MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) **14**, 352–360 doi: 10.1111/1755-0998.12171

Development, validation and high-throughput analysis of sequence markers in nonmodel species

P. ZIELIŃSKI, M. T. STUGLIK, K. DUDEK, M. KONCZAL and W. BABIK
Institute of Environmental Sciences, Jagiellonian University, Gronostajowa 7, 30-387 Kraków, Poland

Abstract

DNA sequences derived from multiple regions of the nuclear genome are essential for historical inferences in the fields of phylogeography and phylogenetics. The appropriate markers should be single-copy, variable, easy to amplify from multiple samples and easy to sequence using high-throughput technologies. This may be difficult to achieve for species lacking sequenced genomes and particularly challenging for species possessing large genomes, which consist mostly of repetitive sequences. Here, we present a cost-effective, broadly applicable framework for designing, validating and high-throughput sequencing of multiple markers in nonmodel species without sequenced genomes. We demonstrate its utility in two closely related species of newts, representatives of urodeles, a vertebrate group characterized by large genomes. We show that over 80 markers, *c.* 600 bp each, developed mainly from 3′ untranslated transcript regions (3′UTR) may be effectively multiplexed and sequenced. Data are further processed using standard, freely available bioinformatic tools, producing phase-resolved sequences. The approach does not require barcoded PCR primers, and the cost of library preparation is independent of the number of markers investigated. We hope that this approach will be of broad interest for researchers working at the interface of population genetics and phylogenetics, exploring deep intraspecific genetic structure, species boundaries and phylogeographies of closely related species.

Keywords: haplotype phasing, newt, nuclear markers, sequencing, transcriptome

Received 13 June 2013; revision received 23 August 2013; accepted 13 September 2013

Introduction

DNA sequences derived from multiple regions of the nuclear genome are essential for historical inferences about demography, gene flow and phylogeny of closely related species (Brito & Edwards 2009; Degnan & Rosenberg 2009; Heled & Drummond 2010; Hickerson *et al.* 2010; McCormack *et al.* 2013; Sousa & Hey 2013). Therefore, there has been a continuing interest in approaches that allow the straightforward development of numerous (tens to hundreds) single-copy nuclear sequence markers (Lemmon & Lemmon 2012; McCormack *et al.* 2013; O'Neill *et al.* 2013). Such markers are then amplified from multiple samples and sequenced using next-generation sequencing (NGS) technologies. Ideal markers should exhibit substantial polymorphism and amplify in closely related species, enabling exploration of species boundaries and genealogy-based analysis of interspecific gene flow and introgression (Twyford & Ennos 2012).

Correspondence: Wieslaw Babik, Fax: +48 12 664 69 12; E-mail: wieslaw.babik@uj.edu.pl

Advances in NGS have led to the development of large-scale, genome-wide approaches to reduced-representation sequencing that may provide thousands of markers densely covering the genome, such as restriction-site-associated DNA sequencing (RAD-seq), genotyping by sequencing (GBS) and others (reviewed in Davey *et al.* 2011). These markers are however usually used to identify and genotype SNPs, or in applications which rely on the availability of a reference genome – in such cases, they are basically methods for partial genome resequencing. While tremendously useful for analyses of genetic structure, genome scans for selection and other population genetics-based inferences, RAD-seq or GBS may be less useful as the level of polymorphism and/or divergence increases, because the proportion of restriction sites shared between samples decreases. Also, these markers usually provide relatively short sequences adjacent to restriction sites, of limited use for genealogical inferences. The latter problem may be partially alleviated by paired-end RAD (PE-RAD) approaches which generate longer contigs (Willing *et al.* 2011; Hohenlohe *et al.* 2013).

Species with very large, unknown genomes pose particular challenges. Reduced-representation techniques

© 2013 John Wiley & Sons Ltd

(Davey *et al.* 2011; Lemmon & Lemmon 2012) in such species require a large amount of sequencing, and the highly repetitive nature of large genomes and fast evolution of noncoding sequences make the development of robust sets of sequence markers for such organisms difficult. Targeted sequence capture methods (Mamanova *et al.* 2010; Faircloth *et al.* 2012) may be helpful in such cases, but these are typically used when a very high number of markers is needed and pose challenges of their own, such as variation in coverage among targeted regions and sometimes low fraction of reads on target (Bi *et al.* 2012; Lemmon & Lemmon 2012). In many situations, a more modest number of high-quality sequence markers is sufficient for robust inference.

Here, we present a generally applicable method which utilizes transcriptome sequences to develop DNA sequence markers (Fig. 1) and demonstrates its utility in two closely related, naturally hybridizing urodele species, the smooth (*Lissotriton vulgaris*) and Carpathian (*Lissotriton montandoni*) newts (Zieliński *et al.* 2013). We show that the markers developed from 3′ untranslated transcript regions (3′UTR) can be effectively multiplexed and sequenced using NGS technology. Data are further processed using standard, freely available bioinformatic tools, producing phase-resolved sequences for a large number of individuals. The approach does not require barcoded PCR primers, and the cost of library preparation is independent of the number of markers investigated.

