Optimization of ddRAD-like data leads to high quality sets of reduced representation single copy orthologs (R2SCOs) in a sea turtle multi-species analysis.

MAXIMILIAN DRILLER[1,2], SIBELLE TORRES VILAÇA[1,2,#], LARISSA SOUZA ARANTES[1,2], TOMÁS CARRASCO-VALENZUELA[1,2,3], FELIX HEEGER[1,4], DAMIEN CHEVALLIER[5], BENOIT DE THOISY[6,7], AND CAMILA J MAZZONI[1,2*]

[1]Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Berlin, Germany
[2]Evolutionary Genetics Department, Leibniz-Institut für Zoo- und Wildtierforschung (IZW), Berlin, Germany
[3]Universität Potsdam, Brandenburg, Potsdam, Germany
[4]Department Materials and Environment, Federal Institute for Material Research and Testing, Berlin, Germany
[5]Université de Strasbourg, CNRS, Strasbourg, France
[6]Institut Pasteur de la Guyane, Cayenne, French Guiana, France
[7]Kwata NGO, Cayenne, French Guiana, France
[#]current address: Universitá degli studi di Ferrara, Italy

*Corresponding author: mazzoni@izw-berlin.de

Short running title: Reduced-representation single copy orthologs

Abstract

Reduced representation libraries present an opportunity to perform large scale studies on non-model species without the need for a reference genome. Methods that use restriction enzymes and fragment size selection to help obtain the desired number of loci - such as ddRAD - are highly flexible and therefore suitable to different types of studies. However, a number of technical issues are not approachable without a reference genome, such as size selection reproducibility across samples and coverage across fragment lengths. Moreover, identity thresholds are usually chosen arbitrarily in order to maximize the number of SNPs considering arbitrary parameters. We have developed a strategy to identify *de novo* a set of reduced-representation single-copy orthologs (R2SCOs). Our approach is based on overlapping reads that recreate original fragments and add information about coverage per fragment size. A further *in silico* digestion step limits the data to well covered fragment sizes, increasing the chance of covering the majority of loci across different individuals. By using full sequences as putative alleles, we estimate optimal identity thresholds from pairwise comparisons. We have demonstrated our full workflow with data from five sea turtle species. Locus numbers were similar across all species, even at increasing phylogenetics distances. Our results indicated that sea turtles have in general very low levels of heterozygosity. Our approach produced a high-quality set of reference loci, eliminating a series of biological and experimental biases that can strongly affect downstream analysis, and allowed us to explore the genetic variability within and across sea turtle species.

Keywords: ddRAD, single-copy loci, non-model species, sea turtles, high-throughput sequencing, RAD pipeline.

1. INTRODUCTION

Reduced representation sequencing (RRS) has become very popular in the last few years among scientists studying the genetic variation of non-model organisms (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Beichman, Huerta-Sanchez, & Lohmueller, 2018; Narum, Buerkle, Davey, Miller, & Hohenlohe, 2013), with special emphasis on digestion-based techniques, here commonly referred to as

RAD (restriction-site-associated DNA). Coupled with high-throughput sequencing technologies, RAD libraries produce sequencing data covering thousands to millions of short (<1000 bp) loci throughout the entire genome that can be replicated in large amounts of individuals, within and between related species. The various protocols (reviewed in Andrews et al., 2016) allow for different levels of genome coverage, and are differently suited to studies involving population genetics, phylogenetics, linkage mapping, genomic scans and association mapping. Several of the RAD flavours, such as the double-digest RAD (ddRAD, Peterson, Weber, Kay, Fisher, & Hoekstra, 2012) and other approaches alike (e.g. 3RAD, ddGBS), generate fixed-length (FL) fragments that must start and end at a recognition site of one of the restriction enzymes utilized. The FL (here generalized as ddRAD-like) protocols offer high flexibility in terms of number of loci targeted and have seen an explosion of publications using the various methods (see Campbell, Brunet, Dupuis, & Sperling, 2018 for an etymological analysis of publications involving reduced-representation methods).

The high flexibility of ddRAD-like protocols relies on the combination of the utilized pair of enzymes and the fragment size selection. Once enzymes have been chosen, researchers can still opt for a specific number of loci, as the range of fragment lengths selected will greatly dictate the extension of genome coverage. However, not only the number of loci may vary, but the complete locus repertoire will change across different ranges of fragment lengths. It has been shown that the set of loci covered can be affected by slight changes in library building protocols (DaCosta & Sorenson, 2014) or by simply including individual libraries in different pools during size selection (Franchini, Monné Parera, Kautt, & Meyer, 2017). Large sensitivity to changes in the protocol affects reproducibility and may not only increase costs but also prevent entire datasets from being comparable. Although a few approaches have been suggested in attempts to homogenize the outcome of library building across many specimens (e.g. Franchini et al., 2017), the type of data generated and the lack of a reference genome prevent researchers from understanding what issues may have caused a drop in individual coverage and/or in the set of loci overlapping among samples.

Bioinformatic pipelines commonly used for ddRAD data analysis include Stacks (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013), ipyrad (Eaton & Overcast, 2020) and dDocent (Puritz, Hollenbeck, & Gold, 2014), among several others (LaCava et al., 2020). Such pipelines are usually compared and evaluated on the basis of the amounts of loci and variation (i.e. single nucleotide polymorphisms, SNPs) they output and parameters are usually tested in values and combinations arbitrarily chosen (Shafer et al., 2017; Díaz-Arce & Rodríguez-Ezpeleta, 2019). Shafer et al. (2017) have shown that changing parameters and pipelines may strongly affect downstream analysis and hypothesized that the biggest differences among methods is caused by the crucial locus definition step. It remains, however, difficult to assess the actual causes for differences among pipelines, since primary data is rarely analysed from a close perspective. Although theoretically possible, even when good (i.e. closely related and well assembled) reference genomes are available, no proper evaluation of the general coverage and differences among individuals is performed. Another problematic fact is that often no deeper analysis on the effects of the chosen parameters are performed during the initial stages of locus catalog building, as recommended for example by Paris, Stevens and Catchen (2017).

Perhaps the most attractive feature of ddRAD is the fact that no reference genome is needed to perform analyses, as a reference locus catalog can be built directly from the data. However, using a reference genome can help improve SNP calling, depending on the genetic distance to the species analyzed (Paris et al., 2017; Shafer et al., 2017). Even if references can theoretically help SNP calling, draft genomes are prone to incompleteness and high levels of mis-assemblies due to - among other reasons - haplotype divergence in homologous regions (Guan et al., 2020) and extensive repetitive regions (Phillippy, 2017). Moreover, it is still largely unknown how genome assembly quality can affect RAD data analysis.

Regardless of the approach chosen - whether *de novo* or reference-based - to analyse RAD data, the main goal is to ensure that variation among individuals is retrieved from comparisons among orthologous loci (or true homologues), in contrast to paralogous loci. A key challenge to identify true homologous alleles is the definition of identity thresholds between sequences, which should be established through an empirically-justifiable protocol (McCartney-Melstad, Gidiş, & Shaffer, 2019). If thresholds are too high, they will over-split orthologous alleles, while too low thresholds will cluster paralogous alleles into the same putative locus. The threshold definition step is considered as a crucial step for *de novo* protocols (Ilut, Nydam, & Hare, 2014), and could also potentially be used to help evaluate the quality of draft genomes. In any case, if loci are mis-identified, the over- or under-clustering of alleles will affect observed heterozygosity and consequently results of downstream analysis that depend on genotype frequencies.

In this study, we present a simple but effective solution to overcome issues in ddRAD-like data that can be transferred to virtually any given species or population. By using overlapping paired-end reads, we allow data to be analyzed using haplotypes that represent the original DNA fragments, producing important information on the fragment length and its associated coverage. Our pipeline can be used to overcome the lack of an appropriate reference genome or to evaluate the reference available. Our approach scrutinizes, in a stepwise fashion, the library building process in the lab as well as the bioinformatics behind the locus catalog building. The locus catalog - here also referred to as the set of reduced-representation single-copy orthologs (R2SCOs) - is built based on a controlled size range and coverage per length. It can serve as a reference for SNP calling, be used for marker development, for inter-specific analysis or even for full haplotype analysis. We have also developed a new approach to define two identity thresholds for any species based on the generated data. We have obtained data for five different sea turtle species and compared each set of R2SCOs to loci identified via an *in silico* analysis of the green sea turtle reference genome. By performing the same exact procedure for different sea turtle species presenting variable genetic distances and splitting times as deep as 100 million years (Naro-Maciel, Le, FitzSimmons, & Amato, 2008), we also analyse the inter-specific potentiality of ddRAD and briefly evaluate the use of draft genomes as references.

2. METHODS

Sample collection

Tissue samples were obtained for five out of the seven sea turtle (superfamily Chelonioidea) species: the leatherback turtle *Dermochelys coriacea*, the green turtle *Chelonia mydas*, the olive ridley turtle *Lepidochelys olivacea*, the loggerhead turtle *Caretta caretta* and the hawksbill turtle *Eretmochelys imbricata*. Species names were abbreviated as Dc, Cm, Lo, Cc and Ei, respectively. The tribe Carettini is represented by Lo, Cc and Ei, and the family Cheloniidae by all three Carettini species plus Cm. Dc is the only representative of the family Dermochelyidae. For each species, two samples were selected (total n=10) from southwestern Atlantic nesting females - with one exception - from areas that were previously shown to belong to different genetic pools: *D. coriacea* (Martinique and Espírito Santo State in Brazil), *C. mydas* (French Guiana and Fernando de Noronha Archipelago in Brazil), *E. imbricata* (Rio Grande do Norte and Bahia States from Brazil), *L. olivacea* (French Guiana and Sergipe State in Brazil) and *C. caretta* (Bahia State in Brazil and Rio Grande Elevation feeding area). Nine out of the ten samples were obtained in nesting beaches and should therefore represent the population of origin.

