MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2014) **14**, 597–605 doi: 10.1111/1755-0998.12206

Applicability of RAD-tag genotyping for interfamilial comparisons: empirical data from two cetaceans

AMÉLIA VIRICEL,* ERIC PANTE,* WILLY DABIN† and BENOIT SIMON-BOUHET*

**Littoral, Environnement et Sociétés (LIENSs) UMR 7266 CNRS, Université de La Rochelle, 2 rue Olympe de Gouges, La Rochelle 17000, France, †Observatoire PELAGIS, UMS 3462 CNRS, Université de La Rochelle, Pôle analytique, 5 allées de l'océan, La Rochelle 17000, France*

Abstract

Restriction-site-associated DNA tag (RAD-tag) sequencing has become a popular approach to generate thousands of SNPs used to address diverse questions in population genomics. Comparatively, the suitability of RAD-tag genotyping to address evolutionary questions across divergent species has been the subject of only a few recent studies. Here, we evaluate the applicability of this approach to conduct genome-wide scans for polymorphisms across two cetacean species belonging to distinct families: the short-beaked common dolphin (*Delphinus delphis*; *n* = 5 individuals) and the harbour porpoise (*Phocoena phocoena*; *n* = 1 individual). Additionally, we explore the effects of varying two parameters in the `Stacks` analysis pipeline on the number of loci and level of divergence obtained. We observed a 34% drop in the total number of loci that were present in all individuals when analysing individuals from the distinct families compared with analyses restricted to intraspecific comparisons (i.e. within *D. delphis*). Despite relatively stringent quality filters, 3595 polymorphic loci were retrieved from our interfamilial comparison. Cetaceans have undergone rapid diversification, and the estimated divergence time between the two families is relatively recent (14–19 Ma). Thus, our results showed that, for this level of divergence, a large number of orthologous loci can still be genotyped using this approach, which is on par with two recent *in silico* studies. Our findings constitute one of the first empirical investigations using RAD-tag sequencing at this level of divergence and highlights the great potential of this approach in comparative studies and to address evolutionary questions.

Keywords: Delphinidae, genomics, interfamilial divergence, Phocoenidae, phylogenetics, RAD sequencing

Received 21 August 2013; revision received 8 November 2013; accepted 13 November 2013

Introduction

Recent parallel DNA sequencing technologies have enabled population genomics studies in nonmodel organisms including characterizing patterns of hybridization and introgression (e.g. Hohenlohe *et al.* 2011), intraspecific phylogeography (e.g. Emerson *et al.* 2010), QTL mapping (e.g. Gagnaire *et al.* 2013) and studying the genetic basis of adaptations (Stapley *et al.* 2010). There is now a growing interest in these methods in the fields of biogeography (Lexer *et al.* 2013) and phylogenetics (McCormack *et al.* 2013).

Among recent genotyping methods using next-generation sequencing, restriction-site-associated DNA tag (RAD-tag) sequencing has become one of the most popular approaches to conduct population genomics studies in nonmodel organisms. To date, however, few studies have explored the applicability of this approach to divergent species to address evolutionary questions at a greater phylogenetic depth. Two *in silico* studies evaluated the suitability of RAD-tag sequencing to address phylogenetic questions (Rubin *et al.* 2012; Cariou *et al.* 2013) using simulated data sets obtained from divergent reference genomes. Rubin *et al.* (2012) used genomes from three taxonomic groups (*Drosophila*, mammals and yeasts) to generate RAD-tag sequences *in silico* and, for each group, assessed whether accurate species phylogenies could be reconstructed from these sequences. Similarly, Cariou *et al.* (2013) simulated RAD-tag sequences from the genomes of 12 species of *Drosophila*, separated by different levels of divergence (5–63 Ma). Both studies suggest that (i) a sufficient number (at least hundreds) of conserved orthologous loci can be obtained even when comparing divergent species within relatively young phylogenetic groups (divergence times of up to 60 Ma), and (ii) RAD-tag loci can be phylogenetically informative and allow reconstruction of accurate species phylogenies.

Correspondence: Amélia Viricel, Fax: +33-05-46-50-76-63; E-mail: amelia.viricel@gmail.com

Few empirical studies have evaluated whether these expectations are verified by including divergent species in their RAD-taq sequencing analysis (Eaton & Ree 2013; Nadeau *et al.* 2013; Stölting *et al.* 2013), and particularly beyond intrageneric (Jones *et al.* 2013; Keller *et al.* 2013; Lexer *et al.* 2013; Wagner *et al.* 2013) or intrafamilial (Bergey *et al.* 2013) comparisons. In the present study, we assessed the applicability of RAD-tag genotyping in the upper bound of these phylogenetic depths by conducting intra- and interfamilial comparisons using two cetacean species: the common dolphin (*Delphinus delphis*, Delphinidae) and the harbour porpoise (*Phocoena phocoena*, Phocoenidae). We analysed generated RAD-tag sequences using the Stacks analysis pipeline (Catchen *et al.* 2011) and evaluated the effects of varying two Stacks parameters on the number of loci and genetic distances obtained.

