Fantastic beasts and how to sequence them:

genomic approaches for obscure model organisms

Mikhail V. Matz
Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712, USA
matz@utexas.edu

Summary:
Application of genomic approaches to "obscure model organisms" (OMOs), meaning species with little or no genomic resources, enables increasingly sophisticated studies of genomic basis of evolution, acclimatization and adaptation in real ecological context. Here, I aim to identify sequencing solutions and data handling techniques best suited for genomic analysis of OMOs.

Trends:

- Adoption of allele frequency spectrum (AFS) analyses for demographic and population structure studies, based on RAD sequencing data.
- Switch from RAD to whole-genome and exome sequencing in "genome scanning" studies.
- Adoption of cost-efficient gene expression quantification methods based on counting transcripts rather than on whole-transcriptome resequencing.
- Use of genetically driven gene expression signatures to examine genotype-phenotype associations and genetic basis of adaptation.
- Studies of ecological dynamics and inheritance of DNA methylation marks based on cost-efficient alternatives to bisulfite sequencing.
- Adoption of "third-generation" sequencing technologies (PacBio and Oxford Nanopore Technologies) for generation of genome and transcriptome references.

Outstanding questions:

- **How the power and accuracy of AFS analysis is affected by the limited number of RAD loci?** Although each RAD locus might contain several SNPs and thus the total SNP count might seem large, SNPs from the same RAD locus are highly correlated and so the number of independent data points is in fact much lower.
- **What are the limits to genotype imputation in natural populations?** Which pilot experiments could help decide whether low-coverage whole genome sequencing with imputation might be a feasible strategy for a particular organism?
- **How to profile genetically determined gene expression in non-clonal organisms?** Cross-tissue analysis is promising but more validation experiments are needed to develop guidelines on how many and which tissues should be profiled for best cost-benefit balance.
- **Can methylated DNA bases be reliably detected by single-molecule sequencing in complex genomes?** Pilot data on bacterial DNA is very promising but additional validation in complex genomes is required.

Glossary:

- **Allele Frequency Spectrum, AFS** (same as Site Frequency Spectrum, SFS): histogram of the number of segregating variants depending on their frequency in one or more populations.
- **Restriction site-Associated DNA (RAD) sequencing**: family of diverse genotyping methods that sequence short fragments of the genome adjacent to recognition site(s) for specific restriction endonuclease(s).
- **Linkage Disequilibrium (LD)**: in this review, correlation of genotypes at a pair of markers across individuals.
- **LD block**: typical distance between markers in the genome across which their genotypes remain correlated.
- **Genome scan:** profiling of genotypes along the genome looking for unusual patterns. Often used to look for signatures of natural selection or introgression.
- **"Denser-than-LD" genotyping**: genotyping of several polymorphic markers per LD block.
- **Highly contiguous reference**: genome or transcriptome reference sequence containing the least amount of fragmentation.
- **Phased data**: data showing which SNP alleles belong to the same homologous chromosome copy.
- **Cross-tissue gene expression analysis**: looking for individual-specific shifts in gene expression detectable across multiple tissues. Such shifts are predominantly genetic in nature.

The focus of this review is mainly on the type of sequencing data required and how to obtain it in the most cost-efficient way rather than on analytical approaches. That said, I could not help but mention some highly promising analytical methods that are not yet broadly adopted by OMO researchers, such as demographic inference based on allele frequency spectra and annotation-independent analyses of gene expression data.

I will start with the summary of general types of questions in OMO studies and corresponding data types required. We might be interested in the following four layers of genomic information, each requiring a specific type of experimental and reference data:

1. Genome-wide patterns of neutral variation. This data can elucidate population structure, population sizes, and migration rates, as well as changes of these parameters through time. This analysis benefits from high quality genotype calls but does not require dense genome coverage; it can even be performed in the absence of reference genome.

2. Regions in the genome particularly affected by non-drift processes (natural selection, introgression, etc). This type of analysis, typically referred to as "genome scanning", takes genome-wide neutral variation as baseline and looks for regions in the genome exhibiting highly dissimilar patterns. It requires "denser-than-LD" genotyping and a highly contiguous reference (see Glossary) to make sure no signal is overlooked.

3. Genome-wide gene expression, an extremely information-rich resource reflecting both environmental and genetic variation. Streamlined transcript counting methods represent a cost-efficient alternative to the industry-standard RNA-seq for generating quantitative data. Analysis of gene expression does not require a genome reference, although a transcriptome reference must be generated at some point. The reference does not have to be highly contiguous.

4. Epigenetics, here limited to DNA methylation. A variety of methods have been recently developed that can generate data for DNA methylation analysis. For vertebrates, genome reference is needed, but for other animals or plants, in which DNA methylation is much less prevalent and predominantly occurs in exons, transcriptome or exome presents a good cost-efficient alternative. The reference does not have to be highly contiguous.

Genome-wide neutral variation

Allele Frequency Spectrum analysis

Neutral genetic markers are traditionally analyzed using summaries of allele frequency differences between populations, such as F_{ST}. Large amount of markers accessible through next-generation sequencing opened up the possibility to dramatically enhance this approach by modeling the evolution of the whole allele frequency spectrum (AFS, see Glossary). AFS represents a rich source of information to fit alternative models with time-resolved population sizes and migration rates as parameters (Box 1) based on coalescent simulations (*fastsimcoal2*, [1]), diffusion approximation (*dadi*, [2]), or ordinary differential equations (*moments*, [3]). Model selection is then based on likelihood ratio tests or Akaike information criterion. The new *moments* method is particularly promising, as it is substantially faster than its predecessors and includes built-in bootstrap, demographic model plotting, and capacity to analyze up to five populations simultaneously. It is also very helpful that *moments* inherits the python code structure well familiar to *dadi* practitioners.

