Fantastic Beasts and How To Sequence Them: Ecological Genomics for Obscure Model Organisms

Mikhail V. Matz

The application of genomic approaches to ‘obscure model organisms’ (OMOs), meaning species with no prior genomic resources, enables increasingly sophisticated studies of the genomic basis of evolution, acclimatization, and adaptation in real ecological contexts. I consider here ecological questions that can be addressed using OMOs, and indicate optimal sequencing and data-handling solutions for each case. With this I hope to promote the diversity of OMO-based projects that would capitalize on the peculiarities of the natural history of OMOs and could feasibly be completed within the scope of a single PhD thesis.

Why Sequence Strange Creatures?

As sequencing methods continue to diversify and their costs continue to drop [1], full-scale genomic analysis is becoming feasible for an increasingly broad range of study systems and biological questions, far beyond what was accessible using well-established genomic models. This review considers OMOs – organisms with no pre-existing genomic resources. Several recent reviews [2–4] have described the range of methods applicable to ‘non-model organisms’ in general. I identify here optimal approaches that would allow valuable insights to be gained within the scope of a single PhD project. This requirement for short-term return justifies the use of term ‘OMO’ instead of ‘non-model organism’: while the latter is basically a model organism in the making, with the hope of eventually initiating many diverse projects, the former might be of interest for only a single project capitalizing on one specific aspect of the natural history of the OMO. For example, Death Valley pupfishes were studied because they have the smallest species range on earth [5], deep-sea mussels were sequenced to understand adaptation of animals to chemosynthetic environments [6], and in the saker falcon from the Qinghai-Tibetian Plateau both DNA polymorphisms and gene expression were examined to investigate adaptations of a predatory bird to high altitude [7].

Generally, ecological reasons to sequence OMOs include population biology, genomic targets of selection and introgression during adaptation to diverse environments, gene regulation underlying acclimatization and adaptation, and ecological role of epigenetics (Table 1). There is one more reason – phylogenomics – for which I refer the reader to recent reviews [8,9] to keep the focus on ecology.

Population biology issues include population structure, migration rates, history of population splits, and population size changes. These are the obvious reasons for OMO sequencing if the goal of the study is the management and conservation of an OMO (e.g., [10–12]). In addition, such studies can elucidate broadly relevant population biology scenarios not represented by established model organisms. Examples include enormous fecundity and dispersal potential of...
Table 1. Optimal Sequencing Methods and Reference Sequence Requirements

<table>
<thead>
<tr>
<th>Aim of study</th>
<th>Optimal sequencing method</th>
<th>Minimal reference</th>
<th>Optimal reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population subdivision, migration rates, population</td>
<td>RAD</td>
<td>None</td>
<td>Draft genome of a sister species</td>
</tr>
<tr>
<td>size changes through time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genes under selection</td>
<td>Exome sequencing</td>
<td>De novo transcriptome</td>
<td>Well-annotated genome</td>
</tr>
<tr>
<td>Selection signatures and genomic regions underlying</td>
<td>Low-coverage WGS</td>
<td>Draft long reads-based genome</td>
<td>High-contiguity genome</td>
</tr>
<tr>
<td>ecological speciation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Molecular basis of acclimatization, adaptation</td>
<td>Tag-based gene expression analysis</td>
<td>De novo transcriptome</td>
<td>High-contiguity transcriptome</td>
</tr>
<tr>
<td>Genome-wide patterns of DNA methylation (many samples)</td>
<td>RRBS-seq, methyRAD</td>
<td>None</td>
<td>Well-annotated genome</td>
</tr>
<tr>
<td>DNA methylation of specific genes (many samples)</td>
<td>MBD-seq, MeDip</td>
<td>Draft genome (vertebrates), de novo transcriptome (invertebrates, plants)</td>
<td>Well-annotated genome (vertebrates) or high-contiguity transcriptome (invertebrates, plants)</td>
</tr>
<tr>
<td>DNA methylation of specific bases (few samples)</td>
<td>ONT nanopore sequencing</td>
<td>None (de novo genome assembly form the same reads)</td>
<td>Well-annotated genome</td>
</tr>
</tbody>
</table>

*Well-annotated: resolving all or nearly all paralogous gene copies.

