MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 2184–2195 doi: 10.1111/j.1365-294X.2010.04650.x

INVITED REVIEW

Genome-enabled development of DNA markers for ecology, evolution and conservation

ROBERT C. THOMSON, IAN J. WANG and JARRETT R. JOHNSON

Department of Evolution and Ecology and Center for Population Biology, University of California, Davis, CA 95616, USA

Abstract

Molecular markers have become a fundamental piece of modern biology's toolkit. In the last decade, new genomic resources from model organisms and advances in DNA sequencing technology have altered the way that these tools are developed, alleviating the marker limitation that researchers previously faced and opening new areas of research for studies of non-model organisms. This availability of markers is directly responsible for advances in several areas of research, including fine-scaled estimation of population structure and demography, the inference of species phylogenies, and the examination of detailed selective pressures in non-model organisms. This review summarizes methods for the development of large numbers of DNA markers in non-model organisms, the challenges encountered when utilizing different methods, and new research applications resulting from these advances.

Keywords: ANM, EPIC, non-model organism, NPCL, nuclear DNA, primer design, sequencing

Received 26 June 2009; revision received 25 March 2010; accepted date 30 March 2010

Introduction

Advances in genomic biology and the increasing availability of genomic resources have altered research on non-model organisms in several fundamental ways. Most prominently, these changes have made collecting large DNA sequence datasets far more feasible, allowing for more detailed analyses of complex biological processes. The massive influx of genomic data for model organisms has driven development of new informatic methods for organizing and utilizing large datasets and new analytical methods that provide far more power to dissect biological processes in detail than was previously possible (Hey & Machado 2003; Manel *et al*. 2003; Rannala & Yang 2008). Although usually designed for model organisms, these analytical methods are perhaps most useful in non-model systems, in which many evolutionary and ecological processes are studied in natural field systems. All of these new analytical methods require large volumes of data, which has created interest in finding ways to gather these data for wild species.

The diverse questions to which these datasets are applied has led to the publication of marker development methods in a variety of journals, with topics ranging from applied genetics to ecology to phylogenetics, resulting in a fractured literature spread across many disciplines. In this review, we first summarize the types of commonly used markers and provide a guide for researchers planning to develop their own novel marker resources. We then argue that obtaining cost- and time-effective many-marker datasets in nearly any organism is now possible, opening a new class of methods available for research on wild species. Finally, we discuss the new directions that emerging sequencing technologies may lead marker development in the near future. We primarily focus our discussion on DNA sequence markers, as these are the most flexible and widely used. However, many of the methods outlined here are easily adapted for the design of alternative marker types.

Choosing a marker development strategy

The choice of strategy for any marker development project begins by balancing the need for certain marker characteristics (number, variability, type) with the kinds of resources (genetic resources, time, funding) necessary for their development (Box 1, Table 1). Following this decision, marker development itself can be fairly

Correspondence: Robert C. Thomson, Fax: +1 530 752 1449; E-mail: rcthomson@ucdavis.edu

Box 1 Marker types

Nuclear Protein Coding Loci—For applications that require highly conserved sequences, such as deep phylogenetics, nuclear protein coding loci (NPCL) are often the markers of choice (Figs 1 and 2). These markers are functionally constrained and, because they are located in coding regions, generally have the most complete annotation information available. Such functionally constrained regions tend to have low incidences of gene gains and losses as well as low nucleotide divergence, making them easy to align over large phylogenetic distances (Townsend *et al.* 2008). The high level of annotation makes identifying and removing paralogous gene copies and repetitive elements easier than with other marker types. Though these advantages must be balanced with the knowledge that many of these markers may be under strong selection, the growing availability of fully sequenced genomes makes selecting single-copy genes, designing primers, and testing primers in the clade of interest relatively straightforward.

Exon-primed Intron-crossing (EPIC) Markers—One of the central challenges in marker design is finding markers that are variable enough to be highly informative but conserved enough that primer sites do not accumulate substitutions across the phylogenetic span of interest. A solution to this problem is an approach that targets variable sequence regions (introns) flanked by conserved regions (exons) that contain the priming sites. In the genetic mapping literature, markers resulting from this approach have been referred to as comparative anchor tagged sequences (CATS, O'Brien *et al.* 1993; Lyons *et al.* 1997), traced orthologous amplified sequence tags (TOASTs, Jiang *et al.* 1998), and sequence-tagged sites (STS, Venta *et al.* 1996; Perry & Bousquet 1998). These markers have also been referred to as exon-primed intron-crossing sequences (EPICs, Palumbi & Baker 1994) in the phylogenetics and population genetics literature. The 'EPIC' acronym strikes us as the most descriptive of the actual marker type, and so we use this term.

Anonymous Nuclear Markers—Anonymous nuclear markers (ANMs) require the least prior information of any marker class. Because designing ANMs generally consists of using random draws from the genome and most of the genome is non-coding, most ANMs fall into non-coding genomic regions. A benefit of the ANM strategy is that non-coding regions of the genome generally have a high substitution rate, so these markers often contain substantial variation, making them informative for analyses at shallow levels of divergence (Fig. 1). However, a significant downside of this marker type is anonymity. Because of the lack of annotation, the prevalence of repetitive elements present in many genomes, and our poor understanding of the functional role of non-coding DNA, paralogy, and copy-number problems remain a major concern. In many cases, though, genomic resources in non-model systems remain very limited and ANMs may present a feasible option to researchers in need of large numbers of highly variable loci.

straightforward. Here, we attempt to guide researchers through the basic steps involved in designing a marker development strategy and discuss other factors that may need to be tailored to individual studies. Although the characteristics of markers developed in these studies vary in several ways, a few main classes of markers have emerged: markers from protein coding regions of the genome (NPCLs), markers from introns that are primed from flanking exons (EPICs), and anonymous nuclear markers (ANMs; Box 1, Box 2, Table 1). All of these marker classes have been successfully developed in non-model organisms to address a wide range of questions, which makes choosing the optimal method difficult when designing a new project. Here, we discuss marker development strategies and provide recommendations for the ideal types of markers for several fields that molecular ecologists are commonly interested in: phylogenetics, phylogeography, population genetics, and mapping genes of ecological or evolutionary interest (Fig. 2).