Experimental data

The data required for AFS analysis is several thousand biallelic neutral single nucleotide polymorphisms (SNPs). Ideally, SNPs must not be closely physically linked in the genome to represent independent data points, although it is fully appropriate to analyze linked SNPs with AFS methods. The lack of requirement for contiguous SNP coverage makes various flavors of restriction site-associated DNA (RAD) sequencing (recently reviewed in [4,5], see Glossary) well suited for this analysis. In our experience, *dadi* [2] and *moments* [3] work robustly with 5-10 thousand SNPs (a typical RAD output) when analyzing populations individually or in pairs. Fitting models with three (*dadi*) or more (*moments, fastsimcoal2*) populations might be problematic with this relatively low number of SNPs but is usually not required for OMOs (Box 1). Recent population size changes are often of special interest in OMOs; since they predominantly affect rare alleles, their robust detection requires 20 or more high-quality genotypes per population [6]. This preference for more individuals rather than more SNPs per

individual is an additional factor that makes cost-efficient RAD the approach of choice for AFS-based analysis. That said, relatively low number of independent (unlinked) SNPs generated by some RAD protocols might limit the power of the AFS analysis, and a good subject for a future study would be the effect of the number of unlinked SNPs on AFS model selection and uncertainties of parameter estimates. In this regard it is worth noting that RAD flavors differ considerably in the number of unlinked loci in the genome that they interrogate [4,5].

Box 1: AFS models.
In the world of OMOs we are usually dealing with samples from many populations, which would be hard or impossible to model simultaneously; moreover, there are usually many populations left unsampled. To infer meaningful demographic parameters in a sparsely sampled system of many populations, a reasonable practical solution is to perform two-dimensional AFS analysis of all population pairs [10]. Typical hypotheses and corresponding tests are:

- Are the two populations demographically separate?
    - compare model with split to model without split, under which the two compared populations are regarded as independent samples from the same population.
- If yes, is there still gene flow between them?
    - compare split models with and without migration.
- 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.
- Was population size stable or went through changes in the past?
    - compare single-population model with a size change in the past to a standard neutral model.

Simple command-line scripts for AFS plotting and running basic pairwise models in *moments* can be found here: https://github.com/z0on/AFS-analysis-with-moments. To access the full potential of *moments*, however, the user is expected to compose python scripts of their own.

For demographic inference, the AFS data must be filtered to exclude potential sites under selection. Whichever test is used to identify such sites (for example, Bayescan, [7]), for their removal the false discovery rate should be set as high as 0.5 to ensure purging of the majority of non-neutral sites. Although under this setting half of the removed sites would be neutral, their removal will not affect the overall AFS as long as the removed fraction does not comprise more than 1-2% of the total number of sites.

Genotyping quality

In diploids, the most common genotyping error is missing one of the alleles in a heterozygote (i.e., a false homozygote call); and the next most common error is missing the whole SNP locus entirely. Both these “missing data” errors are due to insufficient sequencing coverage, the problem that is pervasive in today’s OMO studies. Such errors strongly affect AFS in the region of rare alleles, which is unfortunate since rare alleles are the most informative about recent population history [6,8]. A telltale sign of poor heterozygote calling is under-representation of singletons, but frequencies of doubletons and higher-order frequency bins are also distorted,

155 which has strong effect on AFS itself and inferred demographic parameters until mean
156 sequencing coverage approaches ~10x [9]. When coverage is 10x or higher a good way to filter
157 data is to select SNPs genotyped in >90-95% of samples [10]; importantly for RAD approach,
158 this would select SNPs that are unlikely to be affected by null alleles due to mutation in the
159 restriction endonuclease recognition site [4]. For obvious reasons, for AFS analysis genotype
160 calls should never be quality-filtered based on allele frequencies (for example, retaining only
161 variants that are detected in a minimum of two individuals or requiring minor allele frequency to
162 exceed some cutoff). A robust empirical way to evaluate the consistency of genotype calls is to
163 compare results for independently processed biological samples of the same genotype [11]. Such
164 genotyping replicates are quite feasible in RAD and are also useful to identify true SNPs for
165 training variant quality score recalibration model of the GATK pipeline [12]. For low-coverage
166 data (<10x), a general solution is provided by the *ANGSD* package [13], which generates AFS as
167 well as other population genetic statistics based on genotype likelihoods without actually calling
168 genotypes [14]. This method generates unbiased single-population AFS even with 2x coverage
169 [9]. Still, there is a concern that high variation in coverage across samples and populations might
170 affect *ANGSD* statistics; to avoid this potential issue it is recommended to discard the lowest-
171 coverage outliers and down-sample reads from highest-coverage outliers (J. Ross-Ibarra, pers.
172 comm.).
173
174 *PCR duplicates*
175
176 Presence of PCR duplicates in many early RAD applications has been repeatedly highlighted as a
177 source of genotyping errors [4,15] due to induced over-dispersion of read counts among alleles
178 and loci. Interestingly, the proportion of PCR duplicates does not depend on the number of PCR
179 cycles performed during library preparation. Instead, it depends on the ratio between the number
180 of reads sequenced (N_r) and the number of unique fragments present in the sample prior to PCR
181 (N_o): the fraction of duplicates is the same as expected when sampling N_r from N_o with
182 replacement. Fortunately, PCR duplicates are easy to identify and remove using degenerate tags
183 ligated to RAD fragments prior to amplification [16]. Most present-day RAD protocols now
184 implement this simple deduplication procedure, including the current version of 2bRAD [11].
185
186 *Genome reference for AFS analysis*
187
188 A great advantage of RAD-based AFS analysis for OMOs is that SNPs can be called based on
189 RAD reads themselves, without the need for genome reference. Several *de novo* RAD genotyping
190 pipelines have been developed, such as STACKS, pyRAD, and UNEAK (see references in [4])
191 that work for most RAD flavors, plus a similarly structured pipeline for 2bRAD
192 (https://github.com/z0on/2bRAD_denovo) that takes into account the fact that in 2bRAD either
193 strand of the locus can be sequenced. Still, using a reference genome to call RAD genotypes
194 provides three important advantages. First, it identifies physically linked (and thus potentially
195 non-independent) groups of SNPs, to be resampled as units during AFS bootstrap. The second
196 advantage is particularly important for OMOs sampled in the field: mapping to reference genome
197 automatically discards reads from contaminant DNA sources (bacteria, ingested food, symbionts
198 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.

