Trends in Genetics

CellPress REVIEWS

Review

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

Mikhail V. Matz[1,*]

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

Trends

Genotyping applications are undergoing a shift from high-coverage, reduced representation sequencing to low-coverage, whole-genome sequencing.

Approaches based on full allele frequency spectrum (AFS) to study population structure, migration rates, and historical population sizes are gaining popularity.

There has been a rise in functional genomics studies of acclimatization and adaptation, powered by cost-efficient methods for genome-wide gene expression and DNA methylation analysis.

'Third-generation' sequencing technologies (PacBio and Oxford Nanopore Technologies) have been proven to produce high-quality genome and transcriptome references, and to directly detect epigenetically modified DNA bases.

[1]Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712, USA

*Correspondence: matz@utexas.edu (M.V. Matz).

Table 1. Optimal Sequencing Methods and Reference Sequence Requirements

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

[a]Well-annotated: resolving all or nearly all paralogous gene copies.
[b]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.

marine species, life cycles featuring alternating sexual and asexual generations, parasitic or symbiotic relationships, extreme longevity, invasiveness, and many others.

Adaptation-related projects aiming to detect genomic signatures of selection, or so-called **genome scanning** (see Glossary) studies, have greatly proliferated in the past decade and their methodologies have been highlighted in several recent reviews [13–15], including an overview of genomic resources [16]. These projects are intrinsically linked to the ecology of the study species, and are often initiated in OMOs to take advantage of their existing adaptation to environmental gradients throughout their range. Genome scans are also used to look for variation in gene flow across the genome to elucidate the process of ongoing ecological speciation which might be accompanying adaptation and/or to detect adaptive introgression [17]. Like scans for selection, these studies are intimately tied to the ecology of a species and can be greatly diversified by using various OMOs as subjects.

One of the most efficient ways to elucidate molecular mechanisms of acclimatization and adaptation is gene expression analysis applied in ecological context [18], in other words comparing gene expression across environmental gradients in response to natural stressors, and in reciprocal transplantation and common garden experiments. These possibilities are not yet fully realized by OMO researchers, mainly because the low-cost gene expression profiling alternatives have only recently become available. Gene expression is also an excellent complement to selection and introgression scans [13] to help to substantiate the special role of genes highlighted by the scans.

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 site (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[i]. 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_r) and the number of unique fragments present in the sample

Box 1. Best Practices for OMO Genomics

(i) Use material from a single individual to generate the reference (either genome or transcriptome) to minimize the effects of natural polymorphisms on assembly [94].

(ii) Remove PCR duplicates, otherwise statistical methods based on read counts will not be applicable [44,45].

(iii) As an initial exploratory step, apply strong filters and examine data using ordination techniques (principal component and related analyses [95,96]). This step is necessary to identify outlier samples, sampling artifacts (wrong species collected or same specimen sampled twice), natural clones, and hidden population structure.

(iv) Try several data-filtering settings to confirm that the results are robust. Report the exact filtering settings used.

(v) As soon as the manuscript or preprint is submitted (and ideally sooner) – share new sequencing data with the research community; announce data availability through professional email lists and social media and make them downloadable; post links to datasets, well-commented scripts, and bioinformatic walkthroughs in an open-access repository (such as GitHub) such that the analyses can be evaluated and reproduced by others.

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_r 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_r).

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 (∂a∂i, [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, δaδi [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 patters 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[ii]. 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[iii] 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.

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?

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.