High-contiguity reference: genome or transcriptome containing the least amount of fragmentation. For a genome this implies Mb-scale contigs or scaffolds, and for a transcriptome in which >90% of protein-coding transcripts contain >90% of the coding sequence of the encoded protein.

Glossary

**Allele frequency spectrum (AFS):** the same as site frequency spectrum (SFS), a histogram of the number of segregating variants binned by their frequency. Can be n-dimensional for n populations [41,52] for illustrations.

**Denser-than-LD genotyping:** genotyping of several polymorphic markers per linkage disequilibrium (LD) block, which is the typical distance between markers in the genome across which their genotypes remain correlated as a result of infrequent recombination.

**Exome:** portion of the genome represented in the mature (spliced) RNA.

**Fuzzy genotyping:** performing analyses based on probabilities of alternative genotypes at each SNP without trying to decide which genotype is true [42,92]. This method is designed for lower-coverage data (as low as 1.5–2×) and is implemented in the software package ANGSD (analysis of next-generation sequencing data) [93].

**Genome scanning:** profiling of genotypes in one or more populations looking for genomic regions exhibiting unusual patterns. Typically used to look for signatures of natural selection or introgression.

**Hard-call genotyping:** identifying the most likely genotype at each SNP site and performing downstream analyses assuming that these genotypes are true. Applicable for data with 10× or better coverage [40].

**Restriction site-associated DNA (RAD) sequencing:** a family of diverse genotyping methods [45,46] that sequence short fragments of the genome adjacent to recognition sites(s) for specific restriction endonuclease(s).

**Third-generation sequencing:** methods for sequencing long individual nucleic acid molecules; these include single molecule real-time (SMRT) sequencing by PacBio and nanopore sequencing by Oxford Nanopore Technologies (ONT).
Last but not least, the role of epigenetics in acclimatization and adaptation is now one of the hottest topics in molecular ecology, often eliciting more excitement and press coverage than is warranted by existing data [19]. Whether environmentally induced epigenetic modifications play any role in acclimatization and adaptation remains unresolved. Among many covalent chromatin modifications, DNA methylation currently receives the most attention in OMOs. While vertebrates show high DNA methylation throughout the genome, invertebrates and plants methylate their genomes sparsely and mostly in protein-coding regions (so-called gene body methylation, GBM [20,21]). The function of this ubiquitous and evolutionarily ancient DNA modification remains unclear [21,22], and the greatest challenge in the next few years will be to decipher it. Once again, diversity of ecological contexts accessible through OMOs makes them ideal subjects for addressing this question.

Reference Sequences

All ecological genomic techniques (with a few exceptions, discussed later in this paper) are organized in a similar fashion: the data come in the form of anonymous sequencing reads obtained from experimental subjects, and these reads must be ‘mapped’ (matched) against some type of reference sequence (genome or transcriptome). The experimental reads must be accurate but can be short, making current Illumina HiSeq the technology of choice for their production. Requirements for the reference sequence vary considerably depending on the biological question being asked (Table 1), but there is one commonality: the reference does not need to be very accurate in terms of per-base error rate; it must only be sufficiently accurate to allow unambiguous mapping of experimental reads. The gold standard of genome sequence quality, 99.9% accuracy (or ‘Q30’ quality score), would not provide appreciable improvement to mapping efficiency compared to a rough draft accuracy of 99%. In addition, polymorphism in OMO populations is typically on the order of 1% or higher, and thus the reference is inevitably inaccurate with respect to the reads anyway. In fact, it is perfectly feasible (and in some cases even preferable, Table 1) to use a deliberately inaccurate reference from a closely related group to ‘polarize’ allelic states into ancestral and derived [23].