Materials and methods

Marker development

Markers were developed from de novo assembled liver transcriptome. Illumina sequencing and de novo assembly with Trinity (Grabherr *et al.* 2011) were performed for six *Lissotriton montandoni* and six *Lissotriton vulgaris* individuals representative of the genetic diversity of both species; the details of the transcriptome assembly and analysis will be provided elsewhere (M. Stuglik & W. Babik, in preparation). Trinity assembly was processed further with a custom bioinformatic pipeline to construct transcriptome-based gene models (TGM, M. Stugluk, W. Babik & J. Radwan, submitted), that is, nonredundant representation of transcribed genomic sequences. This step was necessary because Trinity reconstructs as separate contigs sequences of alternatively spliced forms and sometimes also of divergent alleles, making the assembled transcriptome a redundant representation of the transcribed part of the genome.

We focused on 3′UTRs rather than on protein-coding parts of transcripts for marker development. The vast majority of protein-coding exons in vertebrates are short, therefore targeting such exons would seriously limit the scope of available markers. Additionally, markers located entirely in coding exons are usually less polymorphic and more prone to the effects of selection. The exon-priming-intron-crossing EPIC technique, although possible (Nadachowska & Babik 2009), is expected to produce poor results because introns in salamanders are known to be exceptionally long (Smith *et al.* 2009).

Herein, we build on the observation that the overwhelming majority of 3′UTRs in vertebrates are intronless (Hong *et al.* 2006). They are thus contained entirely within a single exon; such last exons may be very long and consist mostly of noncoding sequence. Although functionally important and sometimes containing highly conserved elements (Siepel *et al.* 2005), 3′UTRs are generally less evolutionarily constrained than protein-coding regions, with average mutation rates comparable to those

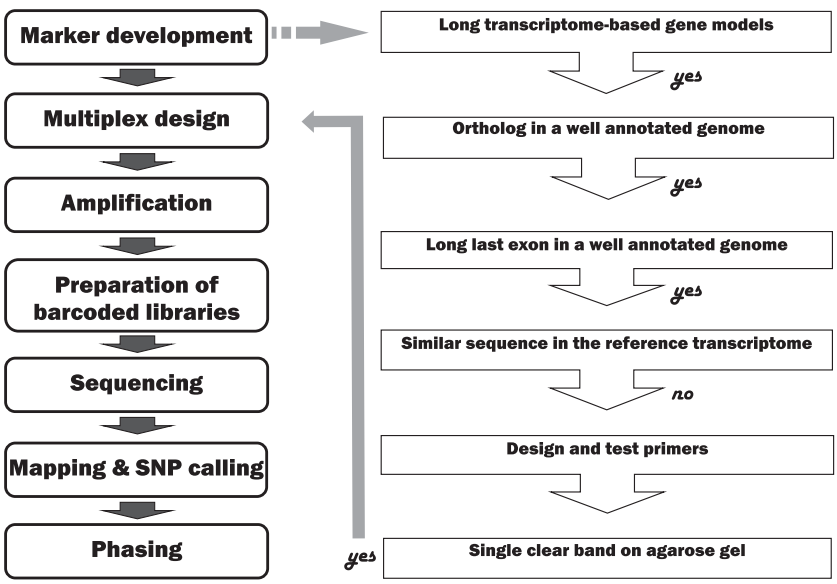

Fig. 1 Workflow illustrating the development and sequencing of the markers.

© 2013 John Wiley & Sons Ltd

of synonymous sites (Makałowski & Boguski 1998). The fact that 3′UTRs are closely associated with genes may be considered as either a strength or a weakness of this approach. However, the advantages appear to outweigh the disadvantages because: (i) it is a common practice to use protein-coding genes in phylogeographic analyses; (ii) some sequence conservation may well be beneficial in terms of marker versatility and broader utility across closely related species; and (iii) the development of markers may be more efficient compared with the situation when random genomic fragments are targeted. The efficiency is illustrated by the observation that in newts attempts to generate markers by random cloning of nuclear DNA fragments suffered from a high failure rate because most fragments were derived from repetitive sequences (W. Babik, unpublished). This may be caused by a high proportion of transposable-element-derived repetitive sequences in very large salamander genomes (Sun *et al.* 2012).

We identified our markers as follows (Fig. 1). We focused on longer TGMs, in the range of 5400–6500 bp, because longer TGMs are more likely to have long 3′ UTRs. The sequences of 3′UTRs are usually poorly conserved at deeper phylogenetic levels; therefore, we could not use blast searches for their identification. Instead, we employed an indirect approach. First, we performed blast search of TGMs against human and *Xenopus* transcripts. Those newt TGMs, which produced unambiguous hits to a single gene or gene family consisting of divergent members, were checked for the presence of a long 3′ portion, extending beyond the coding sequence. If human or *Xenopus* transcripts also had a reasonably long (i.e. at least several hundred bp) 3′UTR, such newt TGMs were classified as candidates for the design of primers. Candidate TGMs were blasted against the newt reference transcriptome to check for the presence of similar sequences which might represent paralogs; we proceeded with primer design only if putative paralogs were not detected. Primers targeting ***c.*** 600-bp fragments of the last exon (encoding mostly 3′UTR) in putative single-copy genes were designed with BatchPrimer3 (You *et al.* 2008). Due to the availability of sequences from several individuals, polymorphisms present in primer-binding sites were identified and incorporated as degenerate positions.