Materials and methods

Tissue samples, DNA extraction and Sanger sequencing

Tissue samples were collected from six dead animals (five short-beaked common dolphins, *Delphinus delphis*, and one harbour porpoise, *Phocoena phocoena*) that were either incidentally caught in pelagic fisheries in the Celtic Sea or Bay of Biscay or stranded on the French Atlantic coast (Table 1). Tissue samples were frozen at −20 °C or stored in ethanol at room temperature. Total genomic DNA was extracted from approximately 15–25 mg of skin or kidney tissue using NucleoSpin® Tissue (Macherey-Nagel EURL, Hoerdt, France) or using DNeasy® Blood & Tissue (Qiagen, Courtaboeuf, France) kits following the manufacturer's protocols. DNA concentration was quantified using a NanoDrop™ 2000 (Thermo Scientific, Illkirch, France). DNA quality was assessed on a 1% agarose gel stained with ethidium bromide and was similar across the six samples: good (high molecular weight as well as shear) to excellent (high molecular weight only). Species identification made in the field using morphological characters was confirmed by sequencing two portions of the mitochondrial genome: (i) the 5′ end of the control region (including a portion of the flanking proline tRNA) was amplified using primers L15824 (5′-CCTCACTCCTCCCTAAGACT-3′; Rosel *et al.* 1999) and H16498 (5′-CCTGAAGTAAGAACCAGATG′-3; Rosel *et al.* 1994); and (ii) a portion of cytochrome *b* was amplified using primers L14724 (5′-TGACTTGAARAACCAYCGTTG-3′; Palumbi *et al.* 1991) and H15149 (5′-CAGAATGATATTTGTCCTCA-3′; Kocher *et al.* 1989). The polymerase chain reaction (PCR) included 10 mM Tris–HCl (pH 8.3), 50 mM KCl, 0.1 % Triton X-100, 1.5 mM $MgCl_2$, 0.3 μM of each primer, 0.15 mM dNTPs (Euromedex, Mundolsheim, France), 2 U Taq polymerase (VWR, Fontenay sous Bois, France) and 50 ng DNA in a 50 μL total volume. PCR profiles were as described in the study by Vollmer *et al.* (2011) for the L15824/H16498 primer pair and Viricel & Rosel (2012) for the L14724/H15149 primer pair. PCR products were sent to Genoscreen (Lilles, France) for purification and Sanger sequencing. Mitochondrial sequences were edited using Sequencher® v. 4.7 (Gene Codes Corp., Ann Arbor, MI, USA) and were aligned using MAFFT v. 7 with default parameters (FFT-NS-i method) (Katoh *et al.* 2002).

Genotyping by sequencing

RAD-tag libraries were prepared by Eurofins Genomics (Ebersberg, Germany) using 1–2 μg of total genomic DNA per individual and using the *Not1* restriction enzyme. Unique barcodes used to differentiate multiplexed individuals were 6–9 nucleotides long and differed by at least two nucleotides. Libraries were sequenced by Eurofins Genomics on two lanes of the Illumina® HiSeq™ 2000 platform (Illumina, Inc., San Diego, CA, USA) with the 1 × 100 base pairs (bp) single-end read module, as part of a larger *D. delphis*

Table 1 Voucher information for one harbour porpoise (*Phocoena phocoena*) and five short-beaked common dolphins (*Delphinus delphis*) used in this study. Voucher identification (ID) numbers correspond to specimen numbers from UMS Pelagis. For samples obtained from stranded animals, the geographical coordinates and date of the stranding event are given. For one common dolphin that was incidentally caught (bycatch) in the tuna fishery, geographical coordinates and date correspond to where and when the dead animal was retrieved from the gear onboard. GenBank Accession nos are given for each mitochondrial DNA portion that was sequenced: control region (CR) and cytochrome *b* (*cytb*)

Species	Voucher ID	Sample type	Sex	Latitude	Longitude	Date	CR Accession no.	*cytb* Accession no.
Phocoena phocoena	10712131	bycatch	M	44.648	−1.316	13-Apr-07	KF727592	KF727598
Delphinus delphis	10307073	stranding	F	46.713	−1.979	29-Jul-03	KF727593	KF727599
Delphinus delphis	10401011	stranding	F	44.404	−1.264	17-Jan-04	KF727594	KF727600
Delphinus delphis	10512077	bycatch	F	48.100	−9.867	3-Sep-05	KF727595	KF727601
Delphinus delphis	9902012	stranding	M	43.955	−1.363	19-Feb-99	KF727596	KF727602
Delphinus delphis	10201010	stranding	F	46.189	−1.429	22-Jan-02	KF727597	KF727603

