INTRODUCTION
Genotyping-by-sequencing in ecological and conservation genomics

SHAWN R. NARUM,* C. ALEX BUERKLE,† JOHN W. DAVEY,§ MICHAEL R. MILLER¶ and PAUL A. HOHENLOHE**
*Department of Fisheries and Wildlife, University of Idaho, Moscow, ID, USA, †Columbia River Inter-Tribal Fish Commission, Hagerman Fish Culture Experiment Station, Hagerman, ID, USA, ‡Department of Botany, University of Wyoming, Laramie, WY, USA, §Department of Zoology, University of Cambridge, Cambridge, UK, ¶Department of Animal Science, University of California, Davis, CA, USA, **Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID, USA

Keywords: genotyping-by-sequencing, next-generation sequencing, RAD-seq, SNP

Received 22 March 2013; revision received 9 April 2013; accepted 10 April 2013

The fields of ecological and conservation genetics have developed greatly in recent decades through the use of molecular markers to investigate organisms in their natural habitat and to evaluate the effect of anthropogenic disturbances. However, many of these studies have been limited to narrow regions of the genome, allowing for limited inferences but making it difficult to generalize about the organisms and their evolutionary history. Tremendous advances in sequencing technology over the last decade (i.e. next-generation sequencing; NGS) have led to the ability to sample the genome much more densely and to observe the patterns of genetic variation that result from the full range of evolutionary processes acting across the genome (Allendorf et al. 2010; Stapley et al. 2010; Li et al. 2012). These studies are transforming molecular ecology by making many long-standing questions much more easily accessible in almost any organism.

When studying the genetics of wild populations, it is desirable to samples tens, hundreds or even thousands of individuals. While it is now possible to sequence whole genomes for tens of individuals with small genome sizes, the sequencing of hundreds of individuals with large genomes remains prohibitively expensive, particularly where the genome sequence is unknown. Further, for the purpose of many studies, complete genomic sequence data for all individuals would be unnecessary and simply inflate the computational and bioinformatic costs. A major recent advance has been the development of genotyping-by-sequencing (GBS) approaches that allow a targeted fraction of the genome (a reduced representation library) to be sequenced with next-generation technology rather than the entire genome, even in species with little or no previous genomic information and large genomes. The subset of the genome to be sequenced in these GBS approaches may be targeted using restriction enzymes or capture probes or by sequencing the transcriptome (reviewed in Davey et al. 2011). In the future, as sequencing technology and computational and bioinformatic methods develop further, whole-genome resequencing may become the predominant method for ecological and conservation genomics. Currently, reduced representation approaches offer the ability to not only discover genetic variants such as SNPs but also genotype individuals at these newly discovered loci in the same data.

This special issue on ‘Genotyping-by-Sequencing in Ecological and Conservation Genomics’ represents a diverse set of empirical and theoretical studies that demonstrate both the utility and some of the challenges of GBS in ecological and conservation genomics. The empirical studies include demonstrations of the utility of GBS for population genomics and association mapping, as well as the development of genomic resources (i.e. large SNP data sets) for target species. The studies also illustrate some of the differences between GBS methods, in particular, aligning paired-end reads to achieve longer consensus sequences in contrast to single-end reads with shorter alignments, and double-digest versus sonication methods to fragment DNA. In addition, several papers describe advanced data pipelines for handling GBS-related sequence data and critically evaluate best practices for GBS methods and potential biases and novel features associated with GBS data. Overall, this compilation of papers emphasizes
that GBS has been quickly adopted by the scientific community and is expected to become a common tool for studies in molecular ecology.