Resources

[i]https://informatics.fas.harvard.edu/best-practices-for-de-novo-transcriptome-assembly-with-trinity.html

[ii]www.lexogen.com/quantseq-3mrna-sequencing/

[iii]https://github.com/z0on/tag-based_RNAseq

References

1. van Dijk, E.L. *et al.* (2014) Ten years of next-generation sequencing technology. *Trends Genet.* 30, 418–426
2. da Fonseca, R.R. *et al.* (2016) Next-generation biology: sequencing and data analysis approaches for non-model organisms. *Mar. Genomics* 30, 3–13
3. Ellegren, H. (2014) Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* 29, 51–63
4. Ekblom, R. and Galindo, J. (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* 107, 1–15
5. Martin, C.H. *et al.* (2016) Diabolical survival in Death Valley: recent pupfish colonization, gene flow and genetic assimilation in the smallest species range on earth. *Proc. R. Soc. B Biol. Sci.* 283, 20152334
6. Sun, J. *et al.* (2017) Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nat. Ecol. Evol.* 1, 121
7. Pan, S. *et al.* (2017) Population transcriptomes reveal synergistic responses of DNA polymorphism and RNA expression to extreme environments on the Qinghai-Tibetan Plateau in a predatory bird. *Mol. Ecol.* 26, 2993–3010
8. McCormack, J.E. *et al.* (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol. Phylogenet. Evol.* 66, 526–538
9. Lemmon, E.M. and Lemmon, A.R. (2013) High-throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Evol. Syst* 44, 99–121
10. Selkoe, K.A. *et al.* (2016) A decade of seascape genetics: contributions to basic and applied marine connectivity. *Mar. Ecol. Prog. Ser.* 554, 1–19
11. Garner, B.A. *et al.* (2016) Genomics in conservation: case studies and bridging the gap between data and application. *Trends Ecol. Evol.* 31, 81–83
12. Benestan, L.M. *et al.* (2016) Conservation genomics of natural and managed populations: building a conceptual and practical framework. *Mol. Ecol.* 25, 2967–2977
13. Lotterhos, K.E. and Whitlock, M.C. (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Mol. Ecol.* 24, 1031–1046
14. Hoban, S. *et al.* (2016) Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *Am. Nat.* 188, 379–397
15. Vatsiou, A.I. *et al.* (2016) Detection of selective sweeps in structured populations: a comparison of recent methods. *Mol. Ecol.* 25, 89–103
16. Manel, S. *et al.* (2016) Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Mol. Ecol.* 25, 170–184
17. Harrison, R.G. and Larson, E.L. (2016) Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. *Mol. Ecol.* 25, 2454–2466
18. Todd, E.V. *et al.* (2016) The power and promise of RNA-seq in ecology and evolution. *Mol. Ecol.* 25, 1224–1241
19. Torda, G. *et al.* (2017) Rapid adaptive responses to climate change in corals. *Nat. Clim. Chang.* 7, 627–636
20. Feng, S. *et al.* (2010) Epigenetic reprogramming in plant and animal development. *Science* 330, 622–627
21. Sarda, S. *et al.* (2012) The evolution of invertebrate gene body methylation. *Mol. Biol. Evol.* 29, 1907–1916
22. Dixon, G.B. *et al.* (2016) Evolutionary consequences of DNA methylation in a basal metazoan. *Mol. Biol. Evol.* 33, 2285–2293
23. Gojobori, J. *et al.* (2007) Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. *Proc. Natl. Acad. Sci. U. S. A.* 104, 3907–3912
24. Jain, M. *et al.* (2017) Nanopore sequencing and assembly of a human genome with ultra-long reads. *bioRxiv* https://www.biorxiv.org/content/early/2017/04/20/128835
25. Kitzman, J.O. (2016) Haplotypes drop by drop. *Nat. Biotechnol.* 34, 296–298
26. Gordon, D. *et al.* (2016) Long-read sequence assembly of the gorilla genome. *Science* 352, aae0344
27. Michael, T.P. *et al.* (2017) High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *bioRxiv* https://www.biorxiv.org/content/early/2017/06/14/149997
28. Jansen, H.J. *et al.* (2017) Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *bioRxiv* https://www.biorxiv.org/content/early/2017/01/20/101907
29. Berlin, K. *et al.* (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* 33, 623–630
30. Vaser, R. *et al.* (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 27, 737–746
31. Walker, B.J. *et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9, e112963
32. Sclavi, B. and Herrick, J. (2016) Genome size variation and species diversity in salamander families. *bioRxiv* https://biorxiv.org/lookup/doi/10.1101/065425
33. Haas, B.J. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512
34. Conesa, A. *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676
35. Huerta-Cepas, J. *et al.* (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* 34, 2115–2122
36. Smith-Unna, R. *et al.* (2016) TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* 26, 1134–1144
37. Huang, X. *et al.* (2016) Comparative performance of transcriptome assembly methods for non-model organisms. *BMC Genomics* 17, 523
38. Hoang, N.V. *et al.* (2017) A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. *BMC Genomics* 18, 395