While per-base accuracy is not a major issue, the majority of OMO questions (with the notable exception of population biology, see below) require a highly contiguous reference. To generate such a reference there are three highly cost-efficient options available at the moment, the choice of which to use should be largely dictated by logistics and access to adequate training. It must be emphasized that, for all these methods, it is crucially important to obtain high molecular weight DNA in fragments exceeding 30–40 kb in length [24]. The technology offered by 10 x Genomics [25] attaches unique barcodes to short reads originating from the same long DNA fragment, and this allows the assembly of short reads into very long (and accurate) haplotypes. A pair of third-generation sequencing methods are capable of producing exceedingly long reads (tens to hundreds of kb) resulting in qualitatively more-contiguous genome assemblies than was previously possible [24,26–29]. Currently, read accuracy and cost of data for PacBio (Sequel system) and Oxford Nanopore Technologies (ONT; R9 flow cell) are equivalent. Although the raw read accuracy of these methods is low (ca 90%), it is sufficient for de novo OMO genome assembly based only on the long-read data [30]. Nonetheless, the best results are achieved by ‘polishing’ (i.e., error-correcting) the long-read assembly using highly accurate Illumina reads [31]; a recent benchmarking study of long-read assembly pipelines is presented in [27].

A reasonably contiguous transcriptome is a viable and cost-efficient alternative to whole-genome sequence for methods that focus on protein-coding regions, such as exome sequencing (for genome scans), gene expression analysis, and gene body methylation [21].
analysis. For OMOs with very large genomes, such as salamanders [32], transcriptome-based approaches might be the only viable genomics option. Ideally the transcriptome should correspond to the same developmental and physiological state, and the same body part, that are the subject of the OMO study. Until recently the standard way to generate a de novo transcriptome was to perform high-coverage RNA-seq of a single individual, assemble the results with Trinity [33], and annotate using blast2GO [34] or the recently introduced eggNOG-mapper [35]. This whole assembly and annotation process can be completed within a day, and produces relatively low-contiguity transcriptomes that are still suitable as references, provided that the potential caveats (occurrence of chimeric, fragmented, or duplicated transcripts, merged paralogous sequences, as well as missing or incorrect annotations) are recognized [36].

For the diversity of methods for de novo transcriptome next-generation sequencing and assembly see [2,37]; see also the detailed walkthrough for RNA-seq data filtering, assembly, and transcriptome quality assessment. With third-generation long-read sequencing technologies coming of age, it will be possible to generate much higher-quality de novo transcriptomes by directly sequencing full-length transcripts [38]. Kits and protocols for full-length third-generation transcriptome sequencing are already available (iso-seq by PacBio, and direct RNA/cDNA sequencing by ONT), and methods to derive de novo gene models from such data are under active development (e.g., [39]).

Sequencing Coverage and PCR Duplicates

There are two ways to perform genotyping – hard-call genotyping for high-coverage data and fuzzy genotyping for low-coverage data. The fuzzy approach is very appealing for OMOs because it provides unbiased estimates of allele frequencies at much lower sequencing coverage than the hard-call approach (1.5–2× versus >10×) [40]. This frees the budget to increase sample size, which in turn results in much higher accuracy of allele frequency estimation in populations [41,42]. The only potential source of bias in fuzzy genotyping is its reliance on the assumption of random mating (Hardy–Weinberg equilibrium) to calculate anticipated proportions of genotypes in a population. This assumption approximately holds for outbred natural populations (for example, humans or mice), and for highly deviant cases such as domesticated or asexually reproducing populations it is possible to account for individual-level inbreeding [43]. Nevertheless, fuzzy analysis will try to smoothen any non-random-mating signals in the data, such as hidden population structure or genetic clines, especially at very low coverage when the data are relatively weak and the method must rely more on the prior assumptions. Whether this ‘smoothing’ tendency could lead to incorrect inferences remains unclear. It also should be noted that hidden population structure or genetic clines would create biases in any analysis if left unaccounted for, and they must therefore be detected at the initial data exploration stage (Box 1).