population genomics RAD-tag sequencing project (total of 92 individuals). Raw Illumina reads were processed using the CASAVA v. 1.8.2 software (Illumina, Inc., San Diego, CA, USA). Illumina read data were de-multiplexed, quality-filtered and assembled using the Stacks tool kit v. 0.99994. A recent study by Davey *et al.* (2013) compared Stacks and RADtools (Baxter *et al.* 2011), another program to analyse RAD-tag sequences without a reference genome, and recommended the use of Stacks as it provides more features. The Stacks pipeline includes four major steps (Catchen *et al.* 2013): reads are first sorted by unique barcode to group together all sequences from each individual (de-multiplexing step) while also excluding sequences that do not pass a set quality score; second, loci are build within each individual by creating stacks of identical reads and assembling unique loci by merging stacks that differ only by a set number of nucleotides (M) to allow polymorphism within individuals; third, loci identified for each individual are compared and catalogued across all individuals, and a set number of nucleotide differences (*n*) is allowed to merge loci from different individuals in the catalog; and fourth, individual genotypes are determined for each locus. The following filters were applied during the first step of the pipeline (process_radtags): one nucleotide mismatch (i.e. one sequencing error) was tolerated within individual barcodes, sequences for which the mean raw Phred quality score dropped below 10 within a sliding window spanning 15% of the read length were discarded. Sequences were truncated to a length of 91 bp. As there is no reference genome available for *D. delphis* or *P. phocoena*, we used the denovo_map.pl program in Stacks to merge stacks (i.e. sets of identical reads) into loci within individuals and to build a catalog of loci across individuals. The minimum number of reads to form a stack (*m*) was set to 3. SNPs were detected while varying two Stacks parameters: the number of mismatches allowed between stacks to be grouped in a unique locus within an individual (*M*, set between 1 and 7), and the maximum distance between loci from distinct individuals to be merged in the population catalog (*n*, set between 1 and 8). Analyses were conducted with 14 different combinations of these parameters setting *M* and *n* at the same value or with one additional mismatch for *n* compared with *M* (see Figs 1 and 2). Highly repetitive sequences were removed or broken down using the 't' option in denovo_map.pl (Catchen, *et al.* 2011, 2013). Additionally, we verified that the final catalog did not contain dimers formed by adapters. The quality of filtered sequences obtained after denovo_map.pl was evaluated using FastQC v. 0.10.1 (Babraham Bioinformatics, http://www.bioinformatics.babraham.ac.uk/projects/). We applied the populations program from Stacks to obtain the final sets of orthologous loci: loci were retained if the locus total depth of coverage was equal or higher than 10 reads per individual, and if they were present in all individuals (i.e. no missing data allowed).

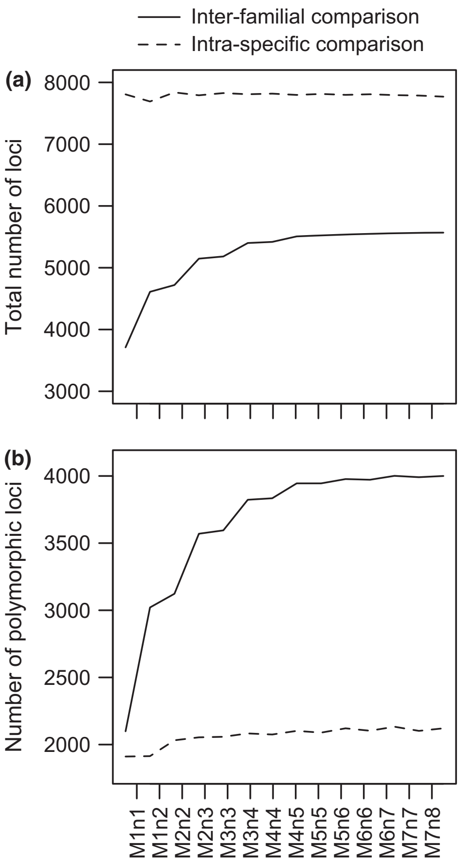

Fig. 1 Influence of Stacks parameters on (a) the total number of loci, and (b) the number of polymorphic loci obtained. Fourteen parameter combinations were evaluated for the whole data set (interfamilial comparison) and for the common dolphin only (intraspecific comparison).

Data analyses

Polymorphism and divergence statistics were calculated using the populations program in Stacks and using the ape (Paradis *et al.* 2004) and adegenet (Jombart 2008) packages in R (R Development Core Team 2013), respectively. Interindividual divergence was assessed using polymorphic sites that are either variable within individuals (in heterozygotes) or fixed within individuals (homozygotes) but variable between individuals. Due to likely heterogeneity in substitution models across loci, genetic distances were calculated as raw p-distances (i.e. proportion of fixed differences between