Samples

In total, 20 individuals were examined. Ten *L. montandoni* from four populations distributed throughout the species range, eight *L. vulgaris* from six populations including subspecies *L. v. lantzi* and *L. v. meridionalis* and single individuals of two other *Lissotriton* species *L. italicus* and *L. boscai*, to function as outgroup (Table S1, Supporting information). This sampling aimed at quantifying amplification efficiency across deep intraspecific and interspecific boundaries.

Laboratory procedures

All designed primer pairs were tested for amplification under identical PCR conditions. PCR reactions were performed in 10 μL; the reaction mix contained: 5 μL of 2 × Hot Start PCR Master Mix (Thermo Scientific), 0.2 μM of each primer, about 100 ng of genomic DNA. The PCR cycling scheme was 95 °C-3.5 min, 35 3-step cycles of 94 °C-30 s, 55 °-30 s, 72 °-45 s and a final extension at 72 °C-3 min. Only markers that produced a single clear band of expected size on the agarose gel were analysed further.

Multiplex Manager (Holleley & Geerts 2009) with complementarity threshold value set to six was used to design multiplex reactions allowing simultaneous amplification of several loci. Two plexity levels were used: (i) multiplexes consisting of 3–5 loci (A and B) and (ii) multiplexes containing 9–10 loci (C). Amplification reactions were performed in 15 μL and contained the following: 7.5 μL of Multiplex PCR Master Mix kit (Qiagen), 0.2 μM of each primer, about 100 ng of genomic DNA. The following cycling scheme was used: 95 °C-15 min, 30 3-step cycles of denaturation 94 °C-30 s, annealing 58 °C-30 s for A and 60 s for B and C, extension 72 °C-90 s followed by a final extension 72 °C-10 min. A, B and C amplification methods were used for eight, four and eight individuals, respectively. Because multiplexes may differ systematically in amplification efficiency, we verified for three individuals that the relative amount of product obtained from various multiplex reactions is approximately constant across individuals using Qubit quantitation. This enabled pooling multiplexes within an individual without the need of measuring DNA concentration in every amplicon. All multiplexes within the individual were pooled, cleaned with Qiagen MinElute PCR cleanup columns, and the resulting pool was used for sequencing.

Indexed sequencing libraries were prepared for each individual with the NexteraXT kit (Illumina) starting from 1 ng of pooled and cleaned PCR product according to the manufacturer's instructions and sequenced as a part of a single 2 × 150 bp MiSeq run. We conducted a BioAnalyzer assay after the NexteraXT PCR Clean-Up step and before bead-based normalization for quality control and to check the library size distribution.

Bioinformatics

The bioinformatic pipeline outlined below aimed at obtaining the phase-resolved sequences of both alleles for each marker and individual (Fig. 1). Although

© 2013 John Wiley & Sons Ltd

individual reads were shorter than the length of the amplified fragments, partial phase information contained in the paired-end reads can be utilized to reconstruct fully phased allele sequences. Because the Nextera tagmentation process causes a lower sequencing coverage of both amplicon ends, we trimmed 50 bp from each end of the markers. The bioinformatic pipeline comprised the following steps:

1 Mapping: reads were mapped to the trimmed reference sequences with Bowtie 2 (Langmead *et al.* 2009) using the following options: –no-unal which suppresses SAM records for reads that failed to align, –local mode which does not require that the entire read aligns from one end to the other, some bases may be clipped to achieve the highest possible alignment score and –very-sensitive alignment settings; the remaining parameters were set at their default values.
2 Local realignment: local realignment was performed to minimize the number of mismatched bases in regions with insertions or deletions (indels). Regions which required realignment were selected by RealignerTargetCreator, and realignment was performed with IndelRealigner, both tools are available in the Genome Analysis Toolkit (GATK) (McKenna *et al.* 2010; Depristo *et al.* 2011), default parameter values were used, except for the option which controls for the maximum interval length selected for the local realignment (-maxIntervalSize), which in our case was set equal to the marker length.
3 SNP calling: SNP calling was performed with GATK Unified Genotyper; the maximum number of alternative alleles per site was limited to two (–max_alternate_alleles 2); and genotyping was performed for all callable sites (–output_mode EMIT_ALL_SITES); PCR error rate was set to 0.005 (–pcr_error_rate 5.0E-3); we excluded all reads with mate unmapped or mapped to a different marker (–read_filter UnmappedRead, –read_filter BadMate); the minimum phred-scaled confidence threshold at which variants were called was set to 20 (–stand_call_conf 20.0); to exclude from variant calling low-quality bases, we set the minimum base quality required to consider it for variant calling to 30 (-mbq 30). Normally, if coverage is limited, multisample SNP calling, which integrates information about polymorphism across samples and incorporates population genetic priors, improves the detection of polymorphism (Depristo *et al.* 2011; Li 2011). However, if SNPs with more than two variants are present, this causes problems with haplotype reconstruction, because such positions are left unphased in all individuals (see below). Therefore, because in our case, coverage was high and we wanted to minimize the proportion of samples with unphased positions, SNP calling was performed separately for each individual.
4 Haplotype reconstruction: haplotype reconstruction through physical phasing of SNP positions was performed with GATK ReadBackedPhasing (Depristo *et al.* 2011). Briefly, the algorithm, which is currently restricted to biallelic SNPs, is based on the assumption that variants spanned by a read (or a pair of reads) exist on the same haplotype (allele). The variant call format (VCF) file containing information about polymorphic positions is used together with the Binary Alignment/Map (BAM) file, which contains the read mapping information, to reconstruct sequences of both haplotypes (alleles) occurring at a Mendelian locus in a particular individual.
5 Generation of allele sequences: information from phased vcf file was extracted to fasta sequence format with a custom Python script.