The third advantage of reference-based genotyping is the possibility to discriminate between ancestral and derived SNP alleles, to attain the best power of AFS-based inference. Counter-intuitively, the best reference for AFS analysis is not a genome of the species under investigation but a genome of a related outgroup species, separated from the focal one by a few million years of evolution, because the SNP state as in the outgroup can be assumed to represent the ancestral state (e.g., [17]). Although some proportion of ancestral states will be misidentified in this way (due to incomplete lineage sorting, convergence or technical artifacts), this error is easy to account for by including a single additional parameter into the model, specifying the proportion of the AFS that needs to be flipped when predicting the data (e.g., [18]). The reference for AFS does not have to be highly contiguous; the contigs should be just long enough to cover a typical LD block for meaningful bootstrapping.

Genome scanning

Since 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 (see Glossary). It has been argued that in most situations, RAD-like approaches would sample the genome too sparsely to satisfy this requirement [19,20]. Although many successful genome scans based on RAD have been published [21], RAD cannot be recommended for genome scanning since it would inevitably leave a considerable fraction of the genome unexplored. Even when LD is known to be extensive enough for RAD to produce "denser-than-LD" genotyping a better solution would be to take full advantage of the extended LD and go for ultra-low coverage whole-genome sequencing (WGS) with imputation instead, to obtain full-genome phased data (Table 1).

The types of sequencing approaches for genome scanning with their pros and cons are summarized in Table 1. Importantly, all of them require highly accurate reads mapped to a reference for confident SNP detection, making short Illumina reads the genotyping data type of choice. Some of the very promising approaches that have not yet been fully adopted for OMOs are exome-seq and ultra-low WGS with imputation. Exome-seq used to be a prerogative of model organisms because of the need for exome-capture platform development, but it has recently been shown that OMO exome can be captured just as efficiently using bead-bound normalized cDNA obtained from the OMO itself (EecSeq Puritz 2017). Such "home-made exome" sequencing could become an excellent alternative to RAD since it would interrogate essentially all the interpretable genetic variation for a comparably low cost. Ultra-low WGS with imputation used to require extensive reference haplotype panels available only for well-established model organisms. However, several methods have been recently developed (most notably STITCH, [22]) that can impute phased genotypes and correct genotyping errors in ultra-low coverage data without relying on reference panels. Still, their applicability for each new OMO must be experimentally confirmed because the success of imputation critically depends on the level of polymorphism relative to the length of LD blocks, and these parameters are typically not known for OMOs *a priori*. Demographic events such as strong recent bottleneck, domestication, or recent

243 colonization would make imputation more efficient because of more extensive LD and small
244 number of founding haplotypes [22], and conversely, in large outbred populations imputation will
245 be less accurate and might require sequencing of a very large number (thousands) of individuals.
246 The accuracy of imputation can be evaluated by sequencing a few individuals at high coverage
247 (>10x) to generate high-confidence genotype calls and then attempting to impute them based on
248 sub-sampled read sets to emulate low coverage. It must be noted that it is inappropriate to
249 measure imputation accuracy by imputing genotype calls masked in high-quality datasets (as in,
250 for example, [23]): masked data do not contain false homozygote calls and therefore do not
251 correctly represent the real-life situation.
252
253 **Table 1. Genotyping approaches for genome scanning.**

Approach	Features	Pros	Cons
Exome-seq	Isolates and sequences only the protein-coding portion of genome .	Dense coverage of genes guarantees that coding variants and variants linked to cis-regulatory mutations are discovered.	Other (arguably less important) types of variation are not profiled (e.g., distant enhancers).
RNA-seq [47,48]	Sequences RNA.	Same as exome sequencing.	Genotyping quality of a gene depends on expression level. Allele-specific expression affects heterozygote calls
Pool-seq [49,50]	Sequences pooled DNA from multiple individuals from each population.	Dense whole-genome coverage with confident determination of allele frequencies in populations.	No possibility for individual–based analysis (such as STRUCTURE) or validation based on genotype-phenotype association across individuals. Must be confident in *a priori* population designations.
Low-coverage whole-genome sequencing (WGS) [51]	Sequences individual genomes at ~1-4x coverage.	Dense whole genome coverage at individual level.	Per-site genotypes are unreliable because of missing data; must use uncertainty-aware analysis such as *ANGSD*.
Ultra-low coverage WGS with imputation [22]	Sequences individual genomes at <2x coverage, imputes missing genotypes and corrects false homozygote calls	Dense whole genome coverage at individual level, phased data enables haplotype-based analysis	Rare alleles (minor allele frequency<0.05) are missed. Requires large sample sizes (depending on LD, hundreds or thousands of individuals). Accuracy of imputation must be experimentally validated for every new OMO.