Sequencing coverage must be calculated after removal of PCR duplicates. Generally, PCR duplicates must be removed for the statistical methods of both hard-call and fuzzy approaches to be applicable [44,45] (Box 1). Identifying PCR duplicates based on identical mapping location is a standard procedure in whole-genome sequencing (WGS); however, this used to be a challenge for restriction site-associated DNA (RAD) sequencing that by design generates reads mapping to the same location [45,46]. In the majority of current RAD implementations, PCR duplicates are identified by degenerate tags that uniquely mark each independently generated fragment of a given locus [47]. The same degenerate tag approach for PCR duplicate removal is implemented in the OMO-friendly gene expression analysis method, TagSeq [48,49]. Counter-intuitively, the proportion of PCR duplicates depends not on the number of PCR cycles performed during library preparation, but on the ratio between the number of reads sequenced (N) and the number of unique fragments present in the sample.
before PCR ($N_0$). Even with an infinite number of PCR cycles, the fraction of duplicates is the same as would be expected when sampling $N_f$ from $N_0$ with replacement. That said, low $N_0$ typically necessitates more PCR cycles during library preparation, but these additional cycles do not cause additional PCR duplicates—they are only an indication that $N_0$ is low and duplicates will be abundant. The only way to reduce the proportion of duplicates is to ensure that the original sample is highly representative (i.e., $N_0$ is not much smaller than $N_f$).

**Filters**

The lack of clear guidelines for genotype filtering is arguably the greatest source of irreproducibility in ecological genomics. It is essential to explore different filter settings to confirm that the final results are robust, and always report the filter settings used (Box 1).

There are two types of filters: those that distort the **allele frequency spectrum** (AFS) and those that do not. The former class, filters giving preference to more common variants (such as snp_pvalue filter in the software package ANGSD – analysis of next-generation sequencing data) are directly or indirectly dependent on allele frequency (AF). They are very powerful in eliminating false SNP calls caused by sequencing errors, which manifest themselves as very rare alleles appearing only once or twice in the whole dataset. AF-based filters should be applied in any situation when rare alleles are irrelevant for analysis. Examples include studies of relatedness, principal component analysis of genetic diversity, genotype–phenotype association, population differentiation based on fixation index ($F_{ST}$), STRUCTURE or ADMIXTURE, and genome scanning.

AFS-preserving filters are less efficient than AF-based filters in removing sequencing errors, but they are the only type that should be used if the main analysis is to be based on the AFS. For other studies, these filters should also be used in combination with the AF-based filter. At the very least, the data should be filtered for mapping quality (uniqueness of read mapping in the genome) and base call quality (probability of erroneous base calls in the read). It is also important to filter against excessive heterozygosity, which is a common OMO artifact if the draft reference contains a single locus instead of several paralogous copies ("lumped paralogs"). One of the most powerful AFS-preserving filters is the genotyping rate filter that requires each SNP site to be genotyped in a specified minimal proportion of all individuals. Ideally, genotyping rate should be kept above 80%, and definitely above 50%. This is particularly important for RAD methods which otherwise would suffer from null alleles ("allele dropout") due
to SNPs in the restriction enzyme recognition sequence [45,50]. Other useful AFS-preserving filters are the strand bias filter, that controls for evenness of direct and reverse strand representation [this filter is not applicable to RAD data except GBS (genotyping by sequencing) and 2b-RAD because of strand-specificity of the reads], and the het-bias filter that controls for evenness of allele representation in a heterozygous individual.

**Population Biology**

OMO population genetics has not yet fully realized the power and versatility of AFS-based demographic inference (Box 2). The undistorted AFS (obtained without AF-based filters, see above) can be used to derive the whole suite of demographic parameters, most importantly population sizes, migration rates, and history of population splits, and statistically test for their significance (Box 1). These methods are based on coalescent simulations (Fastsimcoal2 [51]), diffusion approximation (iaik, [52]), or ordinary differential equations (Moments [53]). The new ‘Moments’ method is particularly promising because it is substantially faster and more flexible than its predecessors. For inferring historical changes in population size from a single-population AFS the model-free StairwayPlot method [54] provides an easy solution.