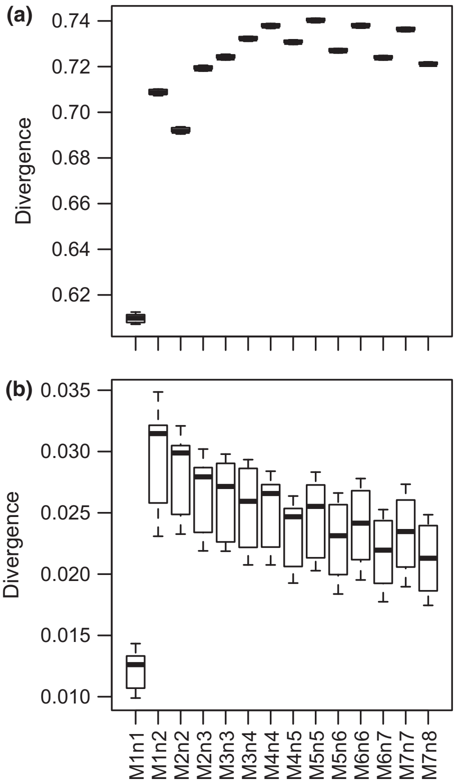

Fig. 2 Influence of Stacks parameters on interindividual sequence divergence (raw p-distances calculated using variable sites) for (a) interfamilial comparisons and (b) intraspecific comparisons. The range of interindividual distances is represented as boxplots.

two sequences). To compare intraspecific and interfamilial data, all analyses were run on (i) the five *Delphinus delphis* individuals and (ii) all six individuals (five *D. delphis* and one *Phocoena phocoena*).

Finally, for the interfamilial comparison, we explored the functions of invariable (conserved) compared with polymorphic loci, at a chosen *M* and *n* combination (*M*3*n*3; see Results section). The goal of this analysis was two-tiered: (i) to investigate whether some functions would be overrepresented in polymorphic compared with invariable loci and (ii) to assess whether putative gene functions can be retrieved from RAD-tag sequences, which could be useful in applications such as studying loci potentially under selection. Identification of RAD-tag sequences (1587 and 3574 tags for conserved and variable loci, respectively) was determined in Blast2GO v. 2.6.6 (public database of August 2012; Conesa *et al.* 2005; Conesa & Götz 2008; Götz, *et al.* 2008, 2011) using the BLASTN program (e-value cut-off of 10^{-3}, HSP cut-off of 33 Altschul, *et al.* 1990, 1997), as the BLASTX program retrieved very little results due to the short length of the corresponding amino acid sequences (<31 amino-acids). While BLASTN can be used to identify tags, it does not allow subsequent mapping and annotation (Blast2GO manual). We therefore used the results of BLASTN to retrieve the sequence of the best match between tags and GenBank sequences from the *Tursiops truncatus* genome (GenBank Bioprojects Accession nos PRJNA189944 and PRJNA20367), a species closely related to *D. delphis* (McGowen *et al.* 2009), similar to the approach employed by Reitzel *et al.* (2013) for the anemone *Nematostella vectensis*. As the percentage identity between tags and *T. truncatus* sequences was very high (84.6–100%), we proceeded to the mapping and annotation steps to obtain Gene Ontology (GO) terms from these longer *T. truncatus* sequences (222–16 700 bp, median size 1549 bp; BLASTX e-value cut-off = 10^{-3} HSP = 33, 20 hits retained; annotation settings: e-value filter = 10^{-6}, annotation cut-off = 55, GO weight = 5, no HSP-hit coverage cut-off). GO terms correspond to groups of genes involved in similar functions such as genes with products involved in cellular components. Genes can be grouped into GO terms at different levels depending on the desired level of precision in the function. Enrichment of GO terms between *T. truncatus* sequence sets corresponding to 'conserved' and 'variable' loci was tested using the Fisher's exact test as implemented in Blast2GO (GOSSIP module, Blüthgen *et al.* 2005, FDR = 0.05).

Results

Species identification of each individual was confirmed using Sanger-sequenced mitochondrial DNA. Mitochondrial sequence alignments encompassed a 425-bp portion of the control region and a 402-bp portion of cytochrome *b* (see Table 1 for GenBank Accession nos). For the five common dolphins, control region sequences were identical to haplotypes published on GenBank (i.e. 100% coverage and 100% identity), which were sequenced from other short-beaked common dolphins from the eastern North Atlantic (NA). Cytochrome *b* sequences for these individuals also supported species identification made in the field as most similar sequences in GenBank belonged to short-beaked common dolphins. For the harbour porpoise, we obtained a perfect haplotype match for the mitochondrial control region sequence, corresponding to another harbour porpoise from the eastern NA. For cytochrome *b*, there was a 1 bp difference between our sequence and a published haplotype (Accession no.: AJ554063) from a harbour porpoise complete mitogenome. The next best match in GenBank was also a harbour porpoise (Accession no.: U13143).

The two Illumina sequencing lanes produced over 2.8 million raw reads per individual (Table 2). On average, 40% of raw reads were removed by the quality filters applied (Table 2). The main reason for removing reads

Table 2 Total number of raw and filtered (i.e. after process-radtags) reads for each sample used in this study. Overall sequence quality was assessed using the mean Phred score after filters from process-radtags

Species	Voucher ID	Raw reads	Filtered reads	Mean Phred score
Phocoena phocoena	10712131	3 102 559	2 455 577	36
Delphinus delphis	10307073	3 125 400	1 614 204	36
Delphinus delphis	10401011	3 185 527	1 846 015	36
Delphinus delphis	10512077	2 814 062	2 091 276	36
Delphinus delphis	9902012	3 859 005	1 466 233	36
Delphinus delphis	10201010	3 712 290	2 105 664	35

was ambiguous barcodes, which could suggest either barcode synthesis errors or a high sequencing error rate. By setting the minimum number of reads to build a stack to 3, the impact of potential sequencing errors on the genotypes and loci we obtained should be very limited. The sequence quality of filtered reads was excellent with a minimum Phred score of 35 (Table 2).