Markers, which exhibited higher than expected heterozygosity in at least one nucleotide position, were removed as they most likely represented paralogous loci. This was based on the observation that the expected heterozygosity at a biallelic SNP cannot exceed 50%. For each species, we identified SNPs with more heterozygotes than expected by chance (at the 0.01 significance level, calculated using binomial distribution) assuming equal allele frequencies.

We tested deviations from genotype frequencies expected under the Hardy–Weinberg equilibrium (HWE) using a sample of 20 *L. montandoni*: six reported in the present study and 14 sequenced for another project (P. Zieliński, K. Nadachowska-Brzyska & W. Babik, in preparation). These newts were sampled from 16 localities in the Eastern Carpathians, where little genetic structuring was detected (Zieliński *et al.* 2013). Population subdivision can only decrease the overall proportion of heterozygotes compared with expectations under random mating. Therefore, the test of HWE performed on a pooled sample should be effective in detecting the presence of null alleles although it may suffer from an increased frequency of false positives caused by population subdivision. Calculations were performed in GENEPOP 4.1.2 (Rousset 2008), and the type I error level for multiple tests was controlled using the Bonferroni correction. Basic statistics of DNA polymorphism were calculated for each marker in DnaSP (Rozas 2009).

Results

We designed 123 primer pairs, 96 (78%) of which produced a single clear band when tested in singleplex PCR. Eighty-seven markers were selected for amplification in multiplexes for 20 individuals (Table S2, Supporting information).

© 2013 John Wiley & Sons Ltd

We obtained 9.9 million paired-end reads, 7.4 million (75%) were mapped to references, and of these, 7.3 million (99%) were properly paired; that is, both reads from a pair were mapped to the same reference and in proper orientation. The average number of reads mapped per marker per individual was 8522 (SD ± 8070). The average per-base coverage per marker was 2222 (SD ± 2147.8). Despite bead-based library normalization, the overall amount of bases sequenced per individual varied by more than one order of magnitude (Fig. S1, Supporting information).

We considered a marker not amplifying in a given individual, if the average per-base coverage was lower than 10. Virtually all markers worked well in *Lissotriton vulgaris/montandoni*, only two and four failed for *L. v. lantzi* and *L. v. meridionalis*, respectively. Additionally, in one *L. vulgaris* individual with a low overall number of reads, indicating problems at the stage of library normalization, three markers also failed to reach the coverage threshold. Four markers (*kpnb, pola, samd8, scap*) were removed because of high frequency of heterozygotes, indicating than reads were derived from more than one genomic location. Six markers (*abl, cask, samdb, usp, ace, pik*) were excluded due to phasing inconsistencies possibly caused by a high incidence of PCR chimeras or the presence of very similar paralogs resulting from recent gene duplications. Considerably more markers failed for the two outgroup species, 30 (34%) for *L. italicus* and 23 (26%) for *L. boscai*. Forty-eight (55%) markers worked in both outgroup species (Table S2, Supporting information).

In *L. vulgaris/montandoni,* coverage varied among markers. Per-base coverage ranged from 209 to 4579× with average 2266× (SD ± 967.6) (Fig. 2a, Table S3, Supporting information). An important measure of marker performance was the relative coverage compared with other markers within an individual. This was expressed as a percentage of the total coverage (PTC) attributable to the marker. The mean PTC for various markers ranged from 0.10% to 2.45% (mean 1.15% SD ± 0.456) (Fig. 2b, Table S4, Supporting information). The results obtained at both plexity levels were comparable (Fig. 3).

Sequences of two alleles present in an individual need to be reconstructed, if more than one position is heterozygous. Read-backed phasing reconstructed the sequences of both alleles in 98% cases. The unresolved 2% were caused by locally low coverage or the presence of indels. In its current version, read-backed phasing uses only biallelic SNPs; therefore, positions containing three variants were left unphased in individuals heterozygous for variants not present in the reference (0.6% of all heterozygous sites).

Significant departures from the Hardy–Weinberg expectations, in all cases, the excess of homozygotes, were detected for only four (*abh, cldn, myo, rab33*) of 75 tested markers (5.3%). These departures may result from the presence of null alleles or population subdivision. Overall, tests of HWE indicate that null alleles are not common in our markers.

The number of segregating sites varied between markers and ranged from 3 to 50 (mean 17.6 SD ± 8.65) for *L. vulgaris* and from 1 to 39 (mean 11.4 SD ± 8.65) for *L. montandoni*. The nucleotide diversity varied extensively, ranging from 0.001 to 0.032 (mean 0.010 SD ± 0.0062) for *L. vulgaris* and from <0.001 to 0.023 (mean 0.006 SD ± 0.0052) for *L. montandoni*. The number of haplotypes identified within species ranged from 2 to 12 (mean 7.1 SD ± 2.32) for *L. vulgaris* and from 2 to 12 (mean 5.5 SD ± 2.58) for *L. montandoni* (Table S5, Supporting information).