Table 1 Common factors to consider when deciding among marker design strategies

			Properties of methods			Properties of markers			
Marker type	Resources used	Example reference	Success rate	Technical difficulty	Resources required	Variability	Likelihood of paralogs	No. potential loci	Phylo-genetic span
NPCL	ESTs and a Genome	Townsend *et al.* 2008	High	Low	Many	Low	Low	High	large
NPCL	Two genomes	Li *et al.* 2007	High	Low	Many	Low	Low	High	large
NPCL	Two EST resources	Putta *et al.* 2004	High	Low	Many	Low	Medium	Medium	large
EPIC	ESTs and genomes	Backström *et al.* 2008a	High	Low	Many	Medium	Low	High	large
EPIC	cDNA library	Whittall *et al.* 2006	High	Medium	None	Medium	Medium	Low	med
ANMs	Small-insert library	Jennings & Edwards 2005	Medium	Medium	None	High	High	Low	low
ANMs	BAC end sequences	Thomson *et al.* 2008	Medium	Low	Few	High	High	Medium	low
ANMs	AFLPs	Brugmans *et al.* 2006	Low	High	None	High	High	Low	low

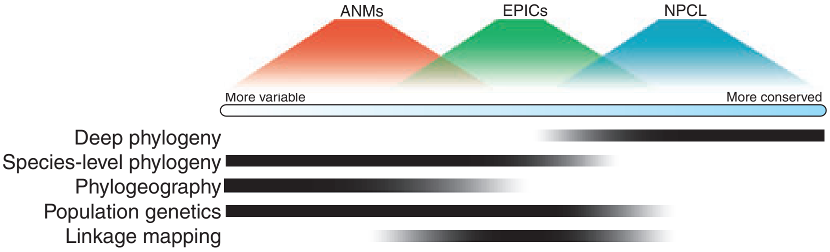

Fig. 1 Schematic of relative variability among marker classes and the amount of variation generally required for different types of research questions.

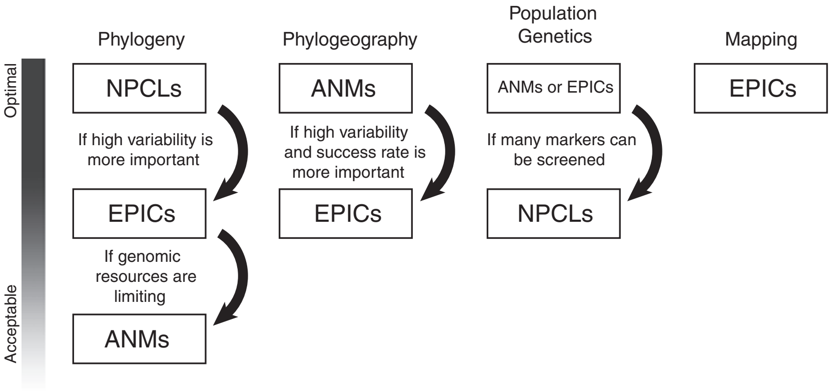
Phylogeny
Phylogeography
Population Genetics
Mapping
Optimal
Acceptable
NPCLs
If high variability is more important
EPICs
If genomic resources are limiting
ANMs
ANMs
If high variability and success rate is more important
EPICs
ANMs or EPICs
If many markers can be screened
NPCLs
EPICs

Fig. 2 Overview of appropriate marker classes for different questions. We outline some of the common issues that affect the choice of marker class, though see text for further discussion.

Box 2 Glossary

ANM—(anonymous nuclear marker) markers that sit in *a priori* unannotated regions of the genome
CATS—(comparative anchor-tagged sequences) synonymous with EPIC markers
EPIC—(exon-primed intron-crossing) markers that have priming sites in conserved exons, but span less-conserved introns
Nested Primer—a primer set that is designed to sit within the amplification product of another primer set. Generally used to perform a second round of amplification using the products of an initial amplification as template.
NPCL—(nuclear protein coding loci) markers that amplify coding regions of genes
Reference Species—the taxon from which markers are designed (generally using available genomic resources in that taxon)
STS—(sequence tagged sites) synonymous with EPIC markers
Test Species—the taxon in which primers designed from a reference taxon are tested for amplification and variability
TOAST—(traced orthologous amplified sequence tags) synonymous with EPIC markers
Universal Primer—a primer designed to anneal to a highly conserved sequence and that will cross-amplify in a wide range of taxa

Phylogenetics

Phylogeneticists have long recognized the problems associated with inference of phylogeny from single gene trees, but only recently has the availability of genomic resources brought massively multilocus datasets to phylogenetics (Edwards 2009). For most interspecific phylogenetics, NPCLs are likely the markers of choice (Murphy *et al.* 2001; Li *et al.* 2007; Rowe *et al.* 2008), because they provide an appropriate level of variation, easy alignment across large phylogenetic distances, and relatively straightforward detection of paralogs (see below: 'Assessing Homology'). One approach to developing NPCLs is outlined by Townsend *et al.* (2008), which used the *Homo* protein database (derived from the human genome) and the pufferfish (*Takifugu*) protein database (derived from pufferfish EST sequences) for reference sequences to design 26 NPCL that amplify across a wide range of vertebrates. After first identifying orthologous genes in human and pufferfish using BLAST searches and filtering sequences that were too small or had too high of a mutation rate, the authors located the chicken (*Gallus*) homologs of their candidate markers using a second round of BLAST searches. Based on comparisons of all available amniote sequences for each marker, Townsend *et al.* (2008) designed primers to amplify regions falling within a single exon and that contained degenerate sites to accommodate the variation observed across the available reference sequences. This approach resulted in the identification of 26 markers that could be used to amplify and sequence the targeted gene region in a set of 10 squamate test taxa and several additional vertebrates.

Li *et al.* (2007) employed a similar approach in ray-finned fishes, using the pufferfish and zebrafish (*Danio*) genomes as comparative data for primer design. In addition, they used a nested PCR design to increase amplification specificity. The utility of employing nested PCR should be balanced against considerations for how many candidate markers are available and the amount of data required from each individual marker. For example, emerging methods for estimation of species phylogenies from multiple gene genealogies can require relatively long DNA reads from each marker to maximize the phylogenetic information available for inferring individual genealogies with confidence (Edwards 2009). Nesting a second set of primers within the first decreases the size of the sequenced region, which may be unacceptable for these applications. So, if the number of potential markers is not limiting, discarding those markers that require nested PCR and designing and screening more markers may be a better approach.