**Population genomics**

Genotyping-by-sequencing methods offer major advantages for population genomics by screening thousands of polymorphisms throughout the genome that are subject to the full range of evolutionary histories (variation in drift, selection, recombination, mutation) and consequences for genetic variation. Historically, most studies in ecological and conservation genetics have relied upon a small number of putatively neutral molecular markers (e.g. allozymes, microsatellites, AFLPs), covering a very limited subset of the genome. These data sets could be used to address questions related to demographic factors that affect the entire genome (e.g. diversity, gene flow and drift, effective population sizes and genetic relationships of populations), but they had limited ability to investigate specific loci that have been subject to selection and adaptive evolution. However, GBS enables researchers to identify specific genomic regions that may have experienced natural selection, in addition to improving the precision of demographic inferences by greatly increasing the number of putatively neutral markers assayed. For example, neutral markers alone may not identify distinct populations that have evolved to become resistant to specific pathogens (Bonneaud et al. 2011) or locally adapted to their habitat (Storz et al. 2009; Narum et al. 2010). Conversely, neutral markers may identify significant differentiation among populations based on limited gene flow or drift, but genomic regions under selection may indicate adaptive similarity that may have been either retained after isolation (Parchman et al. 2013) or evolved in parallel following colonization of new habitats (e.g. Hohenlohe et al. 2010).

Several studies in this issue utilize genome scans to search for potentially adaptive genetic variation in a population genomics context as well as estimate demographic parameters (Table 1). Included are various species of plants, marine invertebrates, marine and freshwater fish, and small mammals, making novel inferences regarding selection in natural populations in addition to measuring demographic parameters using neutral markers (Catchen et al. 2013b; Corander et al. 2013; De Wit & Palumbi 2013; Hess et al. 2013; Hyma & Fay 2013; Keller et al. 2013; Reitzel et al. 2013; Roda et al. 2013; White et al. 2013). Multiple papers demonstrate the utility of GBS for phylogenetic reconstruction across species (Jones et al. 2013; Keller et al. 2013; Ogden et al. 2013; Roda et al. 2013). Additionally, three papers take advantage of GBS to identify genomic regions involved in hybridization (Hohenlohe et al. 2013), speciation (Jones et al. 2013) and divergent adaptation (Keller et al. 2013). Another study (Roesti et al. 2013) investigates stickleback populations to reveal how heterogeneous recombination rates can modulate consequences of selection and influence outlier tests for positive selection. Roesti et al. (2013) also use sex-specific RAD locus coverage to scrutinize sex chromosome divergence and confirm the presence of evolutionary strata in this species. All such population genomics studies face similar challenges in navigating trade-offs in sequencing effort across loci, individuals and populations. Accordingly, Buerkle & Gompert (2013) consider the question of optimizing allocation of sequencing effort in GBS between depth of coverage per locus and larger sample sizes, in order to most effectively use sequence data for population genetics.

**Genome-wide association and QTL mapping studies**

Screening dense markers from the genome has effectively enabled discovery of many candidate loci involved in specific phenotypic traits, either with quantitative trait loci (QTL) mapping or with genome-wide association studies (GWAS). In the last decade, these approaches have been utilized extensively in humans to identify specific genes and pathways involved human health (Hindorff et al. 2009) and to discover disease alleles in model organisms (Flint & Eskin 2012). As GBS does not require previous genomic information, high-density QTL mapping and GWAS studies are now being incorporated to investigate phenotypes related to biological traits in many nonmodel species in natural environments (e.g. Parchman et al. 2012). In this issue, Gagnaire et al. (2013) use RAD-seq to map phenotypic and expression QTL for ecologically relevant traits in lake whitefish (Coregonus clupeaformis). Additionally, RAD-seq was used in GWAS to identify regions of the genome associated with traits such as colour dimorphism in species of cichlid fishes (Takahashi et al. 2013), binary migration patterns in a salmonid fish (Hecht et al. 2013), phenotypic shell variation of land snails (Cepaea nemoralis; Richards et al. 2013) and thermal adaptation of ectothermic fish in desert streams (Narum et al. 2013). These studies illustrate the potential for mapping biologically relevant traits in wild populations to provide novel insight into ecological processes and to facilitate monitoring of species at risk to extinction.