The only sample coming from a feeding area (*C. caretta* from Rio Grande Elevation) presented a mitochondrial haplotype typical from the Indo-Pacific Ocean (Reis et al., 2010). Samples were collected/transported under SISBIO permit 37499-2. Tissue samples were exported from Brazil under CITES permit 14BR015253/DF and imported into Germany under CITES permit E-03346/14. Samples from French Guiana were transported into Germany under institutional CITES between the Leibniz Institute for Zoo and Wildlife Research and Institut Pasteur in French Guiana.

Preliminary *in silico* tests and reference genome

A draft genome of *C. mydas* (CheMyd_1.0, GenBank accession number GCA_000344595.1; Wang et al., 2013) was used to perform preliminary estimations of fragment yield for different enzyme combinations and size ranges before designing the study, using a dedicated python script (RAD_digestion.py) that performs *in silico* digestions according to the ddRAD workflow (Peterson et al., 2012). We have chosen the enzyme pair EcoRI (a 6-bp cutter) and MseI (a 4-bp cutter) based on the number of expected genome fragments (20,000-30,000 loci) within a size range of 400-500 bp.

For analyses performed after the sequencing data was obtained, we have utilized the version of CheMyd_1.0 genome scaffolded by the DNAzoo project (Dudchenko et al., 2017), identified here as CheMyd_1.0_DNAzoo.

Library preparation

DNA extractions were performed using the DNeasy Blood and Tissue kit (Qiagen). The preparation of the ddRAD libraries followed the protocol of Peterson et al. (2012) with modifications of adapter sequences according to Meyer and Kircher (2010). In brief, 1µg of genomic DNA (20-100 ng/µl) was digested with EcoRI and MseI at 37 °C for at least two hours. The ligation of adapters was performed immediately after the digestion. The P5 adapter had inline barcodes of varying sizes (5 to 9 bp) at the restriction site positions and P7 adapter was designed with one of the strands lacking the indexing primer complementary region. Ligated samples were cleaned with 1.8X CleanPCR magnetic beads from CleanNA. A ten-cycle indexing PCR was performed independently for each individual using one of the 50 indexes described by Meyer and Kircher (2010) at the P7 adapter. The PCR was cleaned with 0.8X CleanPCR magnetic beads and concentration measured with Qubit 2.0 using the dsDNA HS assay (Life Technologies) and checked in the Bioanalyzer (Agilent). The indexed libraries were equimolarly pooled before the size selection step.

Size selection and sequencing

The size selection step was performed using either the PippinPrep with a 2% cassette and K2 marker or the BluePippin with a 1.5% cassette and the R2 marker (Sage Science). The initial sequencing runs showed a shift in the size selection among different pools and runs (Fig. S1), and therefore the size range selection was extended for the two last runs to ensure overlapping. Four MiSeq runs were performed with size selections of 495-605, 490-610, 450-650 and 450-650 bp, respectively. Between 129 bp and 134 bp represented adapter sequences, which means that the longest internal fragments should reach ~520 bp. Pool amplification was avoided after size selection, and only performed in one run (Run2) with fewer than 10 cycles and in five independent reactions per library, which were pooled and cleaned. The final libraries were characterized with a qPCR using the KAPA SYBR® FAST (Kapa Biosystems) kit and also checked with the Bioanalyzer. The libraries were run on the in-house Illumina MiSeq using the 600-cycle v3 kit.

Preliminary contamination and homology analysis

Samples were demultiplexed into pools using the Illumina MiSeq Reporter (v2.6.2.1), based on the P7 index. The pools were then demultiplexed into individual samples using the P5 inline barcodes, with Flexbar v.3.0.3 (Roehr, Dieterich, & Reinert, 2017) allowing no mismatches (parameters: -be LEFT_TAIL -u 3). In order to run a preliminary test for major contaminants, a subsample of 50,000 paired-end reads from each individual library was compared against two turtle genomes. For this purpose, the paired-end reads were first mapped against CheMyd_1.0_DNAzoo using bowtie2 v.2.3.0 (Langmead, Trapnell, Pop, & Salzberg, 2009) with parameters adjusted to reach a minimum of ~80% identity (parameters: --mp 10 --score-min L,-1,-2.0 --no-unal). Unmapped read pairs were compared against CheMyd_1.0_DNAzoo using blastn as implemented in NCBI's BLAST+ package (v.2.6.0) with a maximum e-value of 10^{-20} and subsequently against the freshwater turtle *Chrysemys picta* (Shaffer et al., 2013; Chrysemys_picta_bellii-3.0.3, NC_024218.1). Finally, reads that did not produce matches were aligned against the GenBank nucleotide (nt) database using the same parameters. The output from blast was visualized using MEGAN v.6.8.18 (Huson et al., 2018).

Sequence preprocessing

The first step performed with the fastq data was the phiX control library cleaning, as it was spiked into every run. Each demultiplexed pool was mapped to an *Enterobacteria phage* phiX174 reference genome (NC_001422.1) using bowtie2 with default parameters. All read pairs that mapped concordantly were removed from the sample. Subsequently, PEAR v.0.9.11 (Zhang, Kobert, Flouri, & Stamatakis, 2014) was used to merge the paired-end reads (parameters: -v 30 -n 50). The RAD_digestion.py script was used to redigest the samples according to ddRAD (options: --dd --rad --q) to account for any undigested or chimeric sequences. To remove non-targeted loci derived from star activity of the enzymes, sequences with incorrect restriction sites at either end of the locus (MseI and EcoRI, respectively) were discarded using the script checkRestrictionSites.py. Finally, sequences with average Phred quality score below 20 were discarded using Trimmomatic v.0.3.6 (Bolger, Lohse, & Usadel, 2014). The read preprocessing workflow is shown in Fig. S2.

Chimera estimation

Merged sequences including internal restriction sites could represent either undigested sequences due to low efficiency of restriction enzymes or the ligation of independent digested fragments into chimeric sequences during the adapter ligation step. In order to distinguish both cases, fragments generated from *in silico* digested sequences were analyzed in comparison to each other. The details of the workflow are found in the Supporting Information and Fig. S3.

In silico size selection

The *in-silico* size selection of preprocessed reads was based on the widest size range presenting at least 20x estimated coverage (based on the genome digestion, Fig. S4) across all ten individuals, while avoiding very high peaks of coverage that could represent highly repetitive paralogous loci. The range and the best run per individual were selected concomitantly.

Definition of the R2SCO term

We have designed a locus catalog pipeline that is based on two of the most utilized pipelines for RAD data - Stacks and ipyrad - but that incorporates information about coverage from merged sequences based on fragment size and uses the entire sequence as a putative allele. At the end of this pipeline, loci with accepted genotypes are considered as the set of reduced-representation single-copy orthologs (or R2SCOs, pronounced "Artuscos"). The idea to form a set of R2SCOs was inspired on BUSCO (Benchmarking Universal Single-Copy Orthologs, Seppey, Manni, & Zdobnov, 2019), but instead of representing orthologous single-copy coding genes among species, an R2SCO represents a set of orthologous loci produced by a certain combination of enzymes and a specific size range. R2SCOs from this study, for example, are defined as R2SCO-MseI-EcoRI-384-448, in which the last two numbers represent the *in-silico* size selection range.

R2SCO pipeline

A general scheme of the R2SCO pipeline is depicted in Figure 1. The first two general steps of the pipeline are *(i)* the reads pre-processing (Figure S1) and *(ii)* the *in-silico* size selection based on the coverage per fragment size estimated from the merged sequences. We have also established five different thresholds to be used in the pipeline, all based on preliminary analyses of the data. The thresholds and pipeline steps are explained in detail in the Supporting Information and briefly described below:

- *Minimum count per putative allele* (*m*): the minimum coverage per unique sequence within an individual that should represent a putative allele.
- *Minimum locus coverage* (COV_{min}): the minimum coverage per locus within an individual to yield reliable genotypes. It is not used for defining R2SCOs since a locus can be defined by a single allele.
- *Maximum locus coverage* (COV_{max}): limit between the distribution of sequence coverages of single-copy loci and outlier coverages that mostly represent paralogs.
- *Ancestral identity threshold* (T_{ances}): the ancestral threshold represents a minimum identity that connects recent paralogous loci. It can be based on the distance between the most distant samples or in case of an intra-specific analysis, it can be a conservative value that would avoid the successful mapping of sequences to a closely-related paralogous locus (e.g. 90%).
- *Intra-specific identity threshold* (T_{intra}): the identity threshold within species represents an identity value among sequences that clusters the great majority of orthologous alleles and excludes the great majority of paralogs. This threshold is obtained by comparing identities among putative alleles within and between individuals from the same species (Supporting Information).

Following the steps *i* and *ii* of the R2SCO pipeline, four other steps represent the definition of loci within and between samples: (*iii*) the de-replication of the preprocessed and size-selected sequences, performed for each individual separately, (*iv*) the definition of putative alleles, also performed by individual, (*v*) the pairwise comparison of putative alleles, performed among all individuals together and finally (*vi*) clustering and definition of putative loci within and between individuals.

By using the set of loci generated within each individual, the next four steps of the pipeline add a classification to all loci for every individual independently. First, in step *vii* all preprocessed and merged sequences (not de-replicated) are mapped back to the loci of that given individual. For the latter, the merged sequences used must fall within the selected size range with a ±5% extension. Subsequently, in step *viii* a SNP calling is performed for each locus individually and in step *ix* the called SNPs are used to "reconstruct" haplotypes. The last step of the pipeline *(x)* consists of getting all information obtained so far, including putative alleles, mapping results, reconstructed haplotypes and thresholds and running through a decision tree to classify the loci as "accepted" or "removed". The accepted loci will be part of the official set of reduced-representation single-copy orthologs (R2SCOs) of an individual, species or set of species.