NPCLs are typically more conserved than other types of markers, but the variability of loci can be influenced by the marker design strategy. Townsend *et al.* (2008) observed marked variation in the level of conservation of different regions of single exons and selected specific regions that appeared to be most variable for their markers. Similarly, Li *et al.* (2007) also considered variation in exons but compared entire exons (instead of exon regions) and did not require that their markers would actually span the most variable regions within an exon. The Townsend *et al.* (2008) method targeted more variable regions overall; the human–chicken average amino acid similarity for their markers was 72%, compared to an average of 93% in Li *et al.* (2007). Thus, the marker design process can have a strong impact on the amount of variation present in the final marker pool and the phylogenetic breadth across which markers will amplify. Finally, the number of loci required for robust multi-locus phylogenetic inference should be considered when developing markers for a particular system or study. Species tree methods are still in their infancy, and few studies exist on how many markers are required to estimate a species phylogeny with confidence. However, an estimated 20 or more loci appear to be required to resolve phylogenetic incongruence using traditional concatenation approaches, based on both empirical and simulation data (Rokas *et al.* 2003; Spinks *et al.* 2009). Encouragingly, these estimates fall well within the range of the number of markers that the above development strategies have produced.

The amount of annotation information available from genomes makes high throughput NPCL approaches simple and practical for clades in which the appropriate resources exist. Although the number of fully sequenced and annotated genomes is still relatively small (Table 2), the increasing availability of genomic resources means that the number of clades in which these marker development approaches will be effective should increase substantially over the next few years.

Taxonomic group	Number of species	EST sequence reads	Genome projects		
			Complete	Draft assembly	In progress
Animals	1 162 900	31 426 113	4	68	62
Mammals	5400	18 553 673	2	27	19
Birds	10 000	716 723	0	2	1
Amphibians	6300	2 011 357	0	0	2
Reptiles	8200	192 405	0	1	1
Fishes	28 000	4 846 571	0	10	5
Insects	1 000 000	3 571 148	1	20	19
Flatworms	25 000	458 077	0	1	3
Roundworms	80 000	1 076 159	1	7	12
Plants	357 000	19 549 114	2	8	43
Land plants	350 000	19 130 159	2	6	36
Green algae	7000	418 955	0	2	7
Fungi	70 000	2 030 118	10	67	40
Ascomycetes	30 000	1 472 719	8	54	27
Basidiomycetes	20 000	437 908	1	10	7
Other fungi	20 000	119 491	1	3	6
Protists	135 000	872 587	6	24	24
Apicomplexans	4500	454 256	1	11	5
Kinetoplasts	2000	70 254	1	4	3
Other protists	128 500	348 077	4	9	16

Table 2 Availability of genomic resources for major taxonomic groups. For each taxonomic group, we list the number of species it contains, the number of EST sequence reads available on GenBank, the progress of completed and ongoing genome projects

Phylogeography

For phylogeographic studies and phylogenetic analysis of rapid radiations, finding nuclear markers with sufficient variation remains a major challenge (Hare 2001; Brito & Edwards 2009). Currently, EPICs are the most widely used nuclear sequence marker for these studies, however ANMs may actually be more desirable. ANMs typically require fewer resources to develop and contain greater variability than EPICs. For example, SNP frequencies of 1 SNP per 130 bp (Backström *et al.* 2008a) to 400 bp (Aitken *et al.* 2004) have been reported from studies developing EPICs, while estimates in ANM studies are as high as 1 SNP per 17 bp (Jennings & Edwards 2005). Additionally, EPICs are more likely to experience purifying selection via hitchhiking, as they are situated near gene regions that could be under selection, while ANMs typically are not.

A straightforward approach to developing ANMs that requires few starting resources is to create a small insert library from sheared genomic DNA and sequence clone inserts from the library. Rosenblum *et al.* (2007) used this approach to design ANM primer pairs for a phylogeographic/population genetic study in the eastern fence lizard. After sequencing 192 clones and designing primers for 77 inserts, 50 primer pairs amplified PCR products in the target species, and 19 of these had suitable variation for the intended study. These markers contained an average of 3.8 SNPs per 100 bp in a sample of 91 lizards from one geographically restricted area. Even higher levels of variation in ANMs were identified by Lee & Edwards (2008) for an Australian bird from a small insert library. In a comparison of nucleotide diversity (π) across 29 ANMs and 6 EPICs, Lee & Edwards (2008) demonstrated that ANMs (π = 0.016) were significantly more variable than introns (π = 0.009). Thus, this approach provides an effective way to generate a moderate number of highly variable markers in systems that have few or no existing genomic resources.

It is still possible to develop ANMs even when researchers prefer to avoid the added time and expense of creating a small insert library. AFLPs have become one of the standard tools of molecular biology in uncharacterized genomes. Although as dominant markers their utility is debated, several methods have emerged for converting AFLPs into sequence markers (Bradeen & Simon 1998; Shan *et al.* 1999; Meksem 2001; Brugmans *et al.* 2003). While it is possible to simply select AFLP bands, design internal primers, and screen for variation within the new amplification product, this approach is inefficient. AFLP bands are short (typically <500 bp); so the chance of observing polymorphisms within an even shorter amplification product is relatively low. More importantly, nesting primers within the AFLP excludes the variable site that must be present at one of the AFLP's ends (i.e., the site that caused the original amplification product to be identified as an AFLP in the first place). Brugmans *et al.* (2003) described a method for the conversion of AFLP bands into single locus markers that allows for this variable site to be captured. This approach relies on excising and directly sequencing AFLP bands, followed by the design of internal primers and a series of nested and semi-nested PCR amplifications to locate the end of the AFLP containing the variable site. Once the variable site is located, the method uses a PCR-based approach for obtaining the AFLP's flanking sequence (Siebert *et al.* 1995). Because it uses AFLPs as reference data, this method *only* targets variable markers, making the process efficient for locating SNPs between very closely related species.

The primary drawback of ANMs is that they fall in non-coding regions of the genome and a very large fraction of the non-coding regions of most genomes are composed of repetitive elements, such as SINEs, LINEs, and other retroelements. Thus, detecting and avoiding repetitive elements during marker development takes on an even greater importance (see below 'Assessing Homology'). Nevertheless, because of their variability and relative easy development, ANMs provide an excellent class of markers for phylogeographic studies.

Population genetics

The availability of large numbers of markers that span the genomes of non-model organisms has moved the study of population genetics into the realm of population genomics. Fundamentally, population genomics aims to distinguish locus-specific effects—those that might generate variation in only a few loci—from genome-wide effects, such as demographic history, inbreeding, and population structure (Black *et al.* 2001; Luikart *et al.* 2003; Stinchcombe & Hoekstra 2007). Marker development for multi-locus population genetics or population genomics can, in practice, span all three types of markers described above. Researchers looking for SNPs are likely to find sufficient variability in any marker class for their purposes and will be more limited by the total number of markers that they can develop and the ability of different approaches to target single copy markers (see below 'Assessing Homology'). Even researchers focusing on highly conserved NPCLs have found high enough frequencies of SNPs to be useful for population genetic analyses. For example, Putta *et al.* (2004) used EST resources from two closely related emerging model species of ambystomatid salamanders (*Ambystoma tigrinum* and *A. mexicanum*) to design PCR primers for a third, non-model ambystomatid, *A. ordinarium*. Putta *et al.* (2004) selected 123 ESTs that were variable between the two reference taxa, designed PCR primers and tested them in a pooled sample of 10 *A. ordinarium* populations. The authors found that 79% of the markers yielded amplification products of the expected size and that approximately half of the markers contained at least one SNP.