39. Marchet, C. *et al.* (2017) De novo clustering of gene expressed variants in transcriptomic long reads data sets. *bioRxiv*. https://www.biorxiv.org/content/early/2017/11/08/170035
40. Han, E. *et al.* (2014) Characterizing bias in population genetic inferences from low-coverage sequencing data. *Mol. Biol. Evol.* 31, 723–735
41. Robinson, J.D. *et al.* (2014) Sampling strategies for frequency spectrum-based population genomic inference. *BMC Evol. Biol.* 14, 254
42. Alex Buerkle, C. and Gompert, Z. (2013) Population genomics based on low coverage sequencing: how low should we go? *Mol. Ecol.* 22, 3028–3035
43. Vieira, F.G. *et al.* (2013) Estimating inbreeding coefficients from NGS data: impact on genotype calling and allele frequency estimation. *Genome Res.* 23, 1852–1861
44. Andrews, K.R. and Luikart, G. (2014) Recent novel approaches for population genomics data analysis. *Mol. Ecol.* 23, 1661–1667
45. Andrews, K.R. *et al.* (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* 17, 81–92
46. Puritz, J.B. *et al.* (2014) Demystifying the RAD fad. *Mol. Ecol.* 23, 5937–5942
47. Casbon, J.A. *et al.* (2011) A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res.* 39, e81
48. Meyer, E. *et al.* (2011) Profiling gene expression responses of coral larvae (*Acropora millepora*) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. *Mol. Ecol.* 20, 3599–3616
49. Lohman, B.K. *et al.* (2016) Evaluation of TagSeq, a reliable low-cost alternative for RNAseq. *Mol. Ecol. Resour.* 16, 1315–1321
50. Gautier, M. *et al.* (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol. Ecol.* 22, 3165–3178
51. Excoffier, L. *et al.* (2013) Robust demographic inference from genomic and SNP data. *PLoS Genet.* 9, e1003905
52. Gutenkunst, R.N. *et al.* (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5, e1000695
53. Jouganous, J. *et al.* (2017) Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics* 206, 1549–1567
54. Liu, X. and Fu, Y.-X. (2015) Exploring population size changes using SNP frequency spectra. *Nat. Genet.* 47, 555–559
55. Matz, M.V. *et al.* (2017) Potential for rapid genetic adaptation to warming in a Great Barrier Reef coral. *bioRxiv* https://www.biorxiv.org/content/early/2017/06/18/114173
56. Wang, S. *et al.* (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nat. Methods* 9, 808–810
57. Tine, M. *et al.* (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nat. Commun.* 5, 5770
58. Lowry, D.B. *et al.* (2017) Breaking RAD: an evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. *Mol. Ecol. Resour.* 17, 142–152
59. Tiffin, P. and Ross-Ibarra, J. (2014) Advances and limits of using population genetics to understand local adaptation. *Trends Ecol. Evol.* 29, 673–680
60. Catchen, J.M. *et al.* (2017) Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Mol. Ecol. Resour.* 17, 362–365
61. McKinney, G.J. *et al.* (2017) RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry *et al.* (2016). *Mol. Ecol. Resour.* 17, 356–361
62. Therkildsen, N.O. and Palumbi, S.R. (2017) Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Mol. Ecol. Resour.* 17, 194–208
63. Zhulidov, P.A. *et al.* (2004) Simple cDNA normalization using Kamchatka crab duplex-specific nuclease. *Nucleic Acids Res.* 32, e37
64. Davies, R.W. *et al.* (2016) Rapid genotype imputation from sequence without reference panels. *Nat. Genet.* 48, 965–969
65. Kofler, R. *et al.* (2011) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* 27, 3435–3436
66. Futschik, A. and Schlötterer, C. (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* 186, 207–218