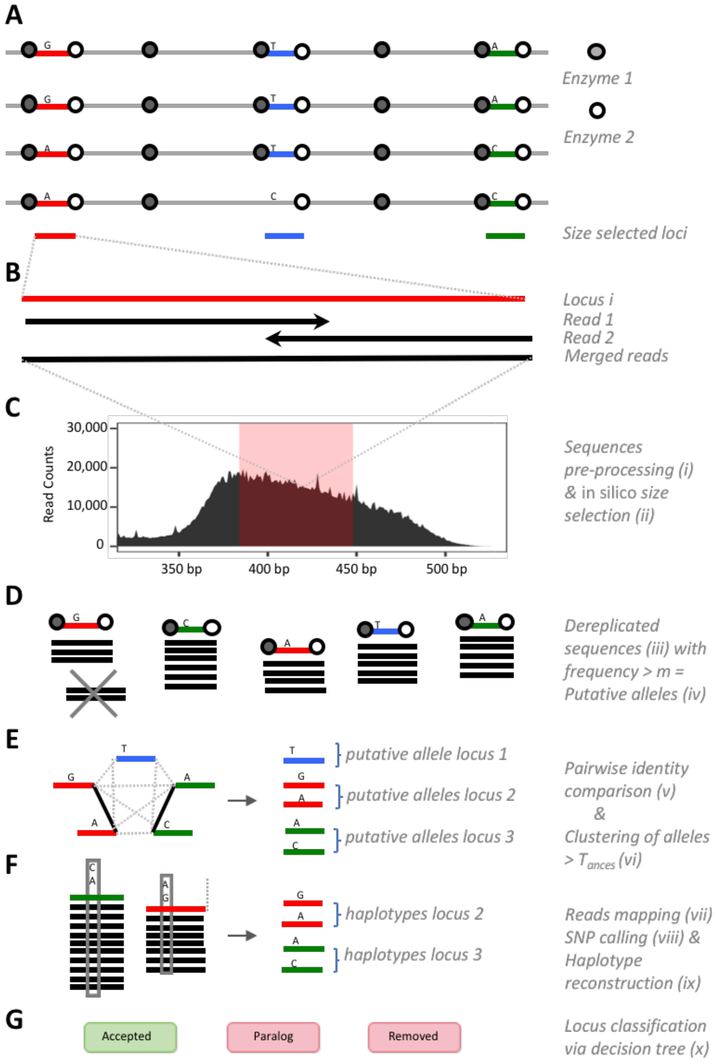
A
G
T
A
Enzyme 1
Enzyme 2
Size selected loci
B
Locus i
Read 1
Read 2
Merged reads
C
Read Counts
30,000
20,000
10,000
0
350 bp
400 bp
450 bp
500 bp
Sequences pre-processing (i) & in silico size selection (ii)
D
Dereplicated sequences (iii) with frequency > m = Putative alleles (iv)
E
} putative allele locus 1
} putative alleles locus 2
} putative alleles locus 3
Pairwise identity comparison (v) & Clustering of alleles > T_{ances} (vi)
F
} haplotypes locus 2
} haplotypes locus 3
Reads mapping (vii) SNP calling (viii) & Haplotype reconstruction (ix)
G
Accepted
Paralog
Removed
Locus classification via decision tree (x)

Fig. 1: Workflow of the reduced-representation single-copy orthologs (R2SCO) catalog building. The general representation of a ddRAD set of loci sequenced across the genome can be seen in **A**. The reconstruction of an entire locus by merging overlapping reads is shown in **B**. The R2SCO pipeline is illustrated from **C-G** with figures that represent each of the ten steps, from *i* to *x* (identified on the right part of the figure).

Locus classification decision tree

The decision tree (Figure 2) is run for each individual independently and the output can be roughly subdivided into three types of classifications: *putative paralogs, removed loci* and *accepted loci*. Loci are mostly removed due to incongruencies between the clustered putative alleles and the reconstructed haplotypes from SNP calling. For performing locus classification, the decision tree will utilize: the identity and coverage thresholds defined above (T_{intra}, T_{ances}, COV_{min}, COV_{max}); the evaluation of the re-mapping of reads against each locus; and the comparison between variants positions derived from the putative alleles and from the SNP calling. More details about the decision tree can be found in the Supporting Information.

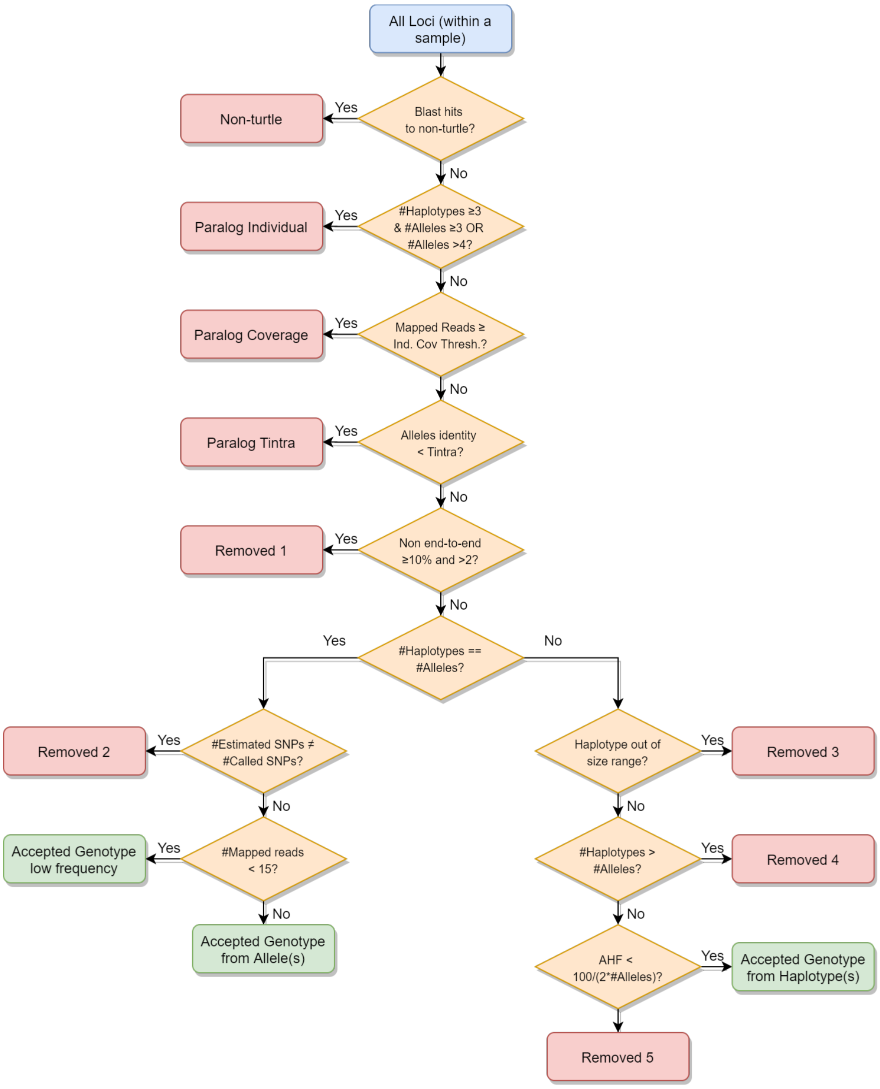
All Loci (within a sample)
Blast hits to non-turtle?
Yes
Non-turtle
No
#Haplotypes ≥3 & #Alleles ≥3 OR #Alleles >4?
Yes
Paralog Individual
No
Mapped Reads ≥ Ind. Cov Thresh.?
Yes
Paralog Coverage
No
Alleles identity < Tintra?
Yes
Paralog Tintra
No
Non end-to-end ≥10% and >2?
Yes
Removed 1
No
#Haplotypes == #Alleles?
Yes
No
#Estimated SNPs ≠ #Called SNPs?
Yes
Removed 2
No
#Mapped reads < 15?
Yes
Accepted Genotype low frequency
No
Accepted Genotype from Allele(s)
Haplotype out of size range?
Yes
Removed 3
No
#Haplotypes > #Alleles?
Yes
Removed 4
No
AHF < 100/(2*#Alleles)?
Yes
Accepted Genotype from Haplotype(s)
Removed 5

Fig. 2: Decision tree for classification of loci clustered within individuals using the R2SCO pipeline. Acceptance of a locus depends on the reliability of the genotype called. The input data (blue square on top) include alleles from all loci defined after clustering above T_{ances} and *sam* files including merged reads mapped to each locus. Orange rhombi represent each checking step, pink squares identify the non-turtle, paralogous and removed loci, and the green squares identify the accepted loci. The additional haplotype frequency (AHF) denotes the frequency of the first haplotype that exceeds the number of alleles when there are more haplotypes than alleles.

Intra and inter-specific set of R2SCOs

Once R2SCOs were defined for each individual, the pairwise results from allele clustering within and between species were used for defining homology between individuals. Homologous loci across individuals were defined by clustering loci that have at least one putative allele between individuals with identity above T_{ances}. A new class of paralog could be identified, in case two or more loci from within one individual were clustered together in any new comparison level: Paralog Species or Paralog Inter-species. The genotypes in comparisons across individuals from within or between species keep the classification of the decision tree, but loci will be removed in case they are classified as Paralogs at the level of comparison tested or if they are not present in one or more species of the set. If a locus is present and accepted in one of the two individuals, it is considered as part of the R2SCO set of that given species.

Comparisons between individual replicates

Whenever the non-selected run (see above-section *in silico* size selection**)** of an individual presented enough coverage (>10x) for at least part of the range selected for the R2SCO pipeline, it was used to evaluate the genotypes obtained and the classifications from the decision tree. A nested size range was selected for each individual depending on the coverage of the non-selected run, and both runs were processed throughout the R2SCO pipeline as independent samples. Loci from the selected runs were only evaluated if they were also present in the non-selected run. Genotypes from the non-selected runs were based solely on sequences with coverage ≥3. The great majority of the genotypes were identical between selected runs and their replicates (AVG=94.92%, SDV=1.35%). In order to evaluate the possible causes for different genotype calling as well as the decision tree classification, we have evaluated four different categories, where: a) both runs called 1 allele, b) both runs called 2 alleles, c) selected runs called 1 allele and non-selected 2 alleles, and d) selected runs called 2 alleles and non-selected 1 allele.