The effect of varying the `denovo_map.pl` parameters *M* and *n* can be contrasted between the intraspecific and interfamilial data sets. For the interfamilial comparison, increasing *M* (intraindividual parameter), and particularly *n* (interindividual parameter), resulted in an increase in the total number of loci until a plateau was reached at parameter combination *M*3*n*3 (Fig. 1a; total number of loci: 5182). This outcome can be explained as follows: in the first step of the analysis (i.e. `denovo_map.pl`), increasing *M* and *n* will decrease the total number of loci present in the catalog as more distinct sequences will be merged into the same locus (Catchen *et al*. 2013). This also results in a greater depth of coverage per locus. Thus, in the second step of the analysis (`populations`), there will actually be an increase in the total number of loci that are kept in the final catalog after the filters are applied (a minimum of 10 reads per locus per individual). Additionally, as *n* is increased, more loci will be in common among all individuals, particularly when including divergent individuals such as here. When *n* is low, fixed differences will be considered as distinct loci that will not be present in all individuals. Therefore, the number of loci kept in the final catalog will also increase when *n* is increased due to the filter of the minimum number of individuals where a locus has to be present (in this study, a locus had to be present in all individuals). A similar trend can be observed for the number of polymorphic loci, which increased as *M* and *n* increased (Fig. 1b). Eventually, increasing *M* and *n* could result in overmerging loci (loci that are not orthologous). By using the set of parameters corresponding to where a plateau for the total number of loci starts, we were most likely to avoid overmerging issues. However, it is possible that overmerging is not detectable by simply observing a plateau in the total number of loci, as the overall decrease in the number of loci by overmerging could be balanced by the discovery of orthologous, yet highly divergent loci. For the intraspecific data set, the effect of varying the two parameters on the number of loci (total and polymorphic) was less striking (Fig. 1a,b). A plateau was quickly reached at the *M*2*n*2 parameter combination (total number of loci: 7838; 2032 polymorphic loci) after an initial small rise in the number of polymorphic loci (Fig. 1b). In terms of divergence, the largest change for both data sets was observed when increasing *n* from 1 to 2 (Fig. 2a,b). This is likely due to an increase in the number of variable sites that are fixed within but variable among individuals. For subsequent data description (e.g. sequence variability) and analyses (`Blast2GO`), we chose the parameter combinations where a plateau was reached in terms of number of loci, which corresponded to *M*3*n*3 for the interfamilial data set and *M*2*n*2 for the intraspecific data set.

Using these parameter combinations, we observed a 34% drop in the total number of loci when analysing all individuals (in the interfamilial comparison) compared with the intraspecific data set. This drop was calculated as the percent difference in the total number of loci found with and without the *P. phocoena* sample in the final data set. This result was not simply an effect of removing any individual from the data set, as excluding a *D. delphis* individual only resulted in a small percent difference (2.2–6.5 %) in the total number of loci. Thus, we can conclude that most of the 34% drop in number of loci was indeed due to including a more divergent individual. The proportion of variable loci was 69% and 26% for the interfamilial and intraspecific data sets, respectively. In terms of sequence polymorphism, we observed one SNP every 292 bp in the intraspecific data set (total sequence length screened: 713 255 bp) compared with one SNP every 71 bp in the interfamilial data set (total sequence length screened: 471 538 bp).

In the interfamilial comparison, sequence identification using `Blast2GO` was significantly higher for conserved tags (994/1587, 63%) than for polymorphic tags (1512/3574, 42%). `BLASTN` searches were reliable, with e-values ranging from 10^{-5} to 10^{-38} for the best hit. Similarity between query and subject sequences of the best hit was high (73–100%, median 100% for conserved tags; 71–100%, median 98% for polymorphic tags). Top species hits included sequences from killer whale (*Orcinus orca*: 774 hits), common bottlenose dolphin (*Tursiops truncatus*: 716 hits), human (*Homo sapiens*: 234 hits) and other mammals.

A majority of the best hit sequences from *T. truncatus* could be mapped and annotated (conserved: 91%; variable: 92%). There was no significant difference in the number of *T. truncatus* sequences per GO term between variable versus conserved loci across the two cetacean families (Fisher's exact test with FDR = 0.05; Fig. S1, Supporting information).