**Genomic resources – SNP discovery**

Development of genomic resources has long been a need in the field of molecular ecology, and NGS
Table 1 Data generated for contributions to this special issue using reduced representation GBS methods

<table>
<thead>
<tr>
<th>Study</th>
<th>Organism</th>
<th>Method</th>
<th># loci analysed</th>
<th># samples</th>
<th># groups</th>
<th>Study goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catchen et al.</td>
<td>Threespine stickleback (Gasterosteus aculeatus)</td>
<td>Single-end RAD-seq*</td>
<td>25 679</td>
<td>578 inds</td>
<td>9 pops</td>
<td>Phylogeography</td>
</tr>
<tr>
<td>Corrander et al.</td>
<td>Herring (Clupea harengus)</td>
<td>Single-end RAD-seq*</td>
<td>5 985</td>
<td>2 pools</td>
<td>2 pops</td>
<td>Population differentiation</td>
</tr>
<tr>
<td>Davey et al.</td>
<td>Caenorhabditis elegans</td>
<td>Paired-end RAD-seq*</td>
<td>24 828</td>
<td>24 pools</td>
<td>1 laboratory strain</td>
<td>Quantification of technical bias</td>
</tr>
<tr>
<td>DeWit &amp; Palumbi</td>
<td>Red abalone (Haliotis rufescens)</td>
<td>Transcriptome sequencing</td>
<td>21 579</td>
<td>39 inds</td>
<td>3 pops</td>
<td>Population structure; identification of outlier loci</td>
</tr>
<tr>
<td>Gagnaire et al.</td>
<td>Lake whitefish (Coregonus clupeaformis)</td>
<td>Single-end RAD-seq*</td>
<td>3438</td>
<td>102 inds</td>
<td>1 hybrid backcross family</td>
<td>QTL mapping</td>
</tr>
<tr>
<td>Hecht et al.</td>
<td>Rainbow/steelhead trout (Oncorhynchus mykiss)</td>
<td>Single-end RAD-seq*</td>
<td>12 073</td>
<td>189 inds</td>
<td>2 pops</td>
<td>Genome-wide association mapping</td>
</tr>
<tr>
<td>Hess et al.</td>
<td>Pacific lamprey (Entosphenus tridentatus)</td>
<td>Single-end RAD-seq*</td>
<td>4439</td>
<td>518 inds</td>
<td>21 pops</td>
<td>Phylogeography; identification of outlier loci</td>
</tr>
<tr>
<td>Hohenlohe et al.</td>
<td>Westslope cutthroat trout (Oncorhynchus clarkii lewisii)</td>
<td>Paired-end RAD-seq*</td>
<td>77 141</td>
<td>97 inds</td>
<td>5 pops</td>
<td>Estimation of admixture</td>
</tr>
<tr>
<td>Hyma &amp; Fay</td>
<td>Yeast (Saccharomyces cerevisiae &amp; S. paradoxus)</td>
<td>Single-end RAD-seq*</td>
<td>5425 + (S.c.); 9809</td>
<td>77 inds</td>
<td>8 pops</td>
<td>Population structure; identification of outlier loci</td>
</tr>
<tr>
<td>Jones et al.</td>
<td>Swordtail fish (Xiphophorus spp.)</td>
<td>Single-end double-digest RAD-seq†</td>
<td>149 362</td>
<td>139</td>
<td>26 species</td>
<td>Phylogenetic reconstruction</td>
</tr>
<tr>
<td>Keller et al.</td>
<td>Cichlid fish (Pundamilia spp. &amp; Mbipia spp.)</td>
<td>Single-end RAD-seq*</td>
<td>10 663</td>
<td>50 inds</td>
<td>5 species</td>
<td>Population structure; phylogenetic reconstruction; identification of outlier loci</td>
</tr>
<tr>
<td>Narum et al.</td>
<td>Redband trout (Oncorhynchus mykiss gairdneri)</td>
<td>Single-end RAD-seq*</td>
<td>10 685</td>
<td>774 inds</td>
<td>2 pops + 1 F1 family</td>
<td>Association mapping; SNP discovery; population structure</td>
</tr>
<tr>
<td>Ogden et al.</td>
<td>Sturgeon (Acipenser spp.)</td>
<td>Paired-end RAD-seq*</td>
<td>48 731</td>
<td>4 pools + 8 inds</td>
<td>4 species from 6 sites</td>
<td></td>
</tr>
<tr>
<td>Reitzel et al.</td>
<td>Sea anemone (Nematostella vectensis)</td>
<td>Single-end RAD-seq*</td>
<td>4065</td>
<td>30 inds</td>
<td>4 pops</td>
<td>Phylogeography; identification of outlier loci</td>
</tr>
<tr>
<td>Richards et al.</td>
<td>Land snail (Cepaea nemoralis)</td>
<td>Single-end RAD-seq*</td>
<td>57 750</td>
<td>26 inds</td>
<td>1 laboratory cross</td>
<td>Mapping of recombination rate; sex chromosome evolution</td>
</tr>
<tr>
<td>Roda et al.</td>
<td>Groundsel (Senecio spp.)</td>
<td>Single-end RAD-seq*</td>
<td>29 307</td>
<td>29 pools</td>
<td>29 pops</td>
<td>Phylogenetic reconstruction; identification of outlier loci</td>
</tr>
<tr>
<td>Roesti et al.</td>
<td>Threespine stickleback (Gasterosteus aculeatus)</td>
<td>Single-end RAD-seq*</td>
<td>1872</td>
<td>282 inds</td>
<td>1 F2 cross</td>
<td>Mapping of recombination rate; sex chromosome evolution</td>
</tr>
<tr>
<td>Senn et al.</td>
<td>Eurasian beaver (Castor fiber)</td>
<td>Paired-end RAD-seq*</td>
<td>30 201</td>
<td>10 inds</td>
<td>3</td>
<td>SNP discovery</td>
</tr>
<tr>
<td>Takahashi et al.</td>
<td>Cichlid fish (Cyprichromis leptosoma)</td>
<td>Single-end RAD-seq*</td>
<td>11 123</td>
<td>14 + 78 inds</td>
<td>F2 cross + 1 wild population</td>
<td></td>
</tr>
</tbody>
</table>