The data required for all these analyses are several thousand biallelic SNPs, aggressively filtered to exclude potential sites under selection [55]. These SNPs can be scattered throughout the genome, and this makes various flavors of RAD sequencing best suited for this analysis. Ideally (although not necessarily) these SNPs must not be physically linked to represent independent datapoints, in which regard it is worth mentioning that RAD flavors differ considerably in the number of unlinked genomic loci that they interrogate [45,46]. In our experience, δαβ [52] and

---

**Box 2. AFS Models**

In the world of OMOs we are usually dealing with samples from many populations, which would be difficult or impossible to model simultaneously; moreover, many populations are usually left unsampled. To infer meaningful demographic parameters in a sparsely sampled system of many populations, a practical solution is to perform 2D AFS analysis of all population pairs [55]. Typical hypotheses and corresponding tests are:

(i) Are the two populations demographically separate?

Compare model with split to model without split, under which the two populations are regarded as two samples from the same one.

(ii) If yes, is there still gene flow between them?

Compare split models with and without migration.

(iii) If yes, is the gene flow symmetric or asymmetric?

Compare split model with two potentially different migration rates to a split model with a single symmetrical migration rate.

(iv) Was population size stable or went through changes?

Reconstruct population size history using StairwayPlot [54].

Simple command-line scripts for AFS plotting and running basic pairwise models in Moments can be found at https://github.com/z0on/AFS-analysis-with-moments. To access the full potential of Moments, however, the user is expected to compose their own python scripts.
Moments [53] work robustly with 10–15 thousand unlinked SNPs (a typical output from 2bRAD [56]) when analyzing populations individually or in pairs [55] (Box 1).

A major advantage of RAD-based AFS analysis for OMOs is that genotyping can be performed based on RAD reads themselves, without the need for a reference genome. Many de novo RAD pipelines exist that can assemble RAD reads into loci [45], followed by hard-call genotyping. Fuzzy genotyping de novo can also be implemented by extracting the assembled RAD loci and using them as the ‘read-based reference’ [56] to map the original reads to. This would generate BAM files that could be fed into the ANGSD software package.

With all the appeal of de novo analysis, using a reference genome to call RAD genotypes provides three important advantages. First, it identifies physically linked (and thus potentially non-independent) groups of SNPs to be resampled as units during bootstrap. Second, mapping to reference automatically discards reads from contaminant DNA sources (viruses, bacteria, ingested food, symbionts, etc). To be able to discard such contaminants in de novo RAD pipeline, the experiment must include at least one sample generated from a ‘clean’ source and consider only the RAD loci observed in that sample. Third, and most importantly, with reference-based genotyping it is possible to discriminate between ancestral and derived SNP alleles, and this gives considerably higher power to the AFS-based inference. Counter-intuitively, the best reference for this purpose is not a genome of the species under investigation but a genome of a related outgroup species that separated from the focal one a few million years ago. SNP states in the outgroup can be assumed to represent the ancestral states (e.g., [23]). Although some proportion of ancestral states will be misidentified owing to incomplete lineage sorting, convergence, or technical artifacts, this error is easy to account for by including a single additional parameter into the model that specifies the proportion of the AFS that needs to be flipped when modeling the data (e.g., [57]).

**Genome Scanning**

Because outlier regions by definition occupy only a small portion of the genome, and typically do not form a single cluster, their confident detection requires denser-than-LD genotyping. It has been argued that RAD-like approaches sample the genome too sparsely to satisfy this requirement [58,59]. Other researchers disagree [60]; indeed, many successful genome scans based on RAD have been published [61]. Nevertheless, RAD cannot be generally recommended for genome scanning because by its very design it leaves a considerable fraction of the genome unexplored. As an alternative, low-coverage WGS data suitable for fuzzy genotyping can be obtained for about $50 per sample, including both library preparation and sequencing costs [62]. Another attractive RAD alternative suitable for even very large genomes is ‘home-made exome’ sequencing: in OMOs, the exome can be efficiently captured using bead-bound normalized cDNA [63] obtained from the OMO itself (EecSeq – expressed exome capture sequencing; [97]). Such sequencing would interrogate essentially all functionally interpretable genetic variation for a cost similar to RAD. Even when linkage disequilibrium (LD) is extensive enough for RAD to produce denser-than-LD genotyping – for example, when the population is known to have gone through a recent bottleneck or represents experimentally generated progeny of a few parental individuals crossed several generations ago – a better option might be to take full advantage of the extended LD and go instead for the ultra-low coverage WGS [62] followed by haplotype imputation [64]. This approach can generate full-genome phased data, an exciting possibility that thus far remains unexplored in OMOs. Sequencing of population pools (PoolSeq [65,66]) is generally not recommended because it does not allow follow-up analyses based on individual genotypes [such as principal component analysis (PCA), STRUCTURE, or genotype-phenotype association], and does not save much cost compared...
to low-coverage WGS [62]. Finally, genotyping based on transcriptomic data (RNA-seq or TagSeq) is certainly possible [67,68], but it is unclear how (and in which cases) to account for variation in genotyping accuracy depending on the expression level of genes and alleles.