Discussion

There has been a continuing interest in approaches that enable simple and cost-effective development of nuclear markers in the fields of phylogeography and phylogenetics (Tewhey *et al.* 2009; Davey *et al.* 2011; Lemmon & Lemmon 2012; Puritz *et al.* 2012). The appropriate markers should be single-copy, informative, easy to amplify from multiple samples and easy to sequence using high-throughput technologies (McCormack *et al.* 2013).

Here, we show that such DNA markers may be effectively identified from transcriptome sequences, multiplexed and sequenced using NGS technology. More than three quarters of primer pairs designed from transcript sequences produced a single clear band of expected length when tested under uniform PCR conditions. An overwhelming majority of markers were successfully amplified in multiplexes, sequenced, and their single-copy nature was confirmed. The use of transcriptome sequences for marker design alleviates the difficulties imposed by the highly repetitive nature of noncoding parts of large salamander genomes (Sun *et al.* 2012) and has already been applied in *Ambystoma* species (O'Neill *et al.* 2013). This approach should also work well in other organisms with large, poorly known genomes (Gregory *et al.* 2007; Dufresne & Jeffery 2011; Gregory 2013). The availability of an assembled transcriptome may seem a serious limitation of our method, but this difficulty should not be exaggerated. Transcriptome sequencing and de novo assembly are not a difficult task, as long as the aim is not to obtain a perfect transcriptome (Martin & Wang 2011; De Wit *et al.* 2012; Singhal 2013), and even assembly far from perfect is sufficient for designing a large number of markers. Even with one-sixth of the Illumina HiSeq lane, the amount of sequences we obtained for a single newt is sufficient to generate an assembly with hundreds of fully reconstructed transcripts containing long 3′UTRs for less than €1000 ($1300) including RNA extraction, library preparation and sequencing.