Reference genome evaluation

CheMyd_1.0_DNAzoo was used in different types of tests. First, the genome was digested using the RAD_digestion.py script and the sequence fragments were size-selected according to the R2SCO pipeline. Size-selected sequences were compared pairwise and clustered using T_{ances}, as described above for the empirical data. Any sequences forming clusters represent putative paralogs in the genome considering only the ddRAD simulated and size-selected sequences (not the entire genome). The putative single-copy and paralogous loci were compared with the locus classification obtained through the R2SCO pipeline for the two individuals of *C. mydas* (Cm1 and Cm2). The clusters obtained for the genome were also used to evaluate the identity among potential paralogs, obtained via the pairwise identity comparison performed using vsearch v.2.8.6 (option: --allpairs_global; Rognes, Flouri, Nichols, Quince, & Mahé, 2016) with an identity defined by T_{ances} (here 0.90).

Statistical analysis and visualization was performed in Python v.2.7.13 (Rossum, 1995) using the biopython library v.1.68 (Cock et al., 2009) and in R v.3.6.1 (Team & Others, 2013) using the packages ggplot2 (Wickham, 2016), gplots (Warnes et al., 2015), circlize (Gu, Gu, Eils, Schlesner, & Brors, 2014) and gridExtra (Auguie, Antonov, & Auguie, 2017).

3. RESULTS

Runs summary and *in silico* size range

Ten individuals (2 from each of 5 sea turtle species) were sequenced in four different MiSeq runs (Fig. S1). Nine out of the ten individuals were sequenced in replicates. The four runs yielded very different distributions, partially due to the increasing size range selection from subsequent runs, but strikingly also due to very different levels of small fragments largely out of the selected range of each run (Figure 3 and Fig. S1). The size range chosen for the selected samples was 384-448 bp, avoiding a very high peak observed in all species but *D. coriacea* at sizes 450 bp and 449 bp. One individual from *L. olivacea* (Lo2) did not reach the desired coverage for the entire range (Figure 3).

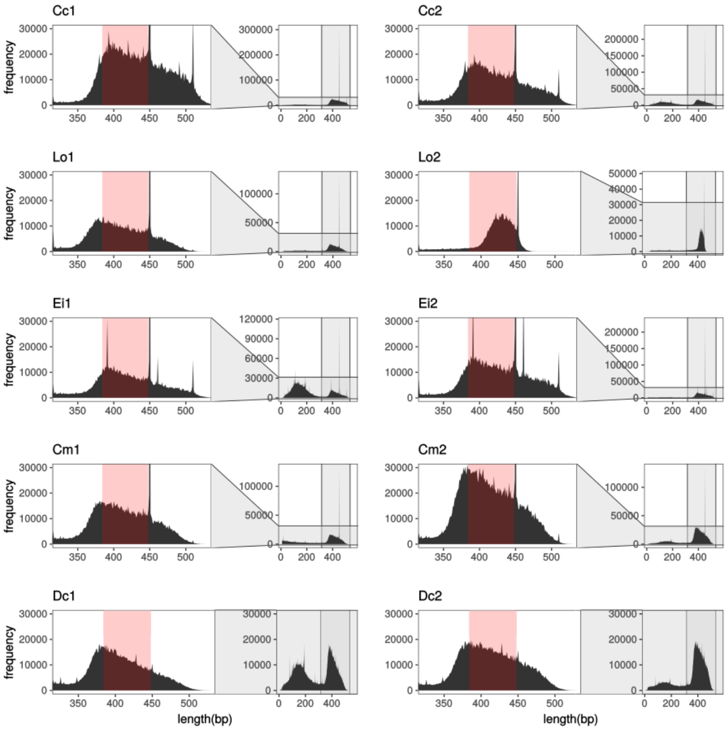

Fig. 3: Fragment size distribution of merged and preprocessed reads for 10 sea turtle individuals from 5 species, including only the selected run for each individual. The zoomed-in area on the left side shows the main distribution area for each individual and highlights in light red the *in-silico* selected size range. The plots on the right side for each individual show the entire fragment size distribution, including the very high peaks at size 450 bp.

Preprocessing

The average number of raw read pairs per individual was 2,968,893 (SD=1,110,900, Table 1). Across all ten selected samples an average of 98.47% (SD=0.17%) of the raw demultiplexed reads remained after all pre-processing steps. For each selected sample, >99% of the reads merged correctly. The *in-silico* digestion showed a frequent occurrence of internal restriction sites within the reads, representing between ~6% and ~37% of the merged sequences across individuals (AVG=15.1%, SD=8.7%). The great majority of the remaining sequences presented the correct restriction sites on the edges (AVG=99.2%, SD=0.16%) and virtually all the remaining ones (>99.97%) presented average quality above Phred score 20. Although at this point all individuals still keep the absolute majority of the initial reads (Table 1), the large amounts of small fragments have caused the proportion of remaining sequences after the *in-silico* size selection (384-448 bp) to drop substantially across individuals. Final preprocessed and size-selected sequences ranged from only ~13% of the initially demultiplexed reads in Ei1 to ~57% in Lo2 (AVG=34.03%, SD=11.69%). The statistics for each step can be seen in Table 1 for all sizes and limited to the size range selected in Figure 4.

Table 1: Sequence counts for each step of the preprocessing workflow and after size selection

Species	Sample	Demultiplexed	Merged	Non-Digested + Digested	Restr. sites filtered	Q>=20	384-448
Caretta caretta	Cc1	3,465,724	3,439,549 (99%)	3,237,364 (93%) + 202,185 (6%)	3407035 (98%)	3,406,031 (98%)	1,357,949 (39%)
	Cc2	3,663,866	3,636,115 (99%)	3,220,418 (88%) + 415,697 (11%)	3,612,355 (99%)	3,611,516 (99%)	897,962 (25%)
Lepidochelys olivacea	Lo1	1,737,259	1,725,492 (99%)	1,412,472 (81%) + 313,020 (18%)	1,715,404 (99%)	1,715,157 (99%)	627,451 (36%)
	Lo2	956,361	947,488 (99%)	771,539 (81%) + 175,949 (18%)	940,205 (98%)	939,981 (98%)	543,961 (57%)
Eretmochelys imbricata	Ei1	4,757,833	4,731,081 (99%)	4,286,842 (90%) + 444,239 (9%)	4,694,583 (99%)	4,693,922 (99%)	631,421 (13%)
	Ei2	2,347,676	2,332,021 (99%)	1,981,023 (84%) + 350,998 (15%)	2,307,027 (98%)	2,306,419 (98%)	872,961 (37%)
Chelonia mydas	Cm1	2,870,709	2,848,253 (99%)	1,796,353 (63%) + 1,051,900 (37%)	2,831,077 (99%)	2,830,546 (99%)	888,641 (31%)
	Cm2	3,964,066	3,933,845 (99%)	3,581,213 (90%) + 352,632 (9%)	3,900,863 (98%)	3,900,245 (98%)	1,545,462 (39%)
Dermochelys coriacea	Dc1	3,297,045	3,275,771 (99%)	2,919,309 (89%) + 356,462 (11%)	3,247,108 (98%)	3,246,868 (98%)	810,048 (25%)
	Dc2	2,628,393	2,606,818 (99%)	2202842 (84%) + 403,976 (15%)	2,588,546 (98%)	2,588,180 (98%)	1,015,233 (39%)

Notes: % Percentages refer to the Demultiplexed sequences; The headers refer to the steps in the preprocessing workflow; 348-448 refers to fragments within the size range 348 bp-448 bp.

Abbreviations: Restr., Restriction; Q, Phred Quality

Chimera detection

The only significant drops in sequence numbers within the *in silico* selected size range (384-448 bp) during the preprocessing step happened after *in silico* digestion (Figure 4A). We have used our newly developed pipeline (Supporting Information and Fig. S3) to differentiate between chimeric sequences and undigested fragments (Table S1). The individual Cm1 presented the highest levels of sequences with internal restriction sites (~37%, Figure 4B), mostly due to incomplete digestion. On average 53% (SD=24.4%) of the sequences with at least one internal restriction site seem to have remained undigested, while an average of ~34.6% of the sequences represented a chimeric construction.

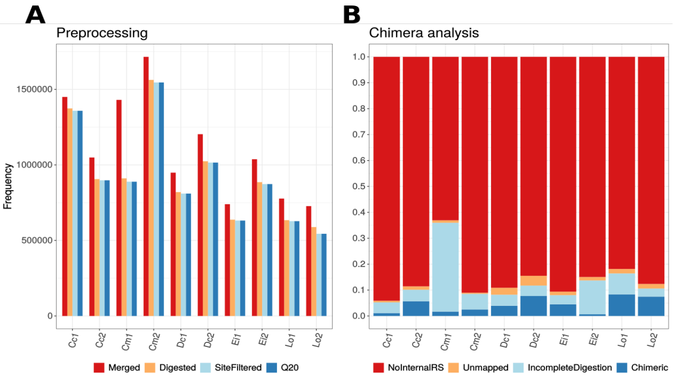

Fig. 4: Results of preprocessing analysis within the selected size range (384-448 bp). **A** shows the preprocessing statistics for the ten different samples and **B** presents the percentage of digested fragments

identified as either chimeras, with an incomplete digestion or inconclusive classification (i.e. unmapped). RS: restriction site; Q: Phred quality.