**Gene Expression**

Gene expression by itself cannot be a conclusive proof of involvement of specific genes and pathways in the process of interest, especially in OMOs, in which gene annotations are tentative at best. Nevertheless, with appropriate experimental design strong and definitive conclusions can be reached by carefully characterizing the overall patterns of gene expression variation. For example, Toth et al. [69] compared gene expression in brains of paper wasp castes to show that the origin of the worker caste involved heterochronic redeployment of maternal care program. Voolstra et al. [70] have demonstrated that in a reef-building coral establishment of symbiosis with an appropriate strain of algae does not elicit a transcriptomic response from the host, and therefore this process depends not on active recognition by the host but on the ability of the algae to enter the host in the ‘stealth’ manner. Reid et al. [71] found that populations of killifish adapting to toxic pollutants repeatedly lost (yes, lost – surprisingly) the capacity to respond to these pollutants at the gene expression level. Kenkel and Matz [72] have shown that overall gene expression plasticity varies among coral populations, and that this variation contributes to local adaptation. Campbell-Staton et al. [73] have shown that a green anole population that experienced a single cold snap acquired stable gene regulatory modifications to make its gene expression patterns more similar to those of cold-adapted populations from higher latitudes. The latter two studies also employed one of the most powerful annotation-independent gene expression analysis methods, WGCNA (weighted gene coexpression network analysis [74]), to reduce the whole-genome gene expression data to a few dozen coregulated gene clusters for easy exploration of complex experimental designs.

One OMO-suitable approach to functionally interpret gene expression patterns is to integrate over large evolutionarily conserved functional groups of genes, which averages over occasionally missing or inaccurate annotations. This approach is implemented in a diverse family of methods based on gene ontology (GO) [75] which test for over-representation of specific functional categories among significantly differentially expressed genes. The need to impose an arbitrary significance cutoff for gene expression is circumvented in methods such as GO_MWU [76] or GSEA [77], which test whether a specific functional category is enriched with either up- or downregulated genes. There is also a method similar to GO_MWU that is based on even broader eukaryotic orthologous gene groups (KOG [35,78]), which is particularly suitable for statistical comparison of very diverse datasets, even from different OMO species. For example, Strader et al. [79] used this method (R package KOGMWU [80]) to demonstrate that the red fluorescent phenotype in coral larvae is associated with a physiological state resembling midge diapause (stress-tolerant quiescent state), which in the coral larvae can be interpreted as adaptation for long-range dispersal.

Typical RNA-seq [81] resequences the whole transcriptome in each sample, but there is a much more economical way to count abundances of protein-coding transcripts: sequence only a single fragment for each transcript molecule and count reads corresponding to each gene. TagSeq [48] sequences a single randomly generated fragment near the 3’-end of the transcript, which is the most economic use of sequencing effort and removes bias towards longer transcripts. In a recent benchmarking study TagSeq was even more accurate than the standard RNA-seq in measuring transcript abundances, despite nearly 10-fold lower cost [49]. The more recently introduced QuantSeq [82] is conceptually similar to TagsSeq but uses a different library preparation procedure, implemented as a kit from Lexogen®. Both TagSeq and QuantSeq
require small quantities of initial material (10 ng of total RNA), and their bioinformatic analysis is highly simplified compared to RNA-seq. TagSeq was originally designed for OMOs and therefore its pipeline uses transcriptome rather than genome as a reference. All that being said, the major strength of TagSeq and QuantSeq – specific focus on quantification of polyadenylated transcripts – is at the same time their major weakness: these methods do not provide any other information about transcriptome than can be extracted from typical RNA-seq data, such as splice isoforms, structural rearrangements of the coding sequence, and non-coding regulatory RNAs.