67. De Wit, P. *et al.* (2015) SNP genotyping and population genomics from expressed sequences – current advances and future possibilities. *Mol. Ecol.* 24, 2310–2323
68. Bay, R.A. and Palumbi, S.R. (2014) Multilocus adaptation associated with heat resistance in reef-building corals. *Curr. Biol.* 24, 2952–2956
69. Toth, A.L. *et al.* (2007) Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science* 318, 441–444
70. Voolstra, C.R. *et al.* (2009) The host transcriptome remains unaltered during the establishment of coral–algal symbioses. *Mol. Ecol.* 18, 1823–1833
71. Reid, N.M. *et al.* (2016) The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science* 354, 1305–1308
72. Kenkel, C.D. and Matz, M.V. (2016) Gene expression plasticity as a mechanism of coral adaptation to a variable environment. *Nat. Ecol. Evol.* 1, 14
73. Campbell-Staton, S.C. *et al.* (2017) Winter storms drive rapid phenotypic, regulatory, and genomic shifts in the green anole lizard. *Science* 357, 495–498
74. Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* 9, 559
75. Ashburner, M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29
76. Wright, R.M. *et al.* (2015) Gene expression associated with white syndromes in a reef building coral, *Acropora hyacinthus*. *BMC Genomics* 16, 371
77. Subramanian, A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550
78. Tatusov, R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinform.* 4, 41
79. Strader, M.E. *et al.* (2016) Red fluorescence in coral larvae is associated with a diapause-like state. *Mol. Ecol.* 25, 559–569
80. Dixon, G.B. *et al.* (2015) Genomic determinants of coral heat tolerance across latitudes. *Science* 348, 1460–1462
81. Morin, R.D. *et al.* (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* 45, 81–94
82. Moll, P. *et al.* (2014) QuantSeq 3′ mRNA sequencing for RNA quantification. *Nat. Methods* 11, 25
83. Lister, R. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462, 315–322
84. Meissner, A. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454, 766–770
85. Wang, S. *et al.* (2015) MethylRAD: a simple and scalable method for genome-wide DNA methylation profiling using methylation-dependent restriction enzymes. *Open Biol.* 5, 150130
86. Serre, D. *et al.* (2010) MBD-isolated genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res.* 38, 391–399
87. Jacinto, F.V. *et al.* (2008) Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. *Biotechniques* 44, 35–43

88. Aberg, K.A. *et al.* (2012) MBD-seq as a cost-effective approach for methylome-wide association studies: demonstration in 1500 case–control samples. *Epigenomics* 4, 605–621

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

90. Flusberg, B.A. *et al.* (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7, 461–465

91. Feng, Z. *et al.* (2013) Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. *PLoS Comput. Biol.* 9, e1002935

92. Nielsen, R. *et al.* (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS One* 7, e37558

93. Korneliussen, T. *et al.* (2014) ANGSD: analysis of next generation sequencing data. *BMC Bioinform.* 15, 356

94. Vinson, J.P. *et al.* (2005) Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. *Genome Res.* 15, 1127–1135

95. Jombart, T. and Ahmed, I. (2011) Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071

96. Dixon, P. (2003) VEGAN, a package of R functions for community ecology. *J. Veg. Sci* 14, 927–930

97. Puritz, J.B. and Lotterhos, K.E. (2017) Expressed Exome Capture Sequencing (EecSeq): a method for cost-effective exome sequencing for all organisms with or without genomic resources. *bioRxiv* https://www.biorxiv.org/content/early/2017/11/23/223735