Contamination and turtle homology

The preliminary contamination analysis performed with a subsample of 50,000 paired-end reads per individual revealed very little contamination for each individual (Fig. S5). In contrast, the analysis indicated very high levels of putative homology among sea turtles. All individuals from the five species presented very high proportions of successful read mapping to the *C. mydas* genome using bowtie2 (between ~94% and ~98% of the subsampled reads, Table S2). By adding the two blast comparisons to the genomes of *C. mydas* and the western painted turtle *Chr. picta*, all ten individuals have reached putative homology levels above 97%.

R2SCO pipeline

Putative alleles were defined as identical sequences within an individual with a minimum coverage of three ($m≥3$) with sizes between 384 and 448 bp, after the *in-silico* size selection. Putative alleles were compared pairwise within and between all 10 individuals in order to define the two identity thresholds, T_{ances} and T_{intra}. The distribution of allele identities between *D. coriacea* and each of the other species was used to define a conservative ancestral threshold (T_{ances}=0.90, Figure 5A and 5C). For each species, the pairwise comparison between putative alleles within individuals and between individuals helped establish the intra-specific threshold (T_{intra}, Figure 5A-B). The T_{intra} values for Cc, Lo, Ei, Cm and Dc were respectively set to 0.989, 0.991, 0.987, 0.98 and 0.991. COV_{max} ranged from 75 to 131 mapped reads depending on the individual (Fig. S6).

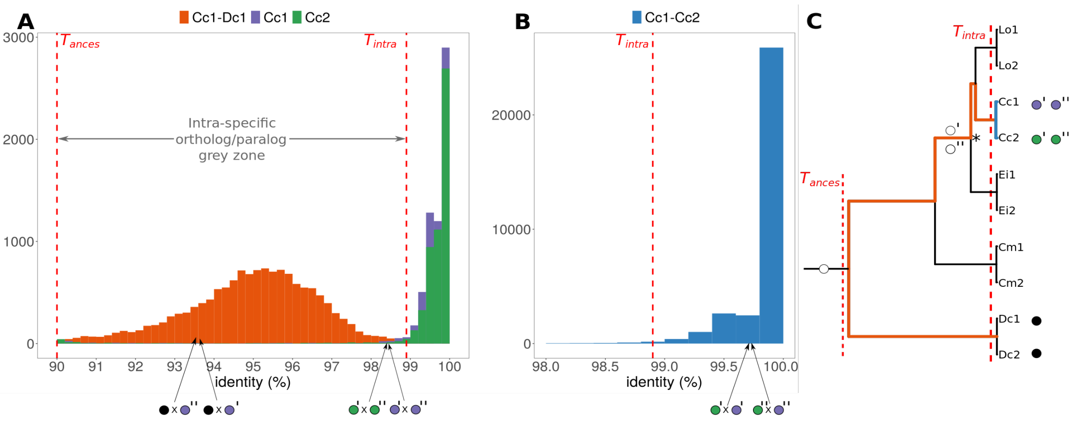

Fig. 5: Definition of identity thresholds (T_{ances} and T_{intra}) based on putative alleles pairwise identity. The figure shows the example of *C. caretta* individuals Cc1 and Cc2 and the comparison with the *D. coriacea* individual Dc1. T_{ances} and T_{intra} are shown as dashed red lines. In **A**, the whole grey zone between thresholds should include all paralogous loci that will be annotated as Paralog Tintra. The comparison between individuals in **B** shows the high amounts of identical alleles shared between them. The tree in **C** represents the sea turtle phylogeny and shows a hypothetical duplication of a locus before the tribe Carettini split (node marked with *). The identity relation between alleles from this locus is shown in all 3 figures, exemplifying where identity between paralogs are supposed to fall.

The number of accepted R2SCOs within each species was very homogeneous ranging between ~24,500 and 25,000 loci. Rejected loci were more numerous in *C. mydas* (n=3,514) followed by *E. imbricata*, *C. caretta*, *L. olivacea* and *D. coriacea* and accounted for 12.45% to 3.48% of the total number of initial loci. Comparing species pairwise within the three main phylogenetic clades (Carettini, Cheloniidae and

Chelonioidea) also revealed very homogeneous numbers of loci in common, while the total number of common loci for all species from each clade decreased as the last common ancestral became more distant (Figure 6, bars with asterisks). The number of accepted R2SCOs for all species within Carettini, Chelonioidea and Cheloniidae were respectively 16,515, 11,796 and 6,337 (Figure 6).

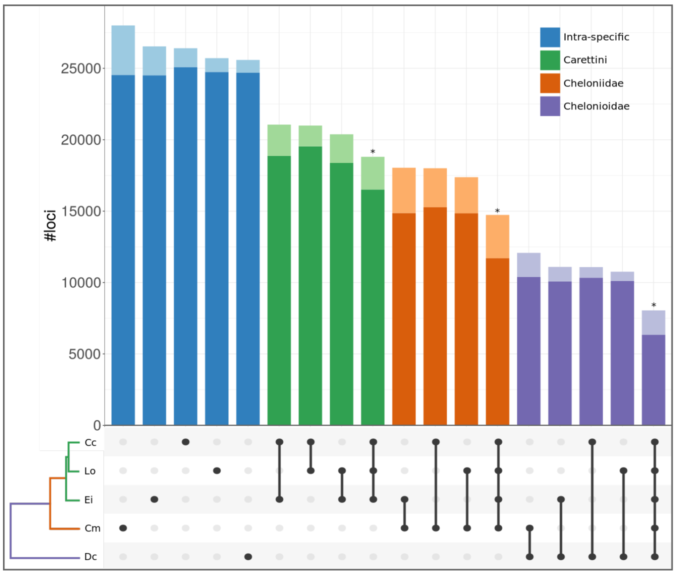

Fig. 6: Number of putative loci within and shared between species. Darker shades represent the accepted loci and lighter shades the loci removed by the decision tree. The * denotes comparisons between all species within the corresponding tribe/family/super-family. The lower part of the figure shows which species were tested together, matching the corresponding bar above.

An analysis of the total locus count per fragment size (Figure 7) revealed that all five species as well as the digested genome present not only similar numbers of loci but also the same trend of decreasing the number of loci as the fragment size increases for the pair of utilized enzymes (MseI-EcoRI). In Figure 7B, the locus counts per fragment size are shown for CheMyd_1.0_DNAzoo digested *in-silico* with and without clustering the fragments using T_{ances}.

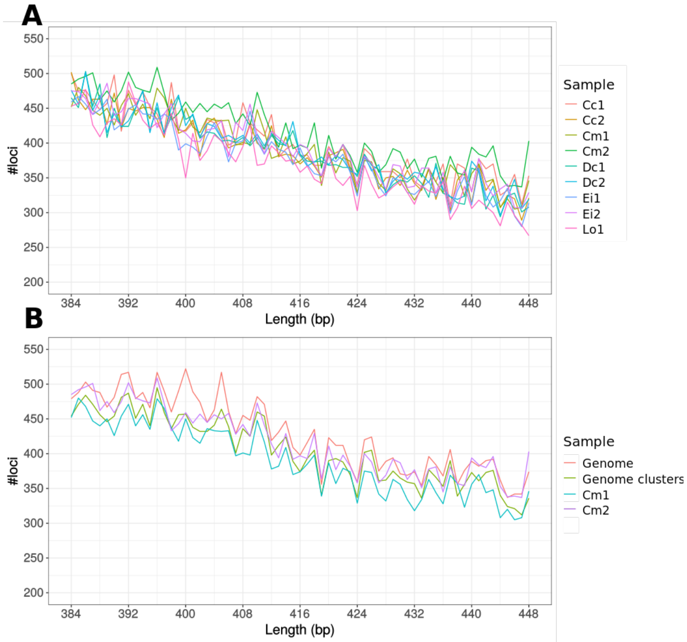

Fig. 7: Number of loci per fragment size estimated after clustering ddRAD data for 9 individuals (**A**) and (**B**) comparing the two *C. mydas* individuals (Cm1 and Cm2) and CheMyd_1.0_DNAzoo digested fragments before (Genome) and after (Genome clusters) clustering them using T_{ances}. The individual Lo2 was removed from this analysis as it presented very low coverage for the first part of the distribution.

When considering the loci for each fragment size, we noticed that the average coverage per locus clearly follows the total number of reads for a given length (Figure 8). It is possible to see in Figure 8, however, that individuals presenting smooth drops in read coverage as fragment size increases, such as Cc1, Ei2 and Cm1, show a stable average coverage per locus across most of the size range distribution. This relation reflects the drop on the number of loci across the size range as seen in Figure 7.

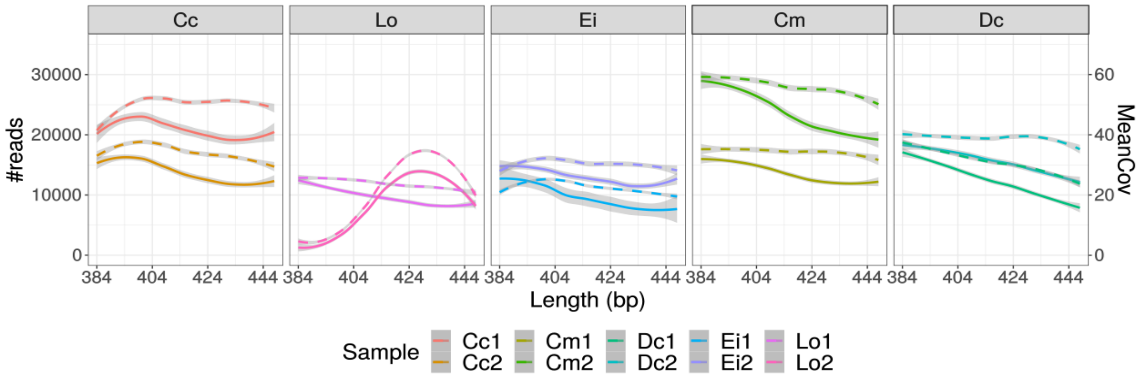

Fig. 8: Total amount of reads (y axis left, full line) vs. average locus coverage (y axis right, dotted line) per length (x axis) for each of the 10 individual selected runs. Lines were smoothed using local regression.

