012. *Standards in Genomic Sciences*, **6**, doi, 10.4056/sigs.2876184.

Gilbert, J. A., Field, D., Huang, Y., et al. (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One*, **3**, e3042.

Gilbert, J. A., Field, D., Swift, P., et al. (2009). The seasonal structure of microbial communities in the Western English Channel. *Environmental Microbiology*, **11**, 3132–9.

Gilbert, J. A., Field, D., Swift, P., et al. (2010). The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. *PLoS One*, **5**, e15545.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–69.

Gnerre, S., MacCallum, I., Przybylski, D., et al. (2010). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 1513–18.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer G. A. C. and Baird, D. (2011). Environmental Barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One*, **6**, e17497.

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. and Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, **8**, R143.

Huse, S. M., Welch, D. M., Morrison, H. G. and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology*, **12**, 1889–98.

Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N. and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, **21**, 1552–60.

JGI (2011). Joint Genome Insitutue (JGI): A 10-Year Strategic Vision. Walnut Creek, CA, US Department of Energy.

King, B. L., Gillis, J. A., Carlisle, H. R. and Dahn, R. D. (2011). A natural deletion of the HoxC cluster in elasmobranch fishes. *Science*, **334**, 1517.

Kryazhimskiy, S. and Plotkin, J. B. (2008). The population genetics of dN/dS. *PLoS Genetics*, **4**, e1000304.

Larsen, P. E., Field, D. and Gilbert, J. A. (2012). Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods*, doi, 10.1038/nmeth.1975.

Linnarsson, S. (2012). Magnetic Sequencing. *Nature Methods*, **9**(4), 339–40.

Lo, I., Denef, V. J., VerBerkmoes, N. C., et al. (2007). Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature*, **446**, 537–41.

Lukjancenko, O., Wassenaar, T. M. and Ussery, D. W. (2010). Comparison of 61 sequenced *Escherichia coli* genomes. *Microbial Ecology*, **60**, 708–20.

Martinez-Garcia, M., Brazel, D., Poulton, N. J., et al. (2012a). Unveiling *in situ* interactions between marine protists and bacteria through single cell sequencing. *The ISME Journal*, **6**, 703–7.

Martinez-Garcia, M., Swan, B. K., Poulton, N. J., et al. (2012b). High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *The ISME Journal*, **6**, 113–23.

Matsen, F. A., Kodnere, R. B. and Armbrust, E. V. (2010). pplacer, linear time maximum-likelihood Bayesian phyogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, **11**, 538.

Moran, M. A. (2009). Metatranscriptomics: eavesdropping on complex microbial communities. *Microbe*, **4**, 329–35.

Mueller, R. S., Denef, V. J., Kalnejais, L. H., et al. (2010). Ecological distribution and population physiology defined by proteomics in a natural microbial community. *Molecular Systems Biology*, **6**, 347.

Porazinska, D. L., Giblin-Davis, R. M. Sung, W. and Thomas, W. K. (2010). Linking operational clustered taxonomic units (OCTUs) from parallel ultra sequencing (PUS) to nematode species. *Zootaxa*, **2427**, 55–63.

Quince, C., Lanzen, A., Davenport, R. J. and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.

Selkoe, K. A., Watson, J. R., White, C., et al. (2010). Taking the chaos out of genetic patchiness: seascape genetics reveals ecological and oceanographic drivers of genetic patterns in three temperate reef species. *Molecular Ecology*, **19**, 3708–26.

Snelgrove, P. V. R., Blackburn, T. H., Hutchings, P., et al. (1997). The importance of marine sediment biodiversity in ecosystem processes. *Ambio*, **26**, 578–83.

Sogin, M. L., Morrison, H. G., Huber J. A., et al. (2006). Microbial diversity in the deep sea and the unexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 12115–20.

Suzuki, Y. (2010). Statistical methods for detecting natural selection from genomic data. *Genes and Genetic Systems*, **85**, 359–76.

Swan, B. K., Martinez-Garcia, M., Preston, C. M., et al. (2011). Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science*, **333**, 1296–300.

Tautz, D., Ellegren, H. and Weigel, D. (2010). Next generation molecular ecology. *Molecular Ecology*, **19**(Suppl 1), 1–3.

Taylor, H. R. and Harris, W. E. (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*, **12**, 377–88.

Thrash, J. C., Temperton, B., Swan, et al. (2014). Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. *The ISME journal*, doi, 10.1038/ismej.2013.243.

Wang, Q., Garrity, G. M., Tiedje, J. M. and Cole, J. R. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–7.

Werner, J. J., Koren, O., Hugenholtz, P., et al. (2011). Impact of training sets on classification of high-throughput bacerial 16S rRNA gene surveys. *The ISME Journal*, doi, 10.1038/ismej.2011.1082.

Wu, D., Hugenholtz, P., Mavromatis, K., et al. (2009). A phylogeny-driven genomic encyclopedia of Bacteria and Archaea. *Nature*, **462**, 1056–9.

Yang, Z. and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, **15**, 496–503.

Yilmaz, P., Kottmann, R., Field, D., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, **29**, 415–20.

Yoon, H. S., Price, D. C., Stepanauskas, R., et al. (2011). Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science*, **332**, 714–17.