In the case of Putta *et al.* (2004), EST library resources for two taxa closely related to the target species allowed for a large number of potential markers to be screened, and therefore the reduced variability of NPCLs was not problematic for the study. When screening such a large numbers of markers is not possible, we recommend that researchers develop EPIC or ANM markers instead, because of an expected increase in variability. A recent study by Backström *et al.* (2008a) shows the power of this approach for deriving markers that are useful across a broad spectrum of species. By comparing the zebra finch (*Taeniopygia*) genome with the chicken genome, Backström *et al.* (2008a) identified a set of 242 EPIC markers spread evenly across the avian genome. These markers have clear utility for genetic mapping of wild avian genomes (see below) but also appear highly variable and useful for population genetic and phylogenetic studies. Backström *et al.* (2008a) sequenced 200 EPIC markers in a series of 10 unrelated collared flycatchers (*Ficedula albicollis*, an exemplar 'wild species' used for several of their tests) and found, on average, 1 SNP for every 130 bp of intron sequence. The authors also attempted to amplify a subset of their markers ($N = 122$) in a panel of five bird species. The proportion of markers that were successfully amplified ranged from 93% in chicken to 34% in Tengmalm's owl (*Aegolius funereus*; overall mean = 73%).

When well-annotated genomic resources are not available, it is still possible to design EPIC markers by utilizing properties of the genome that are known. Whittall *et al.* (2006) sequenced inserts from a cDNA library and designed primers for sequences that exhibited high similarity to known proteins, placing primers near the 3′-end of the coding region and in the 3′-UTR (untranscribed region). This approach may be particularly effective when paralogous gene copies are a concern. When many paralogous copies are present, standard EPIC approaches may fail because the primer sites are conserved across paralogs. By placing one primer in the conserved exon and the other in the relatively less-conserved 3′-UTR, Whittall *et al.* (2006) found it possible to isolate homologs. This result has been observed in several other studies and may be general (Perry & Bousquet 1998; Brown *et al.* 2001; Temesgen *et al.* 2001). Whittall *et al.* (2006) were also able to target highly variable intron sequences by sequencing markers whose amplification products from genomic template DNA were significantly larger than expected based on the initial cDNA sequences (which do not contain introns). Because this approach relies very little on existing genetic resources (only BLAST hits to GenBank were used in helping to establish homology to known genes), it represents a simple, elegant solution to marker limitation in nearly any clade, regardless of the currently available genomic resources.

Mapping genes of interest

Studies of genome-wide genetic variation can also be used to identify genes underlying traits of ecological interest, thereby contributing to the fundamental evolutionary questions of how many and what kinds of genes are involved in adaptive divergence and of what types of nucleotide changes are involved in adaptive genetic differences (Beaumont & Balding 2004; Mitchell-Olds *et al.* 2007; Stinchcombe & Hoekstra 2007; Ellegren 2008). Genomic regions underlying adaptive divergence can be identified through genome scans and the detection of outlier loci (Beaumont & Balding 2004; Vasemagi & Primmer 2005; Ellegren 2008). Loci that differ from the expectations of neutral evolution, which should govern most loci in a genome, or demonstrate a different genetic signature from the genome-wide background, signal regions of the genome that contain candidate genes for traits underlying ecological differentiation (Storz *et al.* 2004; Stinchcombe & Hoekstra 2007; Holderegger *et al.* 2008).

In order to realize the potential for population genomic data sets to elucidate such genes of ecological interest, linkage maps that match markers with specific genomic regions must be developed. Linkage maps are important for assessing linkage disequilibria, for evaluating the extent of genomic coverage of a set of markers, and ultimately for the development of additional markers to increase the fine-scale resolution of particular genomic regions of interest (Stinchcombe & Hoekstra 2007). Linkage maps are frequently constructed using microsatellite markers due to their hypervariability, but microsatellite markers are not as abundant or easily scored as SNPs, and an increase in the availability of genomic resources allows for the development and use of EPIC markers for linkage mapping (Slate *et al.* 2009).

The use of EPICs to construct linkage maps has the potential to provide a greater ability to locate genes under selection in non-model systems because conserved priming sites allow an increased ability to identify homologous regions in the genomes of related species. Furthermore, the development and use of EPIC markers allows for the strategic spacing of markers at regular genomic intervals, which is more difficult to accomplish with markers that were initially developed anonymously, or occur less frequently across the genome (such as microsatellites). The downside of using EPICs is a decrease in variability relative to microsatellites. However, the supplemental use of microsatellites in addition to EPIC markers can help to increase the reliability of linkage maps, while maintaining the benefits achieved by the use of EPICs. For example, Backström *et al.* (2008b) constructed a linkage map for the collared flycatcher using 170 EPIC markers developed from the chicken genome and supplemented those data with 71 microsatellites developed in related species and recovered an estimated 75–80% of the flycatcher genome. Interestingly, Backström *et al.* (2008b) found similar information content in their gene-based EPIC markers and the microsatellites, which illustrates that in some cases sufficient variation for linkage mapping can be obtained from the intronic SNPs present in EPIC markers alone.

Additional considerations in marker development

Success rate

An important consideration when deciding among alternative strategies is the efficiency of each method, measured as the proportion of markers that amplify the correct product across the species of interest. When many candidate sequences are available for marker design (as when using existing large genomic resources), efficiency is much less important than other properties of the marker development strategy because researchers can discard non-ideal candidates and move on. However, when researchers are investing on a per marker basis (e.g. when sequencing additional small-insert clones, or converting additional AFLPs) this becomes an important factor. Fortunately, most methods report success rates that are encouragingly high. Aitken *et al.* (2004) studied the utility of EPIC markers for cross-amplification in distant taxa by screening existing mammalian-derived EPIC markers on a panel of 16 representative mammal species and several chimpanzees with the goals of quantifying the proportion of markers that amplified across the panel and were variable (within the chimpanzee panel). Aitken *et al.* (2004) found that approximately half of the markers yielded amplification products of the expected size across the test panel (ranging from 24% in opposum to 74% in mouse). For chimpanzees, where the larger sample size allowed for discovery of SNPs, the authors found 26 potential SNPs in six loci (1 SNP per 400 bp), though five other loci were invariant. Though the utility of marker sets clearly varies according to which taxa are tested, how closely related they are to the reference taxon, the conditions under which the markers were developed, and the questions to which they are applied; other studies using NPCLs, EPICs, and ANMs have all found similarly high success rates (Fig. 3). In Backström *et al.* (2008a), the lowest observed success rate of 34% (in Tengmalm's owl) would still result in around 80 markers being useful, which is a very large number of markers for phylogenetic and population genetic studies in a non-model species.