Genotypes and Decision tree evaluation

Six out of the nine non-selected runs for each individual were used as replicates in a genotype evaluation analysis. The comparisons were plotted for three individuals (Cc1, Cm2, Ei2) in Figure 9 and include information about allele coverage for both selected and non-selected runs and decision tree classification per locus only for the selected runs. The most important points can be summarized as: 1) The very few paralogs due to coverage in the selected run (Figure 9A) were not always highly covered in the replicate run; 2) True heterozygous in the replicate that presented a single allele in the selected run (Figure 9B, diagonal) were mostly low coverage (i.e. <15 reads) in the latter and got removed by the decision tree; 3) True heterozygous in the selected run that presented a single allele in the replicate (Figure 9C, diagonal) were mostly low coverage (i.e. <15 reads) in the latter; 4) True homozygous in the replicate that presented two putative alleles in the selected run (Figure 9C, two perpendicular lines at low values for one axis) were classified based on SNP calling as homozygous by the decision tree and clearly had very low coverages for the minor putative allele; 5) True homozygous in the selected run that presented two putative alleles in the replicate (Figure 9B, two perpendicular lines at low values for one axis), similarly presented very low coverages for the minor allele in the latter and 6) There seems to be a gradient of coverage per locus when comparing both runs in 9D (diagonal), indicating that despite the constructions of new libraries, the chance of covering each locus might not be random. A more detailed evaluation of the replicate results can be found in the Supporting Information.

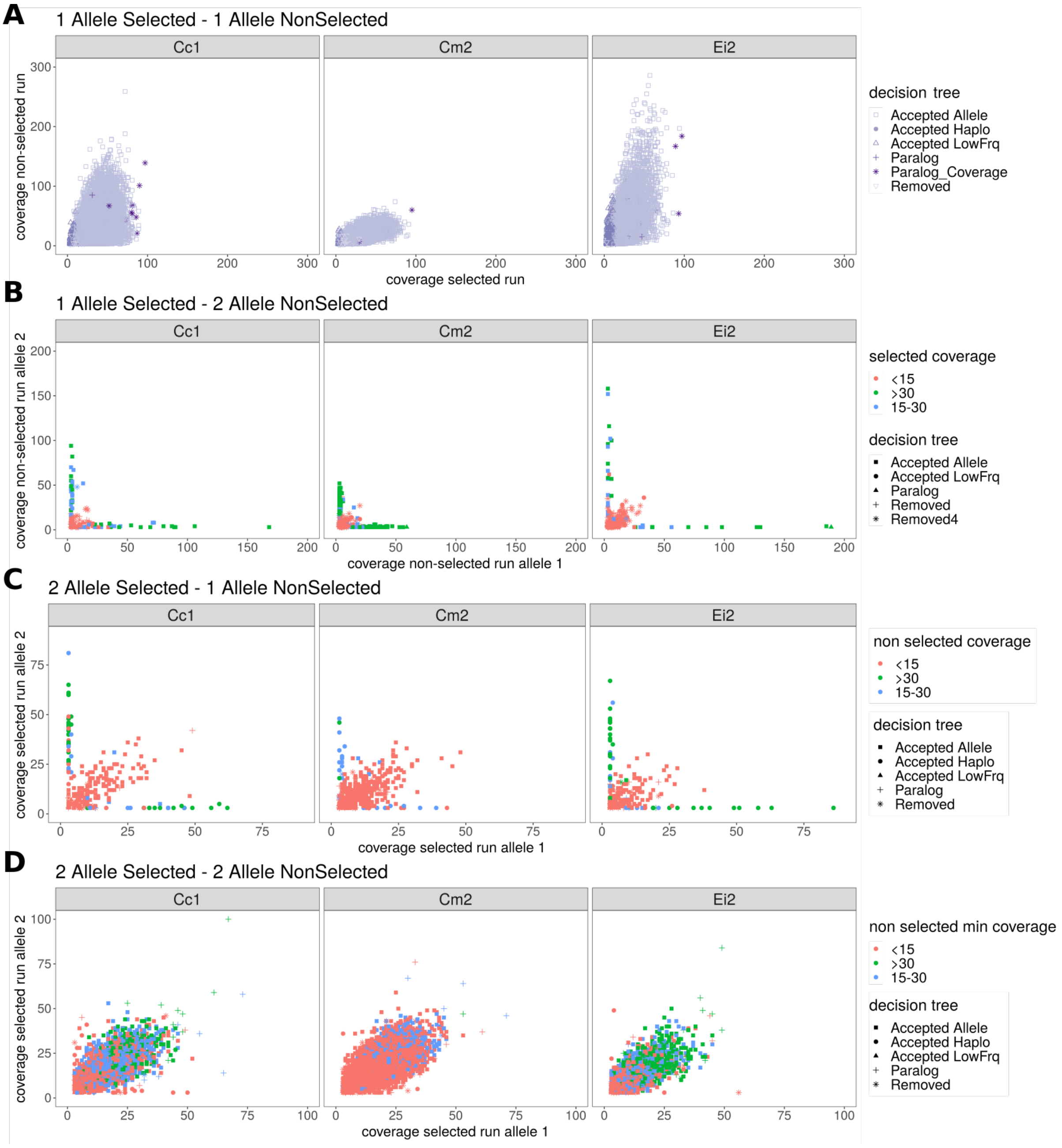

Fig. 9: Analysis of replicate runs (selected and non-selected) for three individuals. Comparisons are shown for **A**) loci where both runs agree on one allele, **B**) cases where the selected run has one allele and the non-selected run has two, **C**) the selected run has two alleles and the non-selected one and **D**) where the selected and non-selected run agree on two alleles. Colours on the bottom figure represent the minimum coverage between the two alleles from the non-selected run. All classifications coming from the decision tree represent only selected runs.

Reference Genome x R2SCOs

We have compared the results from the two *C. mydas* individuals independently (Cm1 and Cm2) and in combination (Cm species) against CheMyd_1.0_DNAzoo after *in silico* digestion with and without a clustering step (Figure 10). The great majority of accepted loci in the two individuals were found as singletons on the genome (88.47% for Cm1, 87.99% for Cm2 and 85.43% of the combined Cm species). The analysis shows that paralogy is very minor compared to the totality of loci obtained for our R2SCO set (Figure 10A and 10B), reaching only 1.04% of the genome digested fragments. In contrast, missing loci between the genome and the two individuals reach up to ~16% of the genome loci, ranging from 1,212 to 4,229 loci, depending on the comparison (Table S3). Similarly, there is a high proportion of removed loci for Cm1 and Cm2 (11.28% and 7.23%, respectively), most of which were found as singletons in the genome. Removed loci within individuals are clusters that - although not identified as paralogous within the first steps of the decision tree - were removed due to at least one type of genotyping issue identified further down in the decision tree. A more detailed analysis of the comparison between Cm individuals and the genome can be found in the Supporting Information.

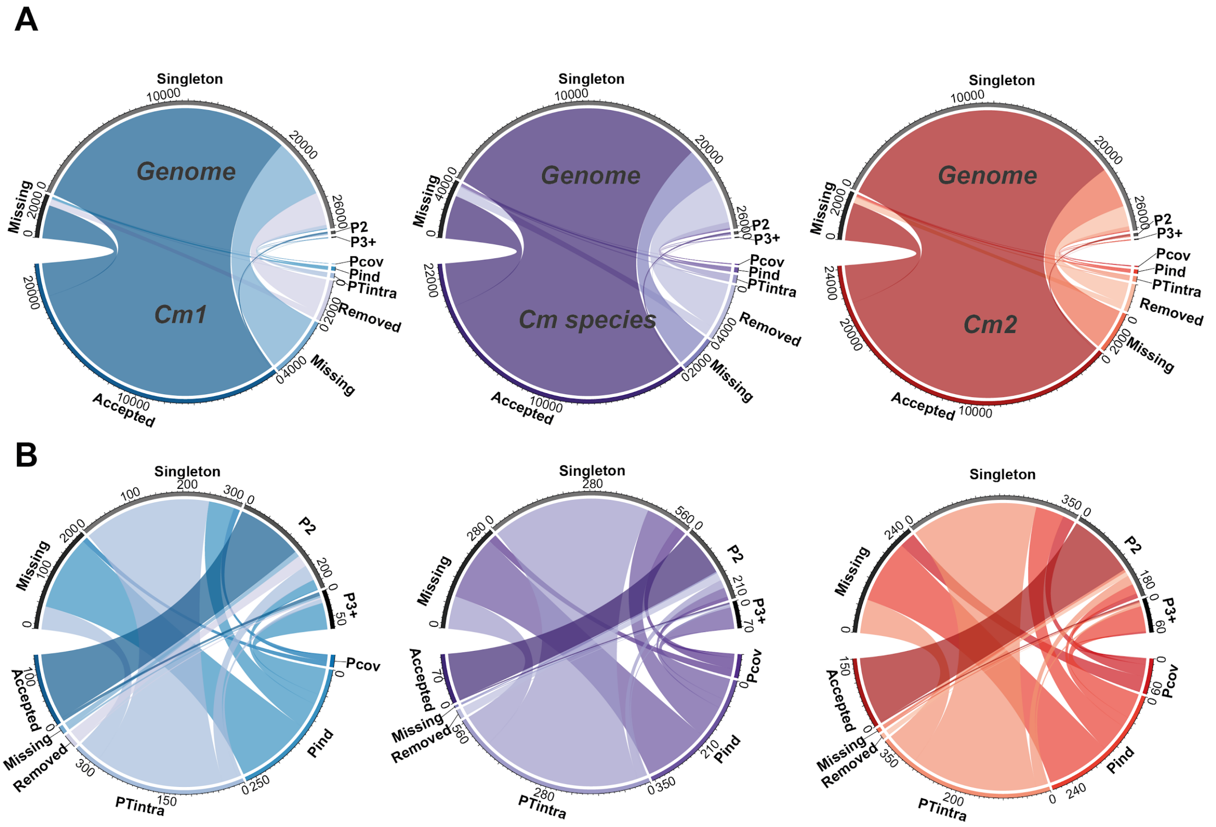

Fig. 10. Circos plots of the comparisons of two *C. mydas* individuals (Cm1 and Cm2) and their combined loci (Cm species) against the genome clusters of digested sequences (**A**). For the genome, loci are either missing (present in the individuals but not in the genome), singleton (cluster of one locus), P2 (clusters with 2 loci) or P3+ (clusters with three or more loci). For the Cm individuals, loci can be accepted (as single copies), missing (if present only in the genome digestion), removed by the decision tree or classified as paralogs based on coverage (Pcov), T_{intra} (PTintra) or the presence of more than 2 alleles (Pind). In **B** only relations to putative paralogs are shown in detail.