Discussion

Modern cetaceans constitute a recent group, which originated approximately 34–35 Ma (Fordyce 1980; Arnason *et al.* 2004) and comprise 14 extant families (Perrin *et al.* 2009). Thus, based on the two *in silico* studies by Rubin *et al.* (2012) and Cariou *et al.* (2013), a large number of conserved orthologous loci should be obtained when comparing species from distinct cetacean families. Indeed, our study confirms this expectation, as only 34% of the loci present, at this sequencing effort, in all common dolphins were lost when including an individual from a distinct cetacean family. The divergence time separating these two families (Delphinidae and Phocoenidae) has been estimated between 14 and 19 Ma based on fossil-calibrated molecular clocks (Arnason *et al.* 2004; McGowen *et al.* 2009; Xiong *et al.* 2009). Comparatively, the proportion of loci that were lost (34%) in our interfamilial comparison was lower than the percentage of loci lost (60%) between two divergent *Drosophila* species pairs, which have been separated for a similar period of time (ca 13 My) (Cariou *et al.* 2013). At similar divergence times, the loss of orthologous loci will depend on the rate of molecular evolution, which varies between taxonomic groups (Britten 1986; Martin & Palumbi 1993). Indeed, *Drosophila* has a high nucleotide substitution rate (e.g. Britten 1986; Chan *et al.* 2012) compared with cetaceans, which generally display slow rates of molecular evolution (Kingston & Rosel 2004; Bininda-Emonds 2007; McGowen *et al.* 2012).

We explored the effects of varying two parameters in the `denovo_map.pl` program of `Stacks` on the number of loci and level of divergence obtained. A plateau was reached, after which the number of loci did not change dramatically. Our results are comparable to those of Keller *et al.* (2013) who observed a decrease in the total number of loci obtained when increasing *M* and *n*, prior to applying filters, and an increase in the number of polymorphic loci after filters were applied. The level of sequence variability we observed in intraspecific comparisons (within *D. delphis*: one SNP every 292 bp) was comparable to the diversity previously observed in *Delphinus* spp. or in closely related species. Thus, Amaral et al. (2010) screened 6537 bp in 17 *Delphinus* spp. individuals and reported a SNP every 272 bp. For the common bottlenose dolphin (*Tursiops truncatus*), Vollmer & Rosel (2012) observed one SNP every 463 bp (total sequence length screened: 70 828 bp in 10 individuals). Note that the five common dolphins analysed here were from the eastern NA, and possibly from the same population. Greater sequence variability would be expected if individuals from distinct regions or populations were analysed as in the study by Vollmer & Rosel (2012) and Amaral *et al.* (2010). A source of sequence variability that we did not consider here is the occurrence of indels. `Stacks` does not allow for indels, and these sequences would appear as distinct loci in our analysis and would not pass the filter of being present in all individuals. Therefore, we may have lost some loci if indels were present. Among alternative pipelines that have been developed to analyse RAD-tag data in absence of a reference genome, `PyRAD` (Eaton & Ree 2013) does accommodate indels. `PyRAD` is based on sequence similarity and alignment, rather than a set number of nucleotide differences.

The number of polymorphic loci and sequence variability (one SNP every 71 bp) observed in our interfamilial comparison outlines the potential benefits of RAD-tag sequencing to solve phylogenetic questions within a group that diversified in multiple rapid and recent radiation events (Steeman *et al.* 2009). Recently, analysis of amplified fragment length polymorphisms (AFLPs) has provided new insights into the phylogeny of cetaceans (Kingston & Rosel 2004; Kingston *et al.* 2009). However, these markers are dominant and anonymous. One advantage of RAD-tag sequencing compared with the approach above is that it provides codominant sequence data, which can be potentially identified and annotated using published databases (e.g. Scaglione *et al.* 2012). Our `Blast2GO` results suggest that insights into putative function can be gained by comparing short (<100 bp) RAD-tag sequences to published sequences. However, GO terms associated with RAD sequences could not be obtained directly using `BLASTX` due to the short length of the corresponding amino acid sequences, and `Blast2GO` does not produce GO terms when `BLASTN` is applied. Thus, obtaining GO terms was achieved indirectly by relying on `BLASTN` hits from the reference genome of a closely related species (*T. truncatus*) and applying `BLASTX` on these longer sequences. While we limited our phylogenetic analyses to comparing GO terms between conserved and polymorphic loci, and calculating genetic distances, `Stacks` produces outputs that allow to conduct other analyses widely used in phylogeny or phylogeography such as building phylogenetic trees and running cluster analyses.

One limitation of RAD-tag sequencing for phylogenetic inferences is that the number of loci is expected to decrease as more taxa are added to the data set (as seen in Lexer *et al.* 2013), which will limit to some extent the size of the phylogenetic data set. Thus, there will be a trade-off between the number of taxa and the total number of orthologous loci analysed. One way to alleviate this issue could be to use a RAD-tag double-digest approach (Peterson *et al.* 2012), which would increase locus representation across individuals. Very recently, new high-throughput genomic sequencing approaches have been developed to specifically target phylogenomics and phylogeographical questions (Carstens *et al.* 2012; Lemmon & Lemmon 2013). These new laboratory methods should be complementary to applications of RAD-tag data. A first approach, developed by Lemmon *et al.* (2012), is based on a sequence capture technique, which relies on probes designed using sequenced reference genomes. This approach, termed *anchored enrichment*, can provide several hundreds of loci, in the form of sequence data, for potentially hundreds of individuals from model and nonmodel organisms and should be applicable at different phylogenetic depths. An advantage of this approach is that it should be applicable to degraded samples. To date, however, its applicability to recently and rapidly evolved groups has not been empirically assessed yet. A second approach was designed to investigate relationships at greater phylogenetic depths by targeting ultraconserved elements (Faircloth *et al.* 2012; McCormack *et al.* 2012). While RAD-tag sequencing may be more appropriate for questions related to species delimitation and phylogeography in rapidly and recently diverged groups (e.g. cetaceans), other approaches such as *anchored enrichment* should be used when the phylogenetic depth in question reaches the limits of utility of RAD-tag data (for a full comparison of these methods, see review by Lemmon & Lemmon 2013). In the near future, we should gain a better sense of the limits of each approach as studies implementing these new methods accumulate.