254
255 **Gene expression**
256
257 There are many aspects to gene expression, of which I here focus on just one: abundance
258 or protein-coding (polyadenylated) transcripts. The reason is that transcript abundance is by far

259 the most interpretable of all aspects of gene expression, and it can be very easily analyzed in
260 OMOs.
261
262 *Counting transcripts instead of resequencing them*
263
264 Typical RNA-seq [24] resequences the whole transcriptome in each sample, but there is a
265 much more economic way to count abundances of protein-coding transcripts: sequence just a
266 single fragment per each transcript molecule and count reads corresponding to each gene. TagSeq
267 [25], for example, sequences a single randomly generated fragment near the 3'-end of the
268 transcript, which is the most economic use of sequencing effort and removes bias towards longer
269 transcripts. In a recent benchmarking study TagSeq was actually more accurate than the standard
270 RNA-seq in measuring transcript abundances, despite nearly tenfold lower cost [26]. More
271 recently introduced QuantSeq [27] is conceptually very similar to TagsSeq: it also sequences a
272 single randomly generated fragment near the 3'-end of each transcript but has a different library
273 preparation procedure, implemented as a kit from Lexogen (https://www.lexogen.com/quantseq-
274 3mrna-sequencing/). Bioinformatics analysis for both TagSeq ad QuantSeq is highly simplified
275 compared to typical RNA-seq. TagSeq was originally designed for OMOs and so its pipeline uses
276 transcriptome (rather than genome) as a reference to attribute reads to genes
277 (https://github.com/z0on/tag-based_RNAseq). One notable feature of the current version of
278 TagSeq pipeline is that it includes removal of PCR duplicates based on adaptor-derived
279 degenerate tags [11], similarly to 2bRAD and for the same reason – to avoid PCR-associated
280 over-dispersion or read counts.
281
282 *Analysis of gene expression "beyond gene lists"*
283
284 The unfortunate tradition that OMO research inherits from the biomedical field is putting
285 too much emphasis on possible functional implications of expression changes of specific genes.
286 For OMOs, this is bound to remain inconclusive because gene annotations are often absent,
287 tentative or based predominantly on similarity to human genes, which may or may not serve the
288 same function in the studied OMO. Even greater problem is interpretation bias: too often
289 researchers focus primarily on genes that "make sense" and ignore the rest. This leads to
290 conclusions reflecting predominantly the researchers' idea of what *should* be going on rather than
291 what is actually happening.
292
293 Table 2 lists alternative ways of objective analysis of gene expression data that are
294 enabled by the large sample sizes feasible with TagSeq or QuantSeq. They either do not require
295 gene annotations or rely sufficiently general functional summaries to be robust to occasional
296 missing or mis-annotations. Particularly useful for OMOs are analyses that use gene expression
297 patterns as anonymous multivariate readouts to compare and classify samples, such as principal
298 coordinate analysis (PCoA) or differential analysis of principal components (DAPC). With
299 appropriate experimental design, these analyses can lead to much more definitive biological
300 conclusions than studies scrutinizing long lists of differentially expressed genes passing a certain
301 significance cutoff.
302

303 **Table 2. Gene expression analyses not relying on accurate gene annotations**

Analysis	What does it do	Software	Applications
Principal coordinate analysis based on Manhattan distances (sum of all log-fold changes across genes)	Characterizes overall transcriptome similarities and differences between experimental groups. Measures fraction of variation attributable to each experimental factor.	R: package *ape*, function *pcoa* [52] package *vegan,* function *adonis* [53]	[54,55]
Differential analysis of principal components (DAPC)	Quantifies transcriptome differences between samples with respect to specified multivariate axis. Good for quantifying overall gene expression plasticity.	R: package *adegenet* [56]	[28]
Weighted gene co-expression network analysis (WGCNA)	Identifies co-regulated groups of genes, which are linked to experimental factors and traits *post hoc*. Method of choice for complex experimental designs (>20 samples) with many quantitative traits measured.	R: package WGCNA [57]	[28,58,59]
Rank-based functional summaries of KOG (euKaryotic Orthologous Groups) classes	Reveals broad functional trends in gene expression. Particularly useful for OMOs since it tolerates sparse and inaccurate annotations. Its main use is for statistical comparison of highly diverse datasets, even from different species.	R: package *KOGMWU* [11].	[11,55]

304

305 *Gene expression as functional summary of genotype*

306

307 Gene expression is best known for its context-dependence reflecting phenotypic plasticity, which
308 is the view inherited from biomedical research dealing with genetically uniform models. In
309 natural populations, one of the most important sources of gene expression variation are genetic
310 differences among individuals, manifested as context-*in*dependent, individual-specific deviations
311 in gene expression. In two reciprocal transplantation experiments that we have performed on
312 different coral species in different oceans, stable between-genotypes differences accounted for
313 more than 50% of total gene expression variation despite transplantation duration of up to a year
314 [28,29]. In non-clonal model organisms such as mice or humans, the direct demonstration of the
315 effect of genetic variation on gene expression are abundant differences in expression between
316 alleles of the same gene [30,31]. Fixed between-population differences are best exemplified by
317 hundreds of genes are differentially expressed between African and European Americans [32].
318 All this suggests that gene expression can be a proxy of not only phenotypic plasticity and
319 acclimatization, but of genetic variation and adaptation. A major advantage of the use of gene
320 expression for these types of studies is that gene expression integrates over many functionally
321 relevant variants in the genome and thus represents a condensed functional summary of the
322 genotype [33].
323