© 2013 John Wiley & Sons Ltd

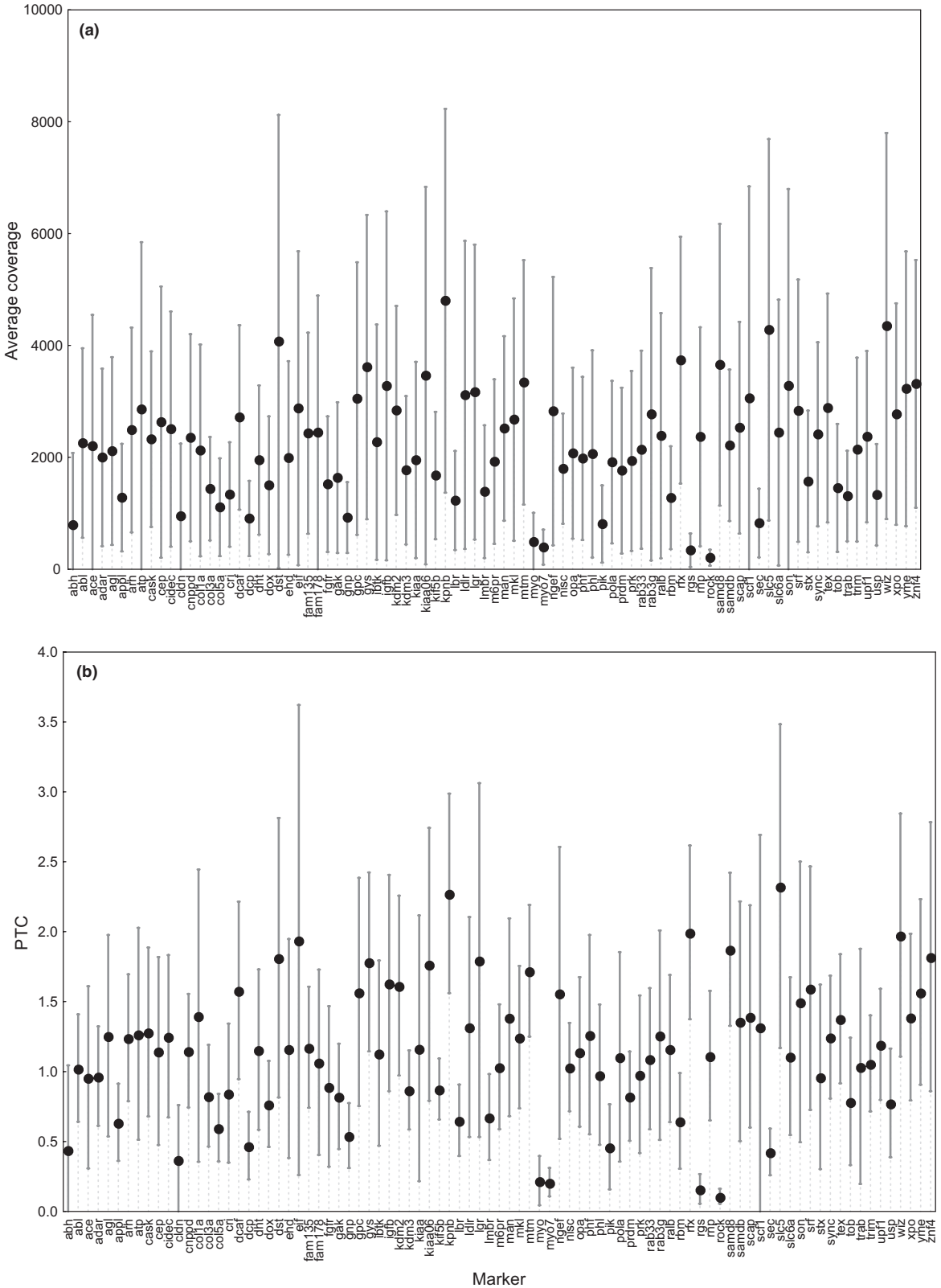

Fig. 2 Sequencing results for individual markers. (a) The per-base coverage averaged over all (*Lissotriton vulgaris*/*montandoni*) individuals, (b) percentage of the total per-base coverage (PTC) attributable to the marker averaged over all (*Lissotriton vulgaris*/*montandoni*) individuals; low PTC values indicate that the marker's coverage was poor compared with other markers. Means ± 1 SD are presented.

© 2013 John Wiley & Sons Ltd

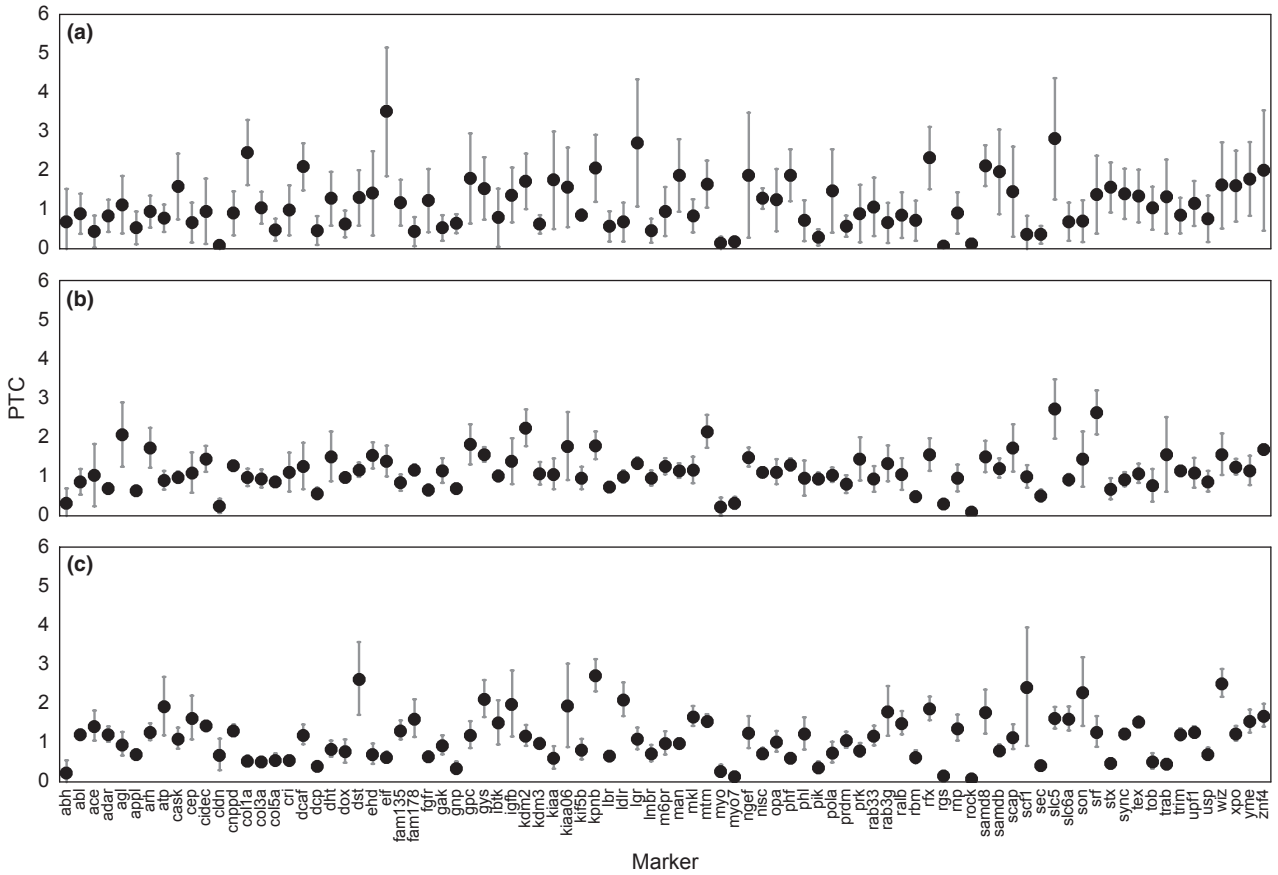

Fig. 3 Mean percentage of the total coverage (PTC) attributable to individual markers in various amplification regimes. (a) multiplexes of 3–5 loci, annealing 30 s (six individuals), (b) multiplexes of 3–5 loci, annealing 60 s (four individuals), (c) multiplexes of 9–10 loci (eight individuals), annealing 60 s. Means ± 1 SD are presented.