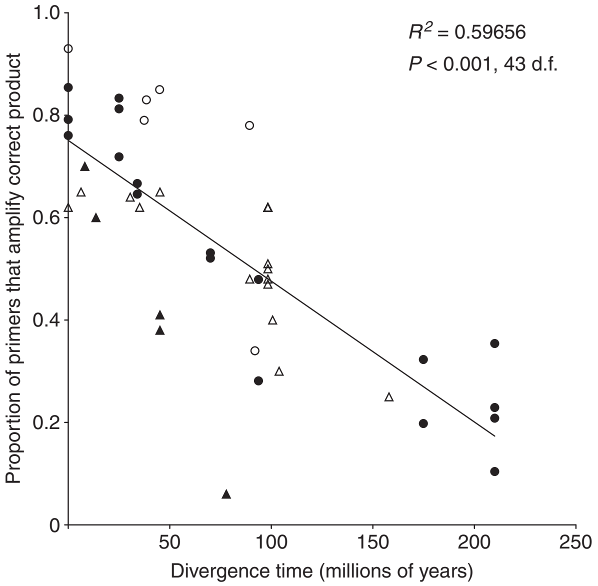

Fig. 3 Linear regression demonstrating the correlation between the proportion of primers that amplify correct products in test species and the divergence time between the test species and reference species. Data are from Thomson *et al.* (2008; closed circles), Aitken *et al.* (2004; open triangles), Peng *et al.* (2009; closed triangles), and Backström *et al.* (2008a; open circles). Divergence times are weighted averages from Hedges *et al.* (2006) and are taken as the divergence between the test taxon and the reference species, or the phylogenetically closest reference species in the case of universal primers. The regression and correlation apply to pooled data.

Ideally, we would like to compare efficiency among the different methods that have been used. However, such comparisons are problematic because of variation in phylogenetic distances, number of taxa tested, and substitution rates between clades among the different studies. Instead of comparing success rates across clades, we can compare the proportion of markers that work in the reference species that markers were developed from. Here, the primers for each marker should be a perfect match or have degenerate sites that allow for a perfect match to the reference species. Thomson *et al.* (2008) developed a set of 96 ANMs for phylogenetics and phylogeography in turtles. This ANM set successfully amplified a specific product of the appropriate size in the reference taxon for 76% of markers (73 out of 96), while the Rosenblum *et al.* (2007) ANMs successfully amplified 65% (50 out of 77). Conversely, the well annotated Backström *et al.* (2008a) EPIC markers had a 93% success rate in the reference taxon. Li *et al.* (2007) do not report an overall success rate for all of the NPCL markers that they examined, but for the 10 markers on which they focused, all amplified a single product of the expected size in both of the reference taxa they used.

A related consideration is the phylogenetic span over which markers work. If researchers wish to maximize phylogenetic span (likely at the expense of marker variability) the obvious approach is to focus on NPCLs using universal primers, However, several existing marker development projects suggest that alternative marker types might be useable across large phylogenetic distances, making it unnecessary to rely exclusively on NPCLs. We compiled examples of novel marker sets that were tested across a large phylogenetic span (> 50 million years [My]) and asked what proportion of markers worked across a given length of evolutionary time (Fig. 3). Too few examples exist in the literature to permit a statistical comparison of different marker development methods, but overall, the decrease in proportion of working markers over time appears to be qualitatively similar. This comparison includes examples of EPIC (Aitken *et al.* 2004; Backström *et al.* 2008a; Peng *et al.* 2009) and ANM (Thomson *et al.* 2008) marker design and, overall, shows that approximately 50% of markers work across a 100 My time span for the clades examined. These results are based on relatively few studies and so may not be general, though what we can take from them is encouraging for studies across relatively modest levels of diversity. For studies examining very large time spans (e.g. phylogenomics of metazoa) these approaches are unlikely to be as fruitful. In these cases, researchers have turned away from marker development *per se* and toward large scale sequencing as a means to develop comparative datasets across very large phylogenetic spans (e.g. Dunn *et al.* 2008).

Assessing homology

For most applications, the inclusion of markers unknowingly designed from paralogs or from repetitive elements can be extremely problematic, thus accurate homology detection is critical. Researchers have used three distinct strategies to detect and avoid paralogous markers: (i) similarity searches; (ii) phylogenetic tests; and (iii) characteristics of the markers themselves. The strategies are not mutually exclusive and, ideally, should be combined, as each has its own strengths and weaknesses.

Marker development in systems with good genomic resources can simply use the existing annotation information of these genomes to design markers using only a known set of orthologous genes (e.g. Lyons *et al.* 1997; Jiang *et al.* 1998). When this information is not available (as is still the case in most non-model organisms), the most widely used strategy relies on BLAST searches. Most commonly, researchers simply BLAST the potential marker sequence against a well-annotated genome and discard potential markers with multiple high scoring hits or hits to a known gene family (Putta *et al.* 2004; Backström *et al.* 2008a; Townsend *et al.* 2008). When the resources are available, more stringent strategies can be employed that test single copy status across multiple genomes, further ensuring against accidental use of non-orthologous markers (Li *et al.* 2007; Peng *et al.* 2009). The utility of this class of approaches depends entirely on the type of genomic resources available for comparison. Though, when the appropriate resources are available, either within the study organisms or closely related species, these are likely the most reliable strategies.