324 In humans, nearly half of all genetic variants affecting gene expression have detectable effects in
325 all tissues [31], and so one feasible way to separate genotype-specific gene expression from
326 context-dependent variation might be to perform "cross-tissue" comparison (see Glossary) to
327 isolate body-wide expression shifts [34]. In the coming years, cross-tissue or similar analysis is
328 likely to become a major approach to study *functional* genetic variation in natural populations.
329
330 **Epigenetics**
331
332 **Table 3. Methods for interrogating DNA methylation**

Method	Features	Pros	Cons
BS-seq [60]	Sequences complete genome after bisulfite conversion	Complete characterization of 5me-cythosine methylation at single-base resolution	High coverage is required to obtain quantitative data. In non-vertebrate OMOs, much sequencing effort is wasted since most of genome is not methylated.
RRBS-seq [61]	RAD-like sequencing of bisulfite-converted DNA	Saves costs compared to BS-seq.	Low genomic resolution. Important signals may be missed in between interrogates sites. Wastes sequencing effort on non-methylated sites.
MBD-seq [62], meDIP [63]	Pull-down and sequencing of methylated DNA.	Optimizes sequencing effort by focusing on methylated DNA. Good for studying per-gene methylation levels.	Resolution equals the length of pulled-down fragments (~150b). Requires relatively expensive library prep (standard Illumina DNA-seq) and >10 million reads per sample.
methylRAD [64]	Flavor of 2bRAD targeting all symmetrically methylated CCGG and CCWGG sites in the genome.	Extremely cost-efficient due to combination of RAD approach and focus on methylated DNA. Relatively high density of interrogated sites.	Low resolution (better than RRBS-seq but worse than MBD-seq); some signals may be missed.
PacBio [65,66]	Direct detection of modified DNA bases during normal SMRT sequencing, based on polymerase lags.	Robust detection of 4-methylcytosine, 8-oxoguanine, and N6-methyladenine. Single-base resolution.	Same as BS-seq. 5-methylcytosine, the most common methylation mark in animals, is not reliably detected.
ONT [67]	Direct detection of modified DNA bases during normal nanopore sequencing, based on conductivity changes.	Detects all marks, including 5-methylcytosine. Single-base resolution.	Same as BS-seq.

333

334 Among many covalent chromatin modifications I will discuss DNA methylation since it currently
335 receives the most attention in OMOs. Still it must be mentioned that in plants histone
336 methylation appears be no less and perhaps even more involved in acclimatization and
337 transgenerational plasticity [35].
338
339 While vertebrates show high methylation throughout the genome, invertebrates and plants
340 methylate their genomes sparsely and mostly in protein-coding regions (so-called gene body
341 methylation, GBM, [36]). The function of his ubiquitous and evolutionarily ancient DNA
342 modification remains unclear [37,38] and the greatest challenge in the next few years will be to
343 decipher it. The most important questions are: (i) Does GBM affect gene expression? (ii) Can it
344 be modified on ecological timescale, to achieve acclimatization to a novel environment? (iii) Can
345 acquired changes in GBM be transmitted across generations? If the answers to all three questions
346 are "yes", then we have a mechanism for transgenerational inheritance of acquired traits, which is
347 an exciting (albeit tentative, [39]) possibility. Table 3 summarizes the methods for generating
348 DNA methylation data. For population-scale studies of OMOs MBD-seq and meDIP provide the
349 best resolution for sequencing effort [37], while methylRAD trades off some of the resolution for
350 extreme cost efficiency and throughput. For studies requiring base-level resolution, the best
351 approach appears to be direct detection by PacBio or ONT – however, these exciting
352 developments still require validation in complex genomes.
353
354 **Generating a reference sequence**
355
356 For all approaches described here, the accuracy of the reference sequence in terms of per-base
357 error rate must only be high enough to allow unambiguous mapping of high-accuracy (Illumina)
358 reads. The gold standard of genome sequence quality, Q30 or 99.9% accuracy, would not provide
359 any benefit compared to a rough draft accuracy of 99%. Occasional errors in the reference would
360 manifest themselves as SNPs that are not polymorphic in the analyzed samples and therefore
361 irrelevant for analysis. This is the same reason why it is possible to use a genome of a related
362 species as a reference.
363
364 For AFS analysis, which does not require highly contiguous reference, even a rough genome draft
365 that can be assembled from a single lane worth of 150b paired-end reads from Illumina HiSeq
366 would be suitable. However, substantially better options are now becoming available for a
367 comparably low price tag. The technology offered by 10x Genomics [40] attaches specific
368 barcodes to short reads originating from the same long DNA fragment, which allows assembling
369 Illumina HiSeq data into very long haplotypes. The two single-molecule long-read "third-
370 generation sequencing" methods, Single Molecule Real Time (SMRT) sequencing by PacBio and
371 nanopore sequencing by ONT, produce reads with broad length distribution, including
372 exceedingly long ones (tens to hundreds of kilobases) resulting in a qualitatively more contiguous
373 genome assemblies [41–44] (Table 4, see [42] for recent benchmarking study of assembly
374 pipelines). At the moment of this writing, read accuracy and cost of data for PacBio (Sequel
375 system) and ONT (R9 flow cell) were equivalent; PacBio generated higher proportion of long
376 reads than ONT; however, PacBio's library prep required ten fold more high-quality DNA than
377 ONT. Both for PacBio and ONT it is critically important to obtain high molecular weight DNA in

fragments exceeding 20kb in length. For new OMOs, it is also essential to confirm that the DNA is accessible to enzymatic modifications by trying to digest it with a frequent-cutting restriction endonuclease.