In conclusion, our empirical study supports expectations that the applicability of RAD-tag genotyping is not limited to closely related species. Using two mammalian species from distinct, but recently evolved, families, we showed that this approach holds great promise for evolutionary studies conducted at this phylogenetic level.

Acknowledgements

We thank the following people and institutions for collecting and providing tissue samples: Oliver van Canneyt, Fabien Demaret and Ghislain Dorémus (UMS Pelagis) and the French stranding network (RNE). DNA samples were prepared at the Molecular Core Facility at the University of La Rochelle. We thank Vanessa Becquet (University of La Rochelle) for laboratory assistance. We would like to acknowledge Bruce Deagle and three anonymous reviewers for constructive comments on the manuscript. The University of La Rochelle supercomputer 'YMIR' was used to run `Stacks` and was partly funded by the European Regional Development Fund. We thank Mikael Guichard for his help with 'YMIR'. This work was funded by LIENSs; salaries for AV and EP were covered by a grant to the Poitou-Charentes region (*Contrat de Projet État-Région 2007–2013*), by a grant from the *Fond Européen de Développement Régional* (EP) and by *Actions internationales et rayonnement* of the University of La Rochelle (AV).

References

Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Altschul SF, Madden TL, Schäffer AA *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Amaral AR, Silva MC, Möller LM, Beheregaray LB, Coelho MM (2010) Anonymous nuclear markers for cetacean species. *Conservation Genetics*, **11**, 1143–1146.

Arnason U, Gullberg A, Janke A (2004) Mitogenomic analysis provide new insight into cetacean origin and evolution. *Gene*, **333**, 27–34.

Baxter SW, Davey JW, Johnston JS *et al.* (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a nonmodel organism. *PLoS ONE*, **6**, e19315.

Bergey CM, Pozzi L, Disotell TR, Burrell AS (2013) A new method for genome-wide marker development and genotyping holds great promise for molecular primatology. *International Journal of Primatology*, **34**, 303–314.

Bininda-Emonds ORP (2007) Fast genes and slow clades: comparative rates of molecular evolution in mammals. *Evolutionary Bioinformatics*, **3**, 59–85.

Blüthgen N, Brand K, Cajavec B, Swat M, Herzel H, Beule D (2005) Biological profiling of gene groups utilizing gene ontology. *Genome Informatics*, **16**, 106–115.

Britten RJ (1986) Rates of DNA sequence evolution differ between taxonomic groups. *Science*, **231**, 1393–1398.

Cariou M, Duret L, Charlat S (2013) Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecology and Evolution*, **3**, 846–852.

Carstens B, Lemmon AR, Lemmon EM (2012) The promises and pitfalls of next-generation sequencing data in phylogeography. *Systematic Biology*, **61**, 713–715.

Catchen JM, Amores A, Hohenlohe P, CreskoW, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3 (Bethesda)*, **1**, 171–182.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Chan AH, Jenkins PA, Song YS (2012) Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1003090.

Conesa A, Götz S (2008) Blast2go: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics*, **2008**, Article ID 619832.

Conesa A, Götz S, García-Gömez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML (2013) Special features of RAD sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

Eaton DAR, Ree RH (2013) Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). *Systematic Biology*, **62**, 689–706.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16196–16200.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.

Fordyce RE (1980) Whale evolution and Oligocene Southern Ocean environments. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **31**, 319–336.

Gagnaire PA, Normandeau E, Pavey SA, Bernatchez L (2013) Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (*Coregonus clupeaformis*). *Molecular Ecology*, **22**, 3036–3048.

Götz S, Garciá-Gómez JM, Terol J *et al.* (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, **36**, 3420–3435.

Götz S, Arnold R, Sebastián-León P *et al.* (2011) B2G-FAR, a species-centered GO annotation repository. *Bioinformatics*, **27**, 919–924.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11** (Suppl 1), 117–122.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Jones JC, Fan S, Franchini P, Schartl M, Meyer A (2013) The evolutionary history of *Xiphophorus* fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Molecular Ecology*, **22**, 2986–3001.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

Keller I, Wagner CE, Greuter L *et al.* (2013) Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Molecular Ecology*, **22**, 2848–2863.