A second approach uses phylogenetic information in the markers to detect possible paralogs. This approach is less commonly used and is less likely to detect non-single copy markers. However, it also relies less on the availability of well-annotated genomic resources than the more reliable BLAST strategies. Whittall *et al.* (2006) sequenced and aligned markers and then checked (1) that each marker supported two well-accepted phylogenetic relationships in the clade of interest; and (2) that no loci supported phylogenies significantly incongruent with each other. Because variation in gene genealogies is expected to occur, particularly in large sets of markers and in rapid radiations, we do not recommend that researchers utilize this second criterion. However, the first criterion was able to confirm that the markers were behaving as expected. Li *et al.* (2007) sequenced 10 NPCL markers for a panel of 14 taxa. Using BLAST searches (with greatly relaxed stringency settings), they found alignable paralogs for 7 of the 10 genes examined. The authors included these paralogous copies in phylogenetic analyses with the sequence data from their markers and found that the paralogous copies were sister to a monophyletic clade composed of all of the sequenced gene copies, verifying that they were ancient duplications that likely occurred early in the vertebrate lineage and posed little problem for their analysis. These approaches are unlikely to identify all non-single copy markers and so we do not recommend that they be used as the primary basis for a homology detection strategy.

The final class of approaches looks for signatures of non-orthology within the marker sequences themselves.

A common strategy is simply to examine the sequence data for sites that are apparently heterozygous across all individuals sequenced. When primers co-amplify pseudogenes or multiple copy elements, any variation between the multiple copies will appear in the sequence data as apparent heterozygous SNPs. When working within a single species (phylogeography) or very closely related species, this approach is practical (Jennings & Edwards 2005; Rosenblum *et al.* 2007). Across larger evolutionary distances however, it becomes more likely that researchers could preferentially amplify different copies in different taxa due to mutations in the primer site. A second strategy relies on structural information about the markers themselves. Whittall *et al.* (2006) amplified markers that were primed in an exon and the 3′-UTR, amplifying an intervening unconstrained region. They checked that the sequences exhibited the expected pattern of little variation in the constrained exons and 3′-UTR, with more variation in the unconstrained intervening region. Contrastingly, the sequencing of pseudogenes would result in similar amounts of variation across the entire marker.

For ANMs, a larger problem is detecting and avoiding repetitive elements, such as SINEs, LINEs, and other retrotransposons. A very large fraction of the noncoding regions of many genomes are composed of these elements and thus it is essential that researchers working with ANMs pay careful attention to their detection. Thomson *et al.* (2008) attempted to screen for repetitive elements by comparing each sequence to GenBank and to the RepBase vertebrate repetitive element and transposable element libraries (Smit *et al.* 1996–2004). Even after excluding the very common CR1-family of LINE elements, nearly half of the markers examined were flagged as repetitive by at least one of the repeat detection comparisons. Overall, Thomson *et al.* (2008) observed little concordance among methods for identifying repetitive elements, suggesting that multiple methods should be employed for this important step.

We suggest that the ideal strategy for paralog detection should primarily rely on a stringent BLAST search to well-annotated resources whenever possible. This has the highest likelihood of successfully detecting paralogs and is easily implemented. However, researchers should utilize whatever information they can and be cognizant of other potential indicators of multiple copy markers. Fortunately, as the number and quality of genomic resources increase, the detection of multiple copy genetic regions will become substantially easier.

Future directions

Next generation sequencing technologies will undoubtedly change the way that sequence data are collected and analysed in non-model systems. These technologies will increase the speed and decrease the cost of collecting large quantities of sequence data, making the acquisition of many-marker datasets readily feasible. Typically, next generation sequencing is cited as providing faster and cheaper routes to genome or transcriptome assembly. However, this need not be the ultimate goal of large-scale sequencing projects. In fact, next generation sequencing provides avenues to many broader applications, including marker development.

In particular, parallel sequencing on next generation sequencing platforms (e.g., 454 Life Sciences, Branford, CT) enables the collection of homologous DNA sequences from large pools of specimens. Typically, parallel sequencing has applied a ‘shotgun’ approach to large-scale sequencing, in which an individual genome is first broken up into small fragments and then these fragments are randomly sequenced (Hudson 2008; Morozova & Marra 2008; Wheeler *et al.* 2008). This approach has been effective in resequencing individual genomes in a very short period of time (Wheeler *et al.* 2008). However, this technology also allows the sequencing of targeted gene regions simultaneously from a pool of PCR products from different individuals (Binladen *et al.* 2007; Meyer *et al.* 2008; Babik *et al.* 2009; Wegner 2009). In parallel tagged sequencing (PTS), unique 5′-nucleotide tagged primers are applied to individual specimens by either PCR (Binladen *et al.* 2007) or blunt-end ligation (Meyer *et al.* 2008), pools of tagged specimens are sequenced in parallel, and the resulting sequences are traced back to individual specimens through 5′ tag analysis. Because individual gene regions can be amplified while simultaneously applying 5′ tags, this process allows for the targeting of homologous sequences from a large number of specimens (Binladen *et al.* 2007). This sequencing technology clearly holds great potential for enabling studies that require large volumes of sequence data from many individuals. However, because it requires PCR primers that cross-amplify in all species under investigation, it increases, rather than obviates, the importance of high throughput marker discovery methods.

Next generation sequencing methods can also directly contribute to marker discovery itself; specifically through the use of mate-pair sequencing techniques in which two linked reads are separated by a known distance. This application is available in traditional Sanger sequencing, the 454 platform, sequencing by synthesis (SBS; Illumina, Hayward, CA, USA), and SOLiD sequencing (Applied Biosystems, Foster City, CA, USA). Next generation sequencing technologies currently produce shorter read lengths than Sanger sequencing, ranging from 25–50 bp in SBS and SOLiD to 400bp in 454 sequencing (Hudson 2008; Mardis 2008;

Morozova & Marra 2008; Ansorge 2009), and this has been viewed as a limitation in their efficacy for identifying suitable primer regions (although emerging DNA sequencing approaches hold promise to drastically increase read lengths, e.g., Pacific Biosciences SMRT technology). However, if the distance in mate-pair sequencing is fixed at a length typically sought after when designing sequencing primers (500–1500 bp, for instance), then although this intervening region will remain anonymous until fully sequenced, the short sequences achieved by mate-pair sequencing may produce suitable primer sites from which PCR primers can be designed. In many cases, these sequences might not be suitable primer regions. However, even if suitable primers are detected at a low rate, because the cost of collecting these reads using next generation sequencers is so much lower than traditional sequencing, this approach may successfully identify ANMs from across the genome with relatively little expense.