© 2013 John Wiley & Sons Ltd
approaches have greatly enhanced the discovery of SNPs for many nonmodel organisms (e.g. Seeb et al. 2011). In particular, GBS has become a highly reliable approach for identifying SNPs both within and between populations (e.g. Hohenlohe et al. 2011). All 21 of the empirical studies in this issue provide new SNP resources for several species, highlighting the strengths of GBS approaches for providing new polymorphisms. While GBS is clearly powerful in diploid species, two papers in this issue describe attempts to identify SNPs in polyploid species of birch (Betula spp.; Wang et al. 2013) and four species of tetraploid sturgeon (Ogden et al. 2013). While SNP discovery was well demonstrated in both studies, challenges remain for calling SNP genotypes for individual organisms because polyploids may have multiple copies of different alleles. Thus, further advances in SNP genotyping algorithms (e.g. Serang et al. 2012) are needed in order for GBS approaches to be applied for this purpose in polyploids.

Software pipelines

As next-generation sequencers can currently produce tens to hundreds of gigabases of sequence data per run (see Glenn 2011 with a recent update at http://www.molecularecologist.com/next-gen-fieldguide-2013), advanced analysis pipelines have become a necessity to filter, sort and align sequence data. A pipeline for GBS must include steps to filter out poor-quality reads, classify reads by pool or individuals based on sequence barcodes, either identify loci and alleles de novo or align reads to an index to discover polymorphisms, and often score genotypes for each individual included in the study. The most comprehensive pipeline for handling GBS data is Stacks (Catchen et al. 2011), and in this issue, Catchen et al. (2013a) describe new features in Stacks to calculate population genomic statistics (such as \( F_{ST} \) and nucleotide diversity), create smoothed distributions using sliding window averaging across the genome and produce output genotype files specifically formatted for commonly used downstream analysis packages. Senn et al. (2013) describe an extension to the Stacks pipeline, using the assembly program Cortex to assemble paired-end reads at RAD loci and call SNPs in the assembled contigs. Tools for this paired-end assembly step are also explored by Davey et al. (2013) and Hohenlohe et al. (2013). These pipelines provide bioinformatics solutions for GBS studies and are broadly applicable to many species.