DNA Methylation
While the industry-standard method of DNA methylation analysis is the whole-genome bisulfite sequencing (WGBS-seq [83]), much more OMO-friendly approaches exist. In ecology, characterization of general patterns of variation with respect to environment, population, and individual genotype is of prime interest but requires analysis of a large number of individuals. For general partitioning of variation in DNA methylation highly cost-efficient solutions are provided by RRBS-seq (reduced representation bisulfite sequencing-seq [84]) and especially methylRAD [85], which are basically RAD-like reduced representation methods for quantifying CpG methylation. Like genotyping RAD approaches, these methods can be implemented without a reference genome.

The next level of inquiry is to identify specific genes undergoing epigenetic modification, which requires gene-level resolution and must interrogate every gene in the genome. Pull-down methods MBD-seq (methyl-CpG binding domain-based capture and sequencing [86]) and meDIP-seq (methyl-DNA immunoprecipitation and sequencing [87]) provide the most economical solution in this case [88]. Unlike WGBS-seq, they concentrate sequencing effort on the methylated portion of the genome, which saves cost dramatically for non-vertebrate OMOs (invertebrates and plants) because their genomes are only sparsely methylated. In addition, because in non-vertebrate OMOs DNA methylation predominantly occurs in coding regions [20,21], it is feasible to use transcriptome instead of genome data as a reference for these analyses.

Single-base resolution for DNA methylation would rarely be of interest in ecological genomics of OMOs. Its only advantage is potentially better insight into the molecular mechanisms involved, which only becomes relevant after patterns of methylation variation with respect to the environment are established. Nevertheless, if required, such ultimate resolution can be achieved via direct detection of modified bases by nanopore sequencing [89]. This exciting new development still requires validation in complex genomes, however. Although SMRT (single molecule, real-time) sequencing by PacBio also claims the ability to detect methylated DNA bases [90,91], it appears to have low sensitivity for the most ubiquitous DNA methylation mark (5-methylcytosine). It is also important to remember that, as in WGBS-seq, achieving robust per-base methylation quantitation with these technologies will require very high sequencing coverage.

Concluding Remarks
In the past 5–7 years a great diversity of low-cost genomic approaches have emerged, some of them driven by advances in sequencing itself (such as 10x Genomics and ONT nanopore sequencing), and some, such as fuzzy genotyping and AFS-based demographics, by advances in statistical analysis of genome-wide variation. These methods now make it entirely feasible to address fundamental questions of ecological genomics in any organism within a
few years of work, without any prior sequencing resources (see Outstanding Questions). As a result, the power to answer broadly relevant ecological questions now more than ever depends on the good choice of the subject. It is time to go back to studies of natural history to identify the most curious OMOs and capitalize on their peculiarities.

References

32. Sclavi, B. and Herrick, J. (2016) Genome size variation and speciation in natural populations? Which pilot experiments could help decide whether ultra-low coverage WGS with imputation might be a feasible strategy for a particular organism?

Can methylated DNA bases be reliably detected by third-generation sequencing in complex genomes? Pilot data on bacterial DNA are very promising, but additional validation in complex genomes is required.

Outstanding Questions

Are there any biases specific to the probabilistic (‘fuzzy’) genotyping approach based on low-coverage data? Several studies have demonstrated how fuzzy genotyping can eliminate allele frequency biases at low sequencing coverage, but are we sacrificing anything to achieve this?

What are the limits to genotype imputation in natural populations? Which pilot experiments could help decide whether ultra-low coverage WGS with imputation might be a feasible strategy for a particular organism?
Trends in Genetics


40. Han, E. et al. (2014) Characterizing bias in population genetic inferences from low-coverage sequencing data. Mol. Biol. Evol. 31, 723-735


89. Stoiber, M.H. et al. (2017) De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. bioRxiv https://www.biorxiv.org/content/early/2017/04/10/094672