Finally, next generation sequencing technologies are rapidly advancing the discovery of markers for genes of ecological interest in non-model organisms. Complete transcriptomes for non-model species can be sequenced on these platforms from isolated RNA or cDNA (Emrich *et al.* 2007; Toth *et al.* 2007; Morozova & Marra 2008; Vera *et al.* 2008). Annotation of the resulting sequences can then occur by BLAST comparison with existing gene and genome resources (Emrich *et al.* 2007; Morozova & Marra 2008). This method has been successfully used to characterize the transcriptomes of maize (Emrich *et al.* 2007), the Glanville fritillary butterfly (Vera *et al.* 2008), and a paper wasp (Toth *et al.* 2007), demonstrating its potential in non-model systems. Using this method, Toth *et al.* (2007) were able to identify over 3000 genes in the paper wasp, based upon similarity to the honeybee genome even though these species diverged 100–150 Ma. Their assays from certain classes of these genes allowed them to identify candidate genes that contribute to the complex maternal behaviour and eusociality in these wasps. A straightforward extension of these methods could also allow for the discovery of very large numbers of EPICs. Because intron positions tend to be conserved across genomes (Rogozin *et al.* 2003), it should be possible to compare these sequences to closely related fully-sequenced genomes (when they exist) in order to design primers that span the predicted location of introns in the target species. These approaches demonstrate not only the potential for identifying genes underlying complex ecological traits but also the potential of 454 sequencing for identifying genic markers even when the nearest reference genome is relatively distantly related to the non-model organisms of interest.

Recommendations and conclusions

The number of genome-enabled species has been steadily growing, but some taxonomic groups still have proportionately fewer genomic resources than others (Table 2). Several of the methods and studies reviewed here have demonstrated that genomic resources are widely transferrable to related organisms for purposes of gene identification and marker development. Future efforts to establish genomic resources should consider their value not just to the organism from which genomic DNA is isolated but also to closely related organisms that may benefit from resource development. For instance, several of the studies reviewed here have demonstrated the utility of reference genomes that diverged from the species of interest as long as 150–210 Ma (Fig. 3; Toth *et al.* 2007; Thomson *et al.* 2008). Numerous speciose groups have originated less than 100 Ma. For example, the oldest date of divergence in the passerine birds, containing over 5700 species, is approximately 82 Ma (Barker *et al.* 2004), and the superfamily of treefrogs, *Hyloidea*, with over 3000 species, is believed to have emerged less than 100 Ma (Crawford & Smith 2005; Santos *et al.* 2009). The development of genomic resources in groups like these could impact hundreds or even thousands of species.

As the proliferation of genome-scale datasets continues, data types, analytical approaches, and questions that were once unavailable for research in non-model systems are becoming routine. Many of the recently described marker development methods rely only on molecular biology tools found in standard genetic laboratories and on simple bioinformatic techniques. Moreover, several of these studies demonstrate that marker development can be very efficient in both time and cost, regardless of the particular resources already available within the clade under study. Several studies have also demonstrated that markers developed in one system may cross-amplify in others. This greatly enhances the utility of existing markers and provides fast routes to marker development in many systems. For this reason, we encourage researchers to contribute primer and cross-amplification information to publically available databases (e.g. the Molecular Ecology Resources Database or the Dryad Data Repository).

Overall, the disappearance of marker limitations represents a qualitative shift in the way that researchers can now approach questions and design analysis strategies. Rather than being constrained to the research approaches that existing markers allow, researchers can now focus on deciding which questions to ask and how to develop the appropriate markers for those studies. This change is a boon to research on wild species that should be fully realized as genomic data become easier to acquire and analytical methods capable of fully utilizing these data mature.

Acknowledgements

We thank Anthony Barley, Louis Bernatchez, Dean Leavitt, Greg Pauly, Brad Shaffer, Phil Spinks, and two anonymous reviewers for comments on this manuscript. The authors were supported by the UC Davis Center for Population Biology, an NSF Doctoral Dissertation Improvement Grant (to RCT), and an NSF Graduate Research Fellowship (to IJW).

References

Aitken N, Smith S, Schwarz C, Morin P (2004) Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach. *Molecular Ecology*, **13**, 1423–1431.

Ansorge WJ (2009) Next-generation DNA sequencing techniques. *New Biotechnology*, **25**, 195–203.

Babik W, Taberlet P, Ejsmond MJ, Radwan J (2009) New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Molecular Ecology Resources*, **9**, 713–719.

Backström N, Fagerberg S, Ellegren H (2008a) Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome. *Molecular Ecology*, **17**, 964–980.

Backström N, Brandström M, Gustafsson L *et al.* (2008b) A gene-based genetic linkage map of the collared flycatcher (*Ficedula albicollis*) reveals extensive synteny and gene-order conservation during 100 million years of Avian evolution. *Genetics*, **179**, 1479–1495.

Barker FK, Cibois A, Schikler P, Feinstein J, Cracraft J (2004) Phylogeny and diversification of the largest avian radiation. *Proceedings of the National Academy of Sciences, USA*, **101**, 11040–11045.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Binladen J, Gilbert MTP, Bollback JP *et al.* (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, **2**, e197.

Black WC, Baer CF, Antolin MF, DuTeau NM (2001) Population genomics: genome-wide sampling of insect populations. *Annual Review of Entomology*, **46**, 441–469.

Bradeen J, Simon P (1998) Conversion of an AFLP fragment linked to the carrot Y2 locus to a simple, codominant, PCR-based marker form. *Theoretical and Applied Genetics*, **97**, 960–967.

Brito PH, Edwards SV (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, **135**, 439–455.

Brown G, Kadel E, Bassoni D *et al.* (2001) Anchored reference loci in loblolly pine (*Pinus taeda* L.) for integrating pine genomics. *Genetics*, **159**, 799–809.

Brugmans B, van der Hulst RGM, Visser RGF, Lindhout P, van Eck HJ (2003) A new and versatile method for the successful conversion of AFLP markers into simple single locus markers. *Nucleic Acids Research*, **31**, e55–e55.

Crawford AJ, Smith EN (2005) Cenozoic biogeography and evolution in direct-developing frogs of Central America (Leptodactylidae: *Eleutherodactylus*) as inferred from a phylogenetic analysis of nuclear and mitochondrial genes. *Molecular Phylogenetics and Evolution*, **35**, 536–555.

Dunn CW, Hejnol A, Matus DQ *et al.* (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**, 745–749.

Edwards S (2009) Is a new and general theory of molecular systematics emerging? *Evolution*, **63**, 1–19.

Ellegren H (2008) Comparative genomics and the study of evolution by natural selection. *Molecular Ecology*, **17**, 4586–4596.

Emrich SJ, Barbazuk WB, Li L, Schnable PS (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, **17**, 69–73.

Hare MP (2001) Prospects for nuclear gene phylogeography. *Trends in Ecology and Evolution*, **16**, 700–706.

Hedges SB, Dudley J, Kumar S (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, **22**, 2971–2972.

Hey J, Machado C (2003) The study of structured populations: new hope for a difficult and divided science. *Nature Reviews Genetics*, **4**, 535–543.