Addressing biases of genotyping-by-sequencing

Genotyping-by-sequencing methods using restriction enzymes (Miller et al. 2007; Baird et al. 2008; van Orsouw et al. 2007; Andolfatto et al. 2011; Elshire et al. 2011; Peterson et al. 2012; Parchman et al. 2012) can produce data with unique characteristics, resulting from factors such as restriction-site polymorphism or correlations of restriction fragment length with read depth. These features of GBS data and the genotyping biases they can produce are reviewed in detail by Davey et al. (2013), while Gautier et al. (2013) and Arnold et al. (2013) focus on the impact of restriction-site polymorphisms on population genetics estimates. Gautier et al. (2013) consider the effect of allele dropout on genotyping and \( F_{ST} \) calculations using both individuals and pools. Arnold et al. (2013) evaluate several additional population genetics statistics, demonstrate that the choice of restriction enzyme and allele dropout can have substantial effects on these estimates, and assess the double-digest RAD-seq method (Peterson et al. 2012) as well as standard RAD-seq. The test of double-digest RAD-seq is particularly useful as this approach should in theory avoid or reduce the bias of fragment length coverage, but Arnold et al. (2013) find that the effects of restriction-site polymorphism on summary statistics are more pronounced with the double-digest method.
All three papers make basic recommendations for data filtering to mitigate the most serious effects of GBS biases, while proposing more sophisticated statistical techniques for identifying and correcting biased genotypes. However, the extensive work of developing these techniques and making them sufficiently general to be applied to a wide range of species and methods remains to be done. Of the empirical papers in this special issue, all apply some type of filter to remove loci with missing genotypes to address the problem of null alleles and other potential biases identified here. While filtering out poor loci is the most common suggestion to address these biases, there are not universal filter criteria that can be applied to all studies, and thus, each of these areas must be evaluated by investigators on a case-by-case basis. As a general guideline for future analyses of GBS data sets, all empirical studies should strive to demonstrate how these potential biases were addressed.

Future needs

While the papers in this issue demonstrate the strength of GBS in ecological and conservation genomics studies, they also highlight areas where further advances are needed. This includes more advanced methods to test for and correct biases associated with GBS, new methods to confront evolutionary theory with population genomic data, additional analytical tools for associating genomic variation with evolutionary processes and histories, and new approaches for visualizing vast amounts of genomic data. These areas are expected to provide better conceptual understanding of selection on organisms in their natural ecosystems, along with improved knowledge of the underlying genetic basis for specific traits related to biological processes. This knowledge will also be utilized to design effective strategies for conserving functional genetic variation to allow for future evolution. The summary information provided in Table 1 also provides a useful context to compare results of different GBS methods.

In addition to advances in theory and analytical tools for genomic data, new technical variations of GBS are expected in the near future that include complete genome typing for individuals and genotyping large numbers of individuals at selected targets that are considered to be biologically relevant. Also, the potential to combine RNA-seq and GBS approaches to identify SNPs in the transcriptome associated with patterns of gene expression offers the potential to strengthen links between genomics, transcriptomics and proteomics. Indeed, GBS has greatly expanded research opportunities in ecological and conservation genomics, and further advances are expected to open nearly endless doors of study to advance our knowledge.

Acknowledgements

We thank Tim Vines, Loren Rieseberg, Jen Gow, Hilary McNaughton and Adrienne Nye for their support during the preparation of this issue.

References

Andolfatto PD, Davison D, Erezylmaz TT et al. (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping, Genome Research, 21, 610–617.


Miller MR, Atwood TS, Eames BF et al. (2007) RAD marker microarrays enable rapid mapping of zebrafish mutations. Genome Biology, 8, R105.


All authors are actively involved in the field of ecological and conservation genomics and have developed expertise in genotyping-by-sequencing methods for studies in this area. The authors were co-editors of the special issue on this subject.