Kingston SE, Rosel PE (2004) Genetic differentiation among recently diverged delphinid taxa determined using AFLP markers. *Journal of Heredity*, **95**, 1–10.

Kingston SE, Adams LD, Rosel PE (2009) Testing mitochondrial sequences and anonymous nuclear markers for phylogeny reconstruction in a rapidly radiating group: molecular systematics of the Delphininae (Cetacea: Odontoceti: Delphinidae). *BMC Evolutionary Biology*, **9**, 245.

Kocher TD, Thomas WK, Meyer A *et al.* (1989) Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 6196–6200.

Lemmon EM, Lemmon AR (2013) High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 19.1–19.23.

Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, **61**, 727–744.

Lexer C, Mangili S, Bossolini E *et al.* (2013) 'Next generation' biogeography: towards understanding the drivers of species diversification and persistence. *Journal of Biogeography*, **40**, 1013–1022.

Martin AP, Palumbi SR (1993) Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 4087–4091.

McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Research*, **22**, 746–754.

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.

McGowen MR, Spaulding M, Gatesy J (2009) Divergence date estimation and a comprehensive molecular tree of extant cetaceans. *Molecular Phylogenetics and Evolution*, **53**, 891–906.

McGowen MR, Grossman LI, Wildman DE (2012) Dolphin genome provides evidence for adaptive evolution of nervous system genes and a molecular rate slowdown. *Proceedings of the Royal Society B-Biological Sciences*, **279**, 3643–3651.

Nadeau NJ, Martin SH, Kozak KM *et al.* (2013) Genome-wide patterns of divergence and gene flow across a buttery radiation. *Molecular Ecology*, **22**, 814–826.

Palumbi S, Martin A, Romano S, McMillan W, Stice L, Grabowski G (1991) *The Simple Fool's Guide to PCR*. version 2.0. Tech. Rep., University of Hawaii, Honolulu.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Perrin W, Würzig B, Thewissen J, (eds) (2009) *Encyclopedia of Marine Mammals*, 2nd edn. Academic Press, San Diego.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and nonmodel species. *PLoS ONE*, **7**, e37135.

R Development Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM (2013) Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular Ecology*, **22**, 2953–2970.

Rosel PE, Dizon AE, Heyning JE (1994) Genetic-analysis of sympatric morphotypes of common dolphins (genus *Delphinus*). *Marine Biology*, **119**, 159–167.

Rosel PE, Tiedemann R, Walton M (1999) Genetic evidence for limited trans-Atlantic movements of the harbor porpoise *Phocoena phocoena*. *Marine Biology*, **133**, 583–591.

Rubin BER, Ree RH, Moreau CS (2012) Inferring phylogenies from RAD sequence data. *PLoS ONE*, **7**, e33394.

Scaglione D, Acquadro A, Portis E, Tirone M, Knapp SJ, Lanteri S (2012) RAD tag sequencing as a source of SNP markers in *Cynara cardunculus* L. *BMC Genomics*, **13**, 3.

Stapley J, Reger J, Feulner PDG *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, **25**, 705–712.

Steeman ME, Hebsgaard MB, Fordyce RE, *et al.* (2009) Radiation of extant cetaceans driven by restructuring of the oceans. *Systematic Biology*, **58**, 573–585.

Stölting KN, Nipper R, Lindtke D *et al.* (2013) Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Molecular Ecology*, **22**, 842–855.

Viricel A, Rosel PE (2012) Evaluating the utility of cox1 for cetacean species identification. *Marine Mammal Science*, **28**, 37–62.

Vollmer NL, Rosel PE (2012) Developing genomic resources for the common bottlenose dolphin (*Tursiops truncatus*): isolation and characterization of 153 single nucleotide polymorphisms and 53 genotyping assays. *Molecular Ecology Resources*, **12**, 1124–1132.

Vollmer NL, Viricel A, Wilcox L, Moore MK, Rosel PE (2011) The occurrence of mtDNA heteroplasmy in multiple cetacean species. *Current Genetics*, **57**, 115–131.

Wagner CE, Keller I, Wittwer S *et al.* (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, **22**, 787–798.

Xiong Y, Brandley MC, Xu S, Zhou K, Yang G (2009) Seven new dolphin mitochondrial genomes and a time-calibrated phylogeny of whales. *BMC Evolutionary Biology*, **9**, 20.

A.V. and B.S. designed the research; W.D. provided tissue samples and voucher information; A.V. conducted laboratory work; A.V. and E.P analysed the data; A.V. and E.P wrote the manuscript.

Data Accessibility

Mitochondrial DNA sequences have been submitted to GenBank (see Table 1 for Accession nos). Demultiplexed and filtered (i.e. after process-radtags) sequences (.fq files), R and Stacks codes, and Blast2GO output files were deposited in Dryad (doi: 10.5061/dryad.mk364).

Supporting Information

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

Figure S1 Frequency distribution of Gene Ontology (GO) categories retrieved for *Tursiops truncatus* sequences that best matched conserved and variable RAD-tag loci (interfamilial comparisons).