Holderegger R, Herrmann D, Poncet B *et al.* (2008) Land ahead: using genome scans to identify molecular markers of adaptive relevance. *Plant Ecology and Diversity*, **1**, 273–283.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

Jennings WB, Edwards SV (2005) Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. *Evolution*, **59**, 2033–2047.

Jiang Z, Priat C, Galibert F (1998) Traced orthologous amplified sequence tags (TOASTs) and mammalian comparative maps. *Mammalian Genome*, **9**, 577–587.

Lee J, Edwards S (2008) Divergence across Australia's carpentarian barrier: statistical phylogeography of the red-backed fairy wren (*Malurus melanocephalus*). *Evolution*, **62**, 3117–3134.

Li C, Ortí G, Zhang G, Lu G (2007) A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. *BMC Evolutionary Biology*, **7**, 44.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Lyons LA, Laughlin TF, Copeland NG *et al.* (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nature Genetics*, **15**, 47–56.

Manel S, Schwartz M, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution*, **18**, 189–197.

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, **24**, 133–141.

Meksem K (2001) Conversion of AFLP bands into high-throughput DNA markers. *Molecular Genetics and Genomics*, **265**, 207–214.

Meyer M, Stenzel U, Hofreiter M (2008) Parallel tagged sequencing on the 454 platform. *Nature Protocols*, **3**, 267–278.

Mitchell-Olds T, Willis JH, Goldstein DB (2007) Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics*, **8**, 845–856.

Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*, **92**, 255–264.

Murphy WJ, Eizirik E, Johnson WE *et al.* (2001) Molecular phylogenetics and the origins of placental mammals. *Nature*, **409**, 614–618.

O'Brien S, Womack J, Lyons L *et al.* (1993) Anchored reference loci for comparative genome mapping in mammals. *Nature Genetics*, **3**, 103–112.

Palumbi SR, Baker CS (1994) Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. *Molecular Biology and Evolution*, **11**, 426–435.

Peng Z, Elango N, Wildman DE, Soojin VY (2009) Primate phylogenomics: developing numerous nuclear non-coding, non-repetitive markers for ecological and phylogenetic applications and analysis of evolutionary rate variation. *BMC Genomics*, **10**, 247.

Perry D, Bousquet J (1998) Sequence-tagged-site (STS) markers of arbitrary genes: the utility of black spruce-derived STS primers in other conifers. *TAG Theoretical and Applied Genetics*, **97**, 735–743.

Putta S, Smith J, Walker J *et al.* (2004) From biomedicine to natural history research: EST resources for ambystomatid salamanders. *BMC Genomics*, **5**, 54.

Rannala B, Yang Z (2008) Phylogenetic inference using whole genomes. *Annual Review of Genomicsand& Human Genetics*, **9**, 217–231.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Current Biology*, **13**, 1512–1517.

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 798–804.

Rosenblum EB, Belfiore NM, Moritz C (2007) Anonymous nuclear markers for the eastern fence lizard, *Sceloporus undulatus*. *Molecular Ecology Notes*, **7**, 113–116.

Rowe KC, Reno ML, Richmond DM, Adkins RM, Steppan SJ (2008) Pliocene colonization and adaptive radiations in Australia and New Guinea (Sahul): multilocus systematics of the old endemic rodents (Muroidea: Murinae). *Molecular Phylogenetics and Evolution*, **47**, 84–101.

Santos JC, Coloma LA, Summers K *et al.* (2009) Amazonian amphibian diversity is primarily derived from late Miocene Andean lineages. *PLoS Biology*, **7**, e1000056.

Shan X, Blake T, Talbert L (1999) Conversion of AFLP markers to sequence-specific PCR markers in barley and wheat. *TAG Theoretical and Applied Genetics*, **98**, 1072–1078.

Siebert P, Chenchik A, Kellogg D, Lukyanov K, Lukyanov S (1995) An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Research*, **23**, 1087–1088.

Slate J, Gratten J, Beraldi D *et al.* (2009) Gene mapping in the wild with SNPs: guidelines and future directions. *Genetica*, **136**, 97–107.

Smit A, Hubley R, Green P (1996-2004) RepeatMasker Open-3.0. Available: http://www.repeatmasker.org.

Spinks PQ, Thomson RC, Lovely GA, Shaffer HB (2009) Assessing what is needed to resolve a molecular phylogeny: simulations and empirical data from Emydid turtles. *BMC Evolutionary Biology*, **9**, 56.

Stinchcombe JR, Hoekstra HE (2007) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.

Storz JF, Payseur BA, Nachman MW (2004) Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Molecular Biology and Evolution*, **21**, 1800–1811.

Temesgen B, Brown G, Harry D *et al.* (2001) Genetic mapping of expressed sequence tag polymorphism (ESTP) markers in loblolly pine (*Pinus taeda* L.). *TAG Theoretical and Applied Genetics*, **102**, 664–675.

Thomson RC, Shedlock AM, Edwards SV, Shaffer HB (2008) Developing markers for multilocus phylogenetics in non-model organisms: a test case with turtles. *Molecular Phylogenetics and Evolution*, **49**, 514–525.

Toth AL, Varala K, Newman TC *et al.* (2007) Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science*, **318**, 441.

Townsend TM, Alegre RE, Kelley ST, Wiens JJ, Reeder TW (2008) Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: an example from squamate reptiles. *Molecular Phylogenetics and Evolution*, **47**, 129–142.

Vasemagi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology*, **14**, 3623–3642.

Venta P, Brouillette J, Yuzbasiyan-Gurkan V, Brewer G (1996) Gene-specific universal mammalian sequence-tagged sites: application to the canine genome. *Biochemical Genetics*, **34**, 321–341.

Vera JC, Wheat CW, Fescemyer HW *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Wegner KM (2009) Massive parallel MHC genotyping: titanium that shines. *Molecular Ecology*, **18**, 1818–1820.

Wheeler DA, Srinivasan M, Egholm M *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.

Whittall JB, Medina-Marino A, Zimmer EA, Hodges SA (2006) Generating single-copy nuclear gene data for a recent adaptive radiation. *Molecular Phylogenetics and Evolution*, **39**, 124–134.

Bob Thomson studies systematics and conservation of amphibians and reptiles, emphasizing genomic and informatic approaches in particular. Ian Wang studies the role of landscapes and environments on population structure, gene flow, and adaptation in amphibians and is especially interested in multi-locus phylogeography and landscape genetics. Jarrett Johnson is interested in the evolutionary ecology and conservation of amphibians with an emphasis on adaptation in variable habitats and movements in heterogenous landscapes.