Table 4. Assembly pipelines for PacBio and ONT reads

Pipeline	Required coverage	Features	Pros	Cons
Canu +Quiver* [68]	>30x	Correct and trims reads before assembly.	Best accuracy at base, indel and assembly level.	Very computationally demanding for large genomes. Generates incomplete assemblies at low coverage.
Falcon + Quiver* [69]	>50x	Similar to Canu.	Standard for PacBio.	Very computationally demanding for large genomes. High reliance on reads >20kb. Highly incomplete assemblies at low coverage.
minimap + miniasm + racon [70,71]	<30x	Raw reads are assembled, correction is done post-assembly	Very fast even for large genomes. Works with lower coverage, shorter reads than Canu and Falcon.	The resulting accuracy is lower than with Canu + Quiver.
pilon [72]	NA (error correction method)	Performs additional correction post-assembly.	Boosts accuracy for any assembly.	Requires high-quality Illumina reads.

*Quiver is a consensus polishing software that is now replaced by Arrow to handle PacBio Sequel data (https://github.com/PacificBiosciences/GenomicConsensus). Racon [71] can be used instead of Quiver/Arrow [42].

For genome scanning, gene expression, or invertebrate DNA methylation analyses targeting protein-coding sequences (exome) genome sequence is not be the best reference; instead, a highly contiguous transcriptome assembly would be preferable. Until now the standard way to generate a *de novo* transcriptome was to perform high-coverage RNA-seq and assemble the results with Trinity [45]. In the coming years, it is expected that even higher-quality and lower-cost OMO transcriptomes would be generated by PacBio or ONT sequencing of full-length cDNA (or, for ONT, direct mRNA sequencing). The long-read capacity of these technologies would essentially obviate the need for assembly, leaving only the sequence correction procedure to be performed.

Finally, which tissue or body part to sample for sequencing? For genome sequencing it does not matter much as long as contamination by other DNA sources can be kept to a minimum, but for *de novo* transcriptomics it is not a trivial question, as gene expression varies dramatically across tissues and life cycle stages. In mammals, there is definitely an organ of choice that expresses nearly all genes in the genome: testis. Rather unexpected transcriptome complexity in the testis is putatively due to chromatin re-packaging during spermatogenesis, which results in genome-wide transcription leakage [46]. If so, testis might be a good choice for *de novo* transcriptomics not only for mammals but for any organism that produces compact sperm.

404 **Note on data sharing**
405
406 As we have seen, the best power of ecological genomics in OMOs is achieved using a genome or
407 transcriptome reference. Every new reference dataset enables new biological questions, and the
408 whole OMO field will get a great boost if these resources are promptly shared. Please consider
409 rapidly sharing your reference data, at least as soon as the initial preprint of your paper is posted
410 to bioRxiv and ideally sooner, by distributing the link to data through research-related email list
411 or professional twitter feed.
412
413 **References**

414 1 Excoffier, L. *et al.* (2013) Robust demographic inference from genomic and SNP data.
415 *PLoS Genet.* 9, e1003905
416 2 Gutenkunst, R.N. *et al.* (2009) Inferring the joint demographic history of multiple
417 populations from multidimensional SNP frequency data. *PLoS Genet.* 5, e1000695
418 3 Jouganous, J. *et al.* (2017) Inferring the Joint Demographic History of Multiple
419 Populations: Beyond the Diffusion Approximation. *Genetics* DOI:
420 10.1534/genetics.117.200493
421 4 Andrews, K.R. *et al.* (2016) Harnessing the power of RADseq for ecological and
422 evolutionary genomics. *Nat. Rev. Genet.* 17, 81–92
423 5 Puritz, J.B. *et al.* (2014) Demystifying the RAD fad. *Mol. Ecol.* 23, 5937–5942
424 6 Robinson, J.D. *et al.* (2014) Sampling strategies for frequency spectrum-based population
425 genomic inference. *Bmc Evol. Biol.* 14, 254
426 7 Günther, T. and Coop, G. (2013) Robust identification of local adaptation from allele
427 frequencies. *Genetics* 195, 205–20
428 8 Schiffels, S. *et al.* (2016) Iron Age and Anglo-Saxon genomes from East England reveal
429 British migration history. *Nat. Commun.* 7, 1–9
430 9 Han, E. *et al.* (2014) Characterizing bias in population genetic inferences from low-
431 coverage sequencing data. *Mol. Biol. Evol.* 31, 723–35
432 10 Matz, M. V *et al.* (2017) Potential for rapid genetic adaptation to warming in a Great
433 Barrier Reef coral. *bioRxiv* at <http://www.biorxiv.org/content/early/2017/06/18/114173>
434 11 Dixon, G.B.B. *et al.* (2015) Genomic determinants of coral heat tolerance across latitudes.
435 *Science.* 348, 1460–1462
436 12 Mckenna, A. *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for
437 analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303
438 13 Korneliussen, T. *et al.* (2014) ANGSD: Analysis of Next Generation Sequencing Data.
439 *BMC Bioinformatics* 15, 356
440 14 Nielsen, R. *et al.* (2012) SNP calling, genotype calling, and sample allele frequency
441 estimation from new-generation sequencing data. *PLoS One* 7, e37558
442 15 Andrews, K.R. and Luikart, G. (2014) Recent novel approaches for population genomics
443 data analysis. *Mol. Ecol.* 23, 1661–1667
444 16 Casbon, J.A. *et al.* (2011) A method for counting PCR template molecules with
445 application to next-generation sequencing. *Nucleic Acids Res.* 39, e81
446 17 Gojobori, J. *et al.* (2007) Adaptive evolution in humans revealed by the negative
447 correlation between the polymorphism and fixation phases of evolution. *Proc. Natl. Acad.*
448 *Sci. U. S. A.* 104, 3907–12
449 18 Tine, M. *et al.* (2014) European sea bass genome and its variation provide insights into
450 adaptation to euryhalinity and speciation. *Nat. Commun.* 5, 5770
451 19 Lowry, D.B. *et al.* (2017) Breaking RAD: An evaluation of the utility of restriction site
452 associated DNA sequencing for genome scans of adaptation. *Mol. Ecol. Resour.* 17, 142–