Sea turtle genetic variability

In order to compare the genetic variability across sea turtle species, we used the set of 6,337 R2SCOs for the whole superfamily Chelonioidea (Chelonioidea-R2SCO-MseI-EcoRI-384-448, Figure 6), excluding the accepted genotype low frequency category for each individual. The numbers of loci used in this analysis for each individual can be seen in Table S4.

The levels of heterozygosity found for conspecific individuals were very similar, but diverged greatly among species (Figure 11A). *Chelonia mydas* was the most variable species, almost three-fold higher than the second most heterozygous one, *E. imbricata*. *Caretta caretta* individuals were the third most variable and presented very similar heterozygosity values although belonging to very distant locations (Atlantic vs. Indo-Pacific). *Lepidochelys olivacea* and *D. coriacea* both presented extremely low levels of heterozygosity (<0.05%), roughly 10-fold smaller than *C. mydas*. Among the five sea turtle species evaluated in this study, *D. coriacea* seems to present the lowest genetic variability.

The third histogram bar for each species in Figure 11A shows a comparison of pairwise genetic distances between individuals. For the four species whose both individuals belong to southwestern Atlantic populations there was only a slight increase in the level of variability when comparing both individuals in relation to intra-individual analysis, indicating that both populations share a major part of their variability. This is however not true for *C. caretta*, for which individuals originated from different ocean basins and seem to have quite divergent alleles, reaching twice as much the distance between alleles found within individuals (Figure 11A, Cc1-Cc2 bar).

The amounts of variable loci per individual follow the trend of species heterozygosity levels (Figure 11B), with *C. mydas* individuals presenting the highest proportion of heterozygous loci (AVG=53.61%) while *L. olivacea* and *D. coriacea* present roughly only 10% of variable loci within individual. However, there is a clear tendency to saturation as the average number of SNPs per locus increases (see the theoretical red line in Figure 11C based on *L. olivacea* and *D. coriacea* values). Finally, in order to check if the heterozygosity levels calculated based on the R2SCOs at the Chelonioidea level could be biased due to the conservation of restriction sites across all five species, we have estimated heterozygosity at different sets of loci, within and among species. Very similar values of heterozygosity were found across the different levels of R2SCOs (Fig. S7).

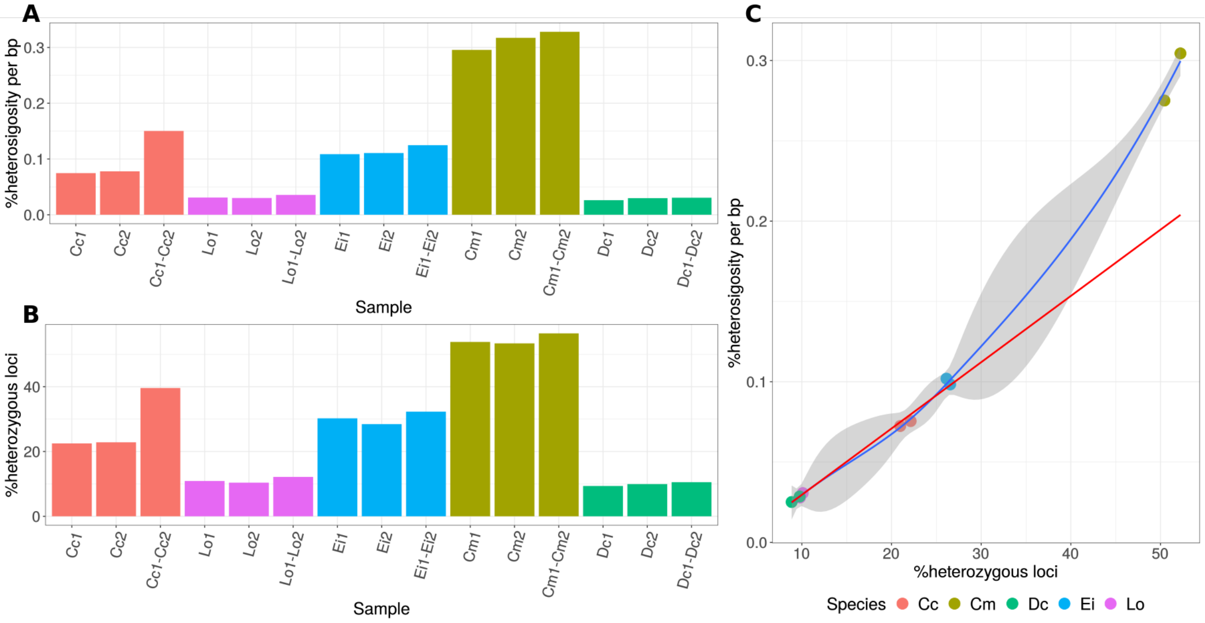

Fig. 11: Heterozygosity levels at Chelonioidea R2SCOs. In **A** and **B**, the third bar for each species corresponds to the analysis of conspecific individuals. In **A**, the heterozygosity was estimated for each individual and genetic distances between individuals. **B** shows the proportion of variable loci for each individual and the combining results from two individuals. In **C**, the relation between heterozygosity per base pair vs. the proportion of heterozygous loci is shown for each individual. The blue line is a loess regression based on all values and the red line is a linear regression based on values for Dc and Lo.

Using the 6,337 Chelonioidea R2SCOs to estimate pairwise genetic distances among all individuals we obtained the true phylogenetic relationships among the five sea turtle species, as seen in a heatmap analysis (Figure 12). *Dermochelys coriacea* showed an average of 4.67% (SD=0.11%) genetic distance against the other four lineages, while showing slightly smaller distances against *C. mydas* (AVG=4.48%). *Chelonia mydas* presented very similar genetic distances (2.32%, 2.39% and 2.35%) against each of the three species from the Carettini tribe. *Lepidochelys olivacea* and *C. caretta* presented the smallest genetic distances (AVG=0.96%), while *C. caretta* had a slightly smaller genetic distance to *E. imbricata* compared to *L. olivacea* (averages 1.08% and 1.15%, respectively). The genetic distance matrix among all individuals from the five species can be found in Table S5.

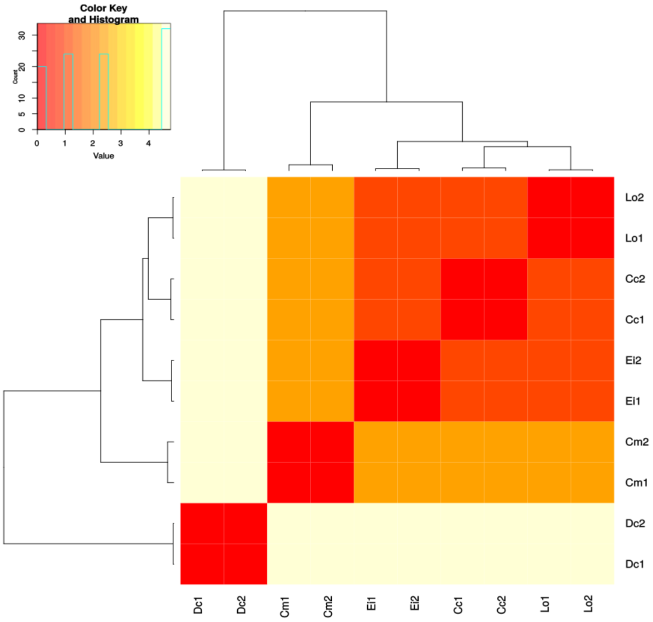

Fig. 12: Heatmap of pairwise genetic distances across all individuals from the five sea turtle species. Genetic distance was calculated based on the total number of differences between the representative sequences of shared loci averaged by the total sequence length.

4. DISCUSSION

A new strategy to build ddRAD-like locus catalogs

Here we present a carefully designed (R2SCO) pipeline, which proved to be a powerful tool to build reliable references to intra and interspecific analysis. Each step of the study, from experimental design to statistical analysis, is carefully planned and analyzed in detail. As the first step, in order to build high-quality locus catalogs that maximize the identification of variation for future population or other types of studies, individuals should represent most of the variation across the set of populations to be analyzed (Davey & Blaxter, 2010). Therefore, we selected two individuals for each species coming from the most genetically distant populations we could identify in our collection.

The biggest novelty in our approach, in terms of methodology, is the use of overlapping Illumina reads. Illumina sequences have very low error rates (Pfeiffer et al., 2018), however, base quality usually drops towards the end of the read. This is especially true for the longest Illumina reads (i.e. 300 bp paired-end reads from MiSeq), as also observed in our run quality controls (data not shown). This problem was completely solved with merging the reads, and quality was generally very high across the whole extent of the merged sequences. This was also verified in our quality filtering step, which barely removed any sequences. The pattern of low removal rates remained the same when we increased the average scores threshold to Q30.