Our markers were located in mostly intronless 3′UTR-containing portions of transcripts, which probably contributed to the high success rate. Because 3′ UTRs of many genes are long, our method offers flexibility regarding the length of markers, and the length can be controlled at the design stage. Moreover, although 3′UTRs often contain conserved regulatory sequences, these regions are usually short (Siepel *et al.* 2005), and thus, 3′UTRs are generally more variable than coding regions (Makałowski & Boguski 1998). A potential disadvantage of 3′UTR markers is their tight linkage to protein-coding portions of genes, making them more prone to the effects of selection at linked sites. However, it is common practice to use coding genes in phylogeographic analyses (Murphy *et al.* 2001; Townsend *et al.* 2008; Shen *et al.* 2012), and most genomic regions may be to some extent affected by selection at linked sites (Charlesworth 2012). As long as selection is mainly purifying, such markers still provide valuable genealogical information. If some markers display aberrant patterns of sequence variation and divergence, these may be immediately associated with respective genes and eventually provide important evolutionary insights.

We did not notice any significant differences in marker amplification between *Lissotriton montandoni* and *L. vulgaris* or between subspecies of *L. vulgaris*. The overall design and amplification success for the outgroup species *L. italicus* and *L. boscai* were lower; however, still 55% of markers were successfully sequenced in both outgroups. If transcriptomes of several individuals from different populations are available, as it was in our case, then to reduce the incidence of null alleles polymorphisms present in primer-binding sites can be incorporated as degenerate positions. This approach may be extended to multiple species by enriching the designed markers into those useful at deeper evolutionary scales, which can be easily carried out by including several focal species during the initial screen for amplification success.

The described approach does not require any marker-specific barcoded PCR primers and employs multiplexing. Both features are attractive because they reduce the cost, labour and logistic complexity of projects. Multiplexing also reduces the amount of template

© 2013 John Wiley & Sons Ltd

DNA required, which may be an important consideration for some projects, if the amount of available tissue/extracted DNA is limiting. Another attractive aspect of our approach is that the per individual cost of library preparation remains constant regardless of the number of markers investigated. Increasing the number of markers requires more sequencing depth, but the amount of sequencing is not large as for current standards (Glenn 2011). A single batch of 96 individuals containing 100 markers of 600 bp each can be sequenced in a single 2 × 150 bp MiSeq run to the average coverage of >500 × /bp/marker/individual (assuming that about ¾ reads map to reference) at a cost of *c.* 1000 € (1300 $). Such coverage should be sufficient to obtain sequences of most markers for most samples even taking into account the observed differences in coverage depth between libraries and between markers within a library. We tested two levels of multiplexing and noticed that both gave comparable variance in coverage between markers. Our observations show that a relatively high plexity can be achieved with practically no optimization.

Sequence data were processed with standard, freely available bioinformatics tools, producing phase-resolved sequences. We demonstrate that read length is not necessarily the limiting factor, and there is no need for costly methods such as 454 sequencing, which produce several hundred bp reads, because individual reads do not need to cover the full marker length for efficient phasing. With paired-end sequencing, the length of markers for which efficient phasing is possible depends on the actual distribution of variable sites across the sequence, coverage and the distribution of library insert sizes, while the read length is less important. Physical phasing is a reliable and efficient method of reconstructing haplotypes and seems to be the best option in species with high polymorphism and high recombination rate or if single samples from divergent and/or structured populations are analysed. Physical phasing may also be combined with computational approaches (Browning & Browning 2011), for example, those implemented in PHASE (Stephens *et al.* 2001; Stephens & Scheet 2005) or HAPLOTYPER (Niu *et al.* 2002). Known haplotypes produced by physical phasing may greatly increase the accuracy of computational phasing of difficult cases (Stephens & Donnelly 2003).

The approach presented in this study provides a cost-effective, broadly applicable framework for designing, validating and high-throughput sequencing of multiple markers in nonmodel species without sequenced genomes. The use of standard, well established and actively developed bioinformatic tools, ensures the availability of state-of-the-art methods for SNP calling and genotype quality assessment. We hope that this approach will be of broad interest for researchers working at the interface of population genetics and phylogenetics, exploring deep intraspecific genetic structure, species boundaries and phylogeographies of closely related species.

Acknowledgements

We thank Maciej Pabijan, Ben Wielstra and two anonymous reviewers for their insightful comments. The work was supported by the Polish National Science Center Grants No. 8171/B/P01/2011/40 and UMO-2012/04/A/NZ8/00662 to WB and by the Jagiellonian University grant DS/WBiNoZ/INoS/762/12. M.T.S is the recipient of DOCTUS stipend.

References

Bi K, Vanderpool D, Singhal S *et al.* (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, **13**, 403.

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

Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, **12**, 703–714.

Charlesworth B (2012) The effects of deleterious mutations on evolution at linked sites. *Genetics*, **190**, 5–22.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

De Wit P, Pespeni MH, Ladner JT *et al.* (2012) The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, **12**, 1058–1067.

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, **24**, 332–340.

Depristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–501.