453 152
454 20 Tiffin, P. and Ross-Ibarra, J. (2014) Advances and limits of using population genetics to
455 understand local adaptation. *Trends Ecol. Evol.* 29, 673–680
456 21 McKinney, G.J. *et al.* (2017) RADseq provides unprecedented insights into molecular
457 ecology and evolutionary genetics: comment on Breaking RAD by Lowry *et al* . (2016).
458 *Mol. Ecol. Resour.* 17, 356–361
459 22 Davies, R.W. *et al.* (2016) Rapid genotype imputation from sequence without reference
460 panels. *Nat. Genet.* 48, 965–969
461 23 Money, D. *et al.* (2015) LinkImpute: Fast and Accurate Genotype Imputation for
462 Nonmodel Organisms. *G3 Genes, Genomes, Genet.* 5, 2383–2390
463 24 Morin, R.D. *et al.* (2008) Profiling the HeLa S3 transcriptome using randomly primed
464 cDNA and massively parallel short-read sequencing. *Biotechniques* 45, 81–94
465 25 Meyer, E. *et al.* (2011) Profiling gene expression responses of coral larvae (Acropora
466 millepora) to elevated temperature and settlement inducers using a novel RNA-Seq
467 procedure. *Mol. Ecol.* 20, 3599–3616
468 26 Lohman, B.K. *et al.* (2016) Evaluation of TagSeq, a reliable low-cost alternative for
469 RNAseq. *Mol. Ecol. Resour.* 16, 1315–1321
470 27 Moll, P. *et al.* (2014) QuantSeq 3 ' mRNA sequencing for RNA quantification. *Nat.*
471 *Methods* 11, 25
472 28 Kenkel, C.D. and Matz, M. V (2016) Gene expression plasticity as a mechanism of coral
473 adaptation to a variable environment. *Nat. Ecol. Evol.* 1, 14
474 29 Dixon, G.B. *et al.* (2014) Bimodal signatures of germline methylation are linked with gene
475 expression plasticity in the coral Acropora millepora. *BMC Genomics* 15, 1109
476 30 Crowley, J.J. *et al.* (2015) Analyses of allele-specific gene expression in highly divergent
477 mouse crosses identifies pervasive allelic imbalance. *Nat. Genet.* 47, 353–60
478 31 The GTEx Consortium (2015) The Genotype-Tissue Expression (GTEx) pilot analysis:
479 Multitissue gene regulation in humans. *Science.* 348, 648–660
480 32 Melé, M. *et al.* (2015) The human transcriptome across tissues and individuals. *Science.*
481 348, 660–665
482 33 Gamazon, E.R. *et al.* (2015) A gene-based association method for mapping traits using
483 reference transcriptome data. *Nat. Genet.* 47, 1091–1098
484 34 Wheeler, H.E. *et al.* (2016) Survey of the Heritability and Sparse Architecture of Gene
485 Expression Traits across Human Tissues. *PLOS Genet.* 12, e1006423
486 35 Lämke, J. and Bäurle, I. (2017) Epigenetic and chromatin-based mechanisms in
487 environmental stress adaptation and stress memory in plants. *Genome Biol.* 18, 124
488 36 Feng, S. *et al.* (2010) Conservation and divergence of methylation patterning in plants and
489 animals. *Proc. Natl. Acad. Sci. U. S. A.* 107, 8689–94
490 37 Dixon, G.B.. *et al.* (2016) Evolutionary Consequences of DNA Methylation in a Basal
491 Metazoan. *Mol. Biol. Evol.* 33, msw100
492 38 Sarda, S. *et al.* (2012) The Evolution of Invertebrate Gene Body Methylation. *Mol Biol*
493 *Evol* 29, 1907–1916
494 39 Uller, T. *et al.* (2013) Weak evidence for anticipatory parental effects in plants and
495 animals. *J. Evol. Biol.* 26, 2161–2170
496 40 Kitzman, J.O. (2016) Haplotypes drop by drop. *Nat. Biotechnol.* 34, 296–298
497 41 Gordon, D. *et al.* (2016) Long-read sequence assembly of the gorilla genome. *Science.*
498 352, aae0344
499 42 Michael, T.P. *et al.* (2017) High contiguity Arabidopsis thaliana genome assembly with a
500 single nanopore flow cell. *bioRxiv* at
501 <http://www.biorxiv.org/content/early/2017/06/14/149997>
502 43 Jansen, H.J. *et al.* (2017) Rapid de novo assembly of the European eel genome from
503 nanopore sequencing reads. *bioRxiv* at

504 <http://www.biorxiv.org/content/early/2017/01/20/101907>
505 44 Berlin, K. *et al.* (2015) Assembling large genomes with single-molecule sequencing and
506 locality-sensitive hashing. *Nat. Biotechnol.* 33, 623–630
507 45 Haas, B.J. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using
508 the Trinity platform for reference generation and analysis. *Nat Protoc* 8, 1494–1512
509 46 Soumillon, M. *et al.* (2013) Cellular Source and Mechanisms of High Transcriptome
510 Complexity in the Mammalian Testis. *Cell Rep.* 3, 2179–2190
511 47 De Wit, P. *et al.* (2015) SNP genotyping and population genomics from expressed
512 sequences - current advances and future possibilities. *Mol. Ecol.* 24, 2310–2323
513 48 Bay, R.A. and Palumbi, S.R. (2014) Multilocus adaptation associated with heat resistance
514 in reef-building corals. *Curr. Biol.* 24, 2952–6
515 49 Futschik, A. and Schlötterer, C. (2010) The Next Generation of Molecular Markers From
516 Massively Parallel Sequencing of Pooled DNA Samples. *Genetics* 186, 207–218
517 50 Kofler, R. *et al.* (2011) PoPoolation2: identifying differentiation between populations
518 using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* 27, 3435–3436
519 51 Therkildsen, N.O. and Palumbi, S.R. (2017) Practical low-coverage genomewide
520 sequencing of hundreds of individually barcoded samples for population and evolutionary
521 genomics in nonmodel species. *Mol. Ecol. Resour.* 17, 194–208
522 52 Paradis, E. *et al.* (2004) APE: Analyses of phylogenetics and evolution in R language.
523 *Bioinformatics* 20, 289–290
524 53 Oksanen, J. *et al.* vegan: Community Ecology Package. , *R package ver. 2.4–3*. (2017)
525 54 Wright, R.M. *et al.* (2015) Gene expression associated with white syndromes in a reef
526 building coral, Acropora hyacinthus. *BMC Genomics* 16, 371
527 55 Strader, M.E.M.E. *et al.* (2016) Red fluorescence in coral larvae is associated with a
528 diapause-like state. *Mol. Ecol.* 25, 559–569
529 56 Jombart, T. (2008) adegenet: a R package for the multivariate analysis of genetic markers.
530 *Bioinformatics* 24, 1403–5
531 57 Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation
532 network analysis. *BMC Bioinformatics* 9, 559
533 58 Bay, R.A. and Palumbi, S.R. (2017) Transcriptome predictors of coral survival and growth
534 in a highly variable environment. *Ecol. Evol.* 7, 4794–4803
535 59 Rose, N.H. *et al.* (2016) Gene Networks in the Wild: Identifying Transcriptional Modules
536 that Mediate Coral Resistance to Experimental Heat Stress. *Genome Biol. Evol.* 8, 243–
537 252
538 60 Lister, R. *et al.* (2009) Human DNA methylomes at base resolution show widespread
539 epigenomic differences. *Nature* 462, 315–22
540 61 Meissner, A. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and
541 differentiated cells. *Nature* 454, 766–770
542 62 Serre, D. *et al.* (2010) MBD-isolated Genome Sequencing provides a high-throughput and
543 comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res.* 38,
544 391–9
545 63 Jacinto, F. V. *et al.* Methyl-DNA immunoprecipitation (MeDIP): Hunting down the DNA
546 methylome. , *BioTechniques*, 44. (2008) , 35–43
547 64 Wang, S. *et al.* (2015) MethylRAD: a simple and scalable method for genome-wide DNA
548 methylation profiling using methylation-dependent restriction enzymes. *Open Biol.* 5,
549 150130
550 65 Flusberg, B.A. *et al.* (2010) Direct detection of DNA methylation during single-molecule,
551 real-time sequencing. *Nat. Methods* 7, 461–5
552 66 Feng, Z. *et al.* (2013) Detecting DNA Modifications from SMRT Sequencing Data by
553 Modeling Sequence Context Dependence of Polymerase Kinetic. *PLoS Comput. Biol.* 9,
554 e1002935

555 67 Stoiber, M.H. *et al.* (2017) De novo Identification of DNA Modifications Enabled by
556 Genome-Guided Nanopore Signal Processing. *bioRxiv* at
557 <http://www.biorxiv.org/content/early/2017/04/10/094672>
558 68 Koren, S. *et al.* (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer
559 weighting and repeat separation. *Genome Res*. 27, 722–736
560 69 Chin, C.-S. *et al.* (2016) Phased diploid genome assembly with single-molecule real-time
561 sequencing. *Nat. Methods* 13, 1050–1054
562 70 Li, H. (2016) Minimap and miniasm: Fast mapping and de novo assembly for noisy long
563 sequences. *Bioinformatics* 32, 2103–2110
564 71 Vaser, R. *et al.* (2017) Fast and accurate de novo genome assembly from long uncorrected
565 reads. *Genome Res*. 27, 737–746
566 72 Walker, B.J. *et al.* (2014) Pilon: An integrated tool for comprehensive microbial variant
567 detection and genome assembly improvement. *PLoS One* 9, e112963
568