Dufresne F, Jeffery N (2011) A guided tour of large genome size in animals: what we know and where we are heading. *Chromosome Research*, **19**, 925–938.

Faircloth BC, McCormack JE, Crawford NG *et al.* (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

Grabherr MG, Pontiller J, Mauceli E *et al.* (2011) Exploiting nucleotide composition to engineer promoters. *PLoS ONE*, **6**, e20136.

Gregory TR (2013) *Animal genome size database*, http://www.genomesize.com.

Gregory TR, Nicol JA, Tamm H *et al.* (2007) Eukaryotic genome size databases. *Nucleic Acids Research*, **35**, D332–D338.

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, **27**, 570–580.

Hickerson MJ, Carstens BC, Cavender-Bares J *et al.* (2010) Phylogeography's past, present, and future: 10 years after Avise, 2000. *Molecular Phylogenetics and Evolution*, **54**, 291–301.

Hohenlohe PA, Day MD, Amish SJ *et al.* (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology*, **22**, 3002–3013.

Holleley CE, Geerts PG (2009) Multiplex Manager 1.0: a cross-platform computer program that plans and optimizes multiplex PCR. *BioTechniques*, **46**, 511–517.

Hong X, Scofield DG, Lynch M (2006) Intron size, abundance, and distribution within untranslated regions of genes. *Molecular Biology and Evolution*, **23**, 2392–2404.

© 2013 John Wiley & Sons Ltd

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Lemmon AR, Lemmon EM (2012) High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. *Systematic Biology*, **61**, 745–761.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Makałowski W, Boguski MS (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2820 orthologous rodent and human sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 9407–9412.

Mamanova L, Coffey AJ, Scott CE *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nature Methods*, **7**, 111–118.

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nature Reviews Genetics*, **12**, 671–682.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

McKenna A, Hanna M, Banks E *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

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

Nadachowska K, Babik W (2009) Divergence in the face of gene flow: the case of two newts (Amphibia: Salamandridae). *Molecular Biology and Evolution*, **26**, 829–841.

Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, **70**, 157–169.

O'Neill EM, Schwartz R, Bullock CT *et al.* (2013) Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology*, **22**, 111–129.

Puritz JB, Addison JA, Toonen RJ (2012) Next-generation phylogeography: a targeted approach for multilocus sequencing of non-model organisms. *PLoS ONE*, **7**, e34241.

Rousset F (2008) GENEPOP '007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Rozas J (2009) DNA sequence polymorphism analysis using DnaSP. *Methods in Molecular Biology*, **537**, 337–350.

Shen XX, Liang D, Zhang P (2012) The development of three long universal nuclear protein-coding locus markers and their application to osteichthyan phylogenetics with nested PCR. *PLoS ONE*, **7**, e39256.

Siepel A, Bejerano G, Pedersen JS *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, **15**, 1034–1050.

Singhal S (2013) De novo transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. *Molecular Ecology Resources*, **13**, 403–416.

Smith JJ, Putta S, Zhu W *et al.* (2009) Genic regions of a large salamander genome contain long introns and novel genes. *BMC Genomics*, **10**, 19.

Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, **14**, 404–414.

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, **73**, 1162–1169.

Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, **76**, 449–462.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Sun C, Shepard DB, Chong RA *et al.* (2012) LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. *Genome Biology and Evolution*, **4**, 168–183.

Tewhey R, Warner JB, Nakano M *et al.* (2009) Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotechnology*, **27**, 1025–1031.

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.

Twyford AD, Ennos RA (2012) Next-generation hybridization and introgression. *Heredity*, **108**, 179–189.

Willing EM, Hoffmann M, Klein JD, Weigel D, Dreyer C (2011) Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics*, **27**, 2187–2193.

You FM, Huo N, Gu YQ *et al.* (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics*, **9**, 253.

Zieliński P, Nadachowska-Brzyska K, Wielstra B *et al.* (2013) No evidence for nuclear introgression despite complete mtDNA replacement in the Carpathian newt (*Lissotriton montandoni*). *Molecular Ecology*, **22**, 1884–1903.

W.B. designed research, performed research, and analyzed data. P.Z. performed research, contributed new reagents or analytical tools, and analyzed data. K.D. performed research. M.T.S. and M.K. contributed new reagents or analytical tools. W.B. and P.Z. wrote the paper (with contribution from other authors).

Data Accessibility

Raw DNA sequences: NCBI BioProject PRJNA 214312.

Transcriptome contigs used for marker design, mapping (BAM) variant calling (VCF) files, alignments of sequenced markers, genotype data for *L. montandoni* from Eastern Carpathians region and custom Python script: Dryad Digital Repository entry doi:10.5061/dryad.94 ms3.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 The per-base average coverage for individuals averaged across markers. Means ± SD are presented.

Table S1 Sampling locality and amplification method used for each individual.

Table S2 Sequence markers developed from transcriptome of *Lissotriton* newts.

Table S3 Average coverage depth per marker per individual.

Table S4 Relative coverage per marker per individual expressed as a percentage of the total coverage (PTC) obtained for individual.

Table S5 Basic statistics of DNA polymorphism in *Lissotriton montandoni* and *L. vulgaris*.

© 2013 John Wiley & Sons Ltd