With high quality sequences representing the original fragment length, we could use the distribution of fragment lengths to evaluate the size selection performed in the lab and to perform a new size selection *in silico*, in order to homogenize coverages across loci and individuals. Our analysis of the fragment length distribution revealed huge variations across ddRAD libraries. However, even if in some cases we had to remove more than 80% of the sequences, we managed to select a range of well covered (>20x) loci across all five species, representing roughly 25,000 single-copy loci per individual. The R2SCO pipeline generated not only well-supported single-copy loci, but also gave indications of whether SNP calling would be able to retrieve the whole variation present in the alleles of each locus. The loci "removed" by our decision tree have issues of either biological (e.g. one allele out of range) or technical background (e.g. erroneous SNP calling, accumulation of sequencing errors, or alleles that are not mappable to representative locus sequences).

The consistency across individuals is also seen in the comparison of expected versus observed average locus coverage per length (Fig. S8). The putative alleles are compared to each other exhaustively (i.e. pairwise for all possible combinations), and a new script was designed to build clusters based on the pairwise alignments above the ancestral threshold. The exhaustive search for identity connections guarantees reproducibility of alleles clusterization within and between individuals. Since the data was extremely reduced to a set of high-quality and high-coverage sequences, the extended computational time demanded for exhaustive searches became a minor issue. We also performed the pairwise comparisons including putative alleles from all 10 individuals at the same time, and the alignment results were used for all intra-individual, intraspecific and interspecific analyses later on.

We have also selected identity and coverage thresholds in an innovative way, using the identities generated by the pairwise comparisons to help establish an ancestral as well as intraspecific thresholds. T_{intra} correlates negatively with the species heterozygosity, as it should be set to the initial point of the distribution of identities between alleles (Figure 5). In fact, *D. coriacea* and *L. olivacea* presented the highest T_{intra} and the lower heterozygosity, while *C. mydas*, the most polymorphic species, presented the lowest T_{intra}. T_{intra} is comparable to the value *M* of Stacks (Catchen et al., 2013), as it is set to avoid over-splitting of alleles, but should still keep the number of allowed differences low considering an intra-specific analysis. In contrast, we can arguably compare T_{ances} with the identity threshold from ipyrad (Eaton & Overcast, 2020), which is set to accommodate homology across a set of different species. The combination of both thresholds allowed us to keep clusters representing paralogous loci and annotate them as putative paralogs within species. This way, we still perform comparison of loci across species using lower thresholds (T_{ances}), even if using an intra-specific identity threshold (T_{intra}) for locus classification for each species independently.

Given our analysis and comparisons to the genome, we believe that our set of R2SCOs for each species is better suited for future population studies than the draft genome available. Although most of the loci were classified as single copy in both *C. mydas* R2SCOs and in the genome, we have identified large numbers of problematic loci that will most likely cause issues during genotyping of population data. We identified a variety of possible explanations, such as: presence of a second allele out of the size selected range, mostly due to internal indels; too many erroneous sequences within a locus; errors in homopolymeric or microsatellite regions; difficulties for the SNP caller to properly identify some longer or complex indel regions; issues with mapping for part of the reads that originally matched one allele within a locus; among others. Finally, the genome has a large amount of missing – *in-silico* digested - loci compared to Cm1 and Cm2. This might be due to the higher heterozygosity levels of *C. mydas* and the fact that the genome sample comes from an individual from the Indo-Pacific (Wang et al., 2013), in contrast to the two individuals from southwestern Atlantic populations. As we have seen in the initial contamination analysis, levels of homology between the Cm individuals and the entire genome are very high (~98.8%, Table S2). This means that the fact that after genome digestion ~10% of the accepted loci for Cm1 and Cm2 are missing in the genome is mostly due to the genetic divergence (in the restriction sites) accumulated between the distant populations.

Possible adjustments for future libraries and analysis limitations

Our data revealed significant carryover of small fragments in the size selection attempted in the lab, as also seen by daCosta & Sorenson (2014). We believe that this is due to a combination of enzyme choice and the protocol used. MseI-EcoRI generates high amounts of small fragments (Fig. S4), mostly due to the fact that MseI is a 4-bp cutter rich in AT (recognition site: AATT) and cuts the genome with a very high frequency. Moreover, we have performed the indexing PCR before size selection, therefore including all digested fragment sizes in the 10-cycle-amplification. As amplification of sequencing libraries tends to favour smaller fragments (Dabney & Meyer, 2012), we believe that the number of small fragments increased so dramatically for some libraries that even precise machines like the PippinPrep did not manage to remove them all. This might be easily solved by performing the indexing PCR after the size selection, as performed in the original protocol (Peterson et al., 2012). Moreover, it is probably sensible to choose enzymes, if possible, that do not produce such a biased ratio towards small fragments. Similarly to daCosta & Sorenson (2014), the fragment analyzer analysis did not show any obvious presence of small fragments after our libraries were ready for sequencing (data not shown).

Our approach has some limitations in the identification of paralogs that are very closely related and present up to two alleles in total. If alleles of such paralogs have identities above T_{intra} and coverage does not reach outlier levels, the only way to identify them is by using either a high-quality assembled genome or population data, since the proportion of heterozygous individuals should be greater for duplicated loci in comparison to singleton loci at any given allele frequency (McKinney, Waples, Seeb, & Seeb, 2017). This is more prone to happen in species with low variability, as clustering two paralogous loci together would not give indication of a third allele. When testing several individuals across populations, the chance of identifying a third allele within an individual is also increased.

Sea turtle genetic variability and possible applications

We have used ~6.6K putative single copy loci distributed across the sea turtle genomes to estimate heterozygosity levels per bp. Since our approach used entire sequences from merged paired-end reads, we could also use the invariable positions to estimate levels of genetic variation. The heterozygosity levels (0.028% and 0.031%) found for the two *C. mydas* individuals are slightly higher but not significantly different from the heterozygosity (AVG=0.024%, SD=0.018%) estimated across the draft *C. mydas* genome (Fitak & Johnsen, 2018). *Chelonia mydas* was substantially more variable compared to the other four sea turtle species, but its heterozygosity levels were comparable to the saltwater crocodile *Crocodylus porosus* genome (Green et al., 2014). Green et al. (2014) found low levels of heterozygosity in all three crocodilian genomes they analyzed when compared to avian and mammalian genomes. The least variable species, the American alligator *Alligator mississippiensis*, presented heterozygosity levels around 0.01%, which is at least three times higher than what we found for *L. olivacea* and *D. coriacea*, and comparable to *E. imbricata*, which presented the second highest heterozygosity levels in our study (Figure 11). The reduced genetic diversity observed for *L. olivacea* and *D. coriacea* compared to other species is concordant with their lower haplotypic diversity based on mitochondrial control region data for worldwide populations (Reid, Naro-Maciel, Hahn, FitzSimmons, & Gehara, 2019).

Assuming that the consistency across locus numbers and heterozygosity levels in our dataset make values comparable across species, we can evaluate the genetic distances among sea turtle species.

Chelonia mydas presented smaller distances to *D. coriacea* in comparison to all three members of the Carettini tribe, which might indicate one of two things: 1) a small acceleration of genetic divergence within Carettini, maybe during the short period in which speciation events across lineages took place or 2) a deceleration of genetic divergence in the C. *mydas* lineage, which does not agree with the current higher levels of genetic variation for this species. Similarly, *E. imbricata* presents consistently lower (albeit with a minor difference) genetic distances to *C. caretta* in comparison to *L. olivacea*. Although this might be a small analytical artefact, it is worth mentioning that *C. caretta* and *E. imbricata* have a proven capacity to generate fertile hybrids in large scale (Soares et al., 2018, 2017; Vilaca et al., 2012). Despite the fact that this is most likely a common phenomenon within the Carettini tribe, the presence of gene flow across lineages might influence levels of genetic distances and deserve a closer look in future studies involving genome-wide analysis.

When comparing individuals from different populations, we could see a clear difference between populations that are geographically close (i.e. from southwestern Atlantic) compared to populations from different oceans (*C. caretta* individuals). Although it is not possible to confirm whether a slight increase in distances (Figure 11) between alleles from different populations in comparison to within populations (i.e. within individual) represent population structuring, it might indicate a trend that should be evaluated in further studies using reduced-representation methods and more specifically the same set of R2SCOs we established for each species.

The data generated in this study is highly comparable across sea turtle species and can potentially be readily used to develop markers for different purposes and that could fit one or even all five species at the same time. By presenting loci with entire allele sequences that are alignable and with lengths between 384 and 448 bp, our data can be used to develop primers for small to large assays for different purposes and that could include Sanger sequencing, SNP panels or new microsatellite markers. Digging deeper into the data will reveal variable microsatellites within and across species, as we have observed ourselves. The least variable species, *D. coriacea* and *L. olivacea*, present only about 10% of the loci variable within individual. This knowledge should also be useful to guide future studies designing reduced-representation or genome-wide analysis for these species, since the level of invariable loci should be considered. Because R2SCO loci are conserved across sea turtle species, they can be potentially used for reconstructing the evolutionary history of species, discovery of genes related to adaptation and sex determination, estimate the genetic diversity and population structure, development of methods to assign individuals to nesting population, among other uses that can benefit sea turtle conservation.

Application to other species

Using sea turtle data, Chow et al. (2019) showed that reference-based SNP discovery using genomes of more closely related species allows identifying more SNPs. Similar to several other studies, they demonstrate the importance of establishing a proper reference for the target species. Here we present a method to build a locus catalog that is comparable to results from a conspecific genome but yet better suited for related species and even populations divergent from the genome. Our approach can be potentially transferred for any other species. We have successfully tested it in birds, mammals and freshwater turtles (manuscripts in preparation). However, given the level of locus and even allele dropout that might be expected in species with higher heterozygosity levels, we recommend and use ourselves this approach for species with low levels of heterozygosity. Although the method was very successful for all sea turtle species, locus dropout was clearly more present in *C. mydas*, which still presents a fairly low level of heterozygosity of ~0.3%.

Low coverage results in decreased ability to confidently detect heterozygotes, affecting the analysis involving individual genotypes (Barbanti et al., 2020; Chow et al., 2019). This should specially be avoided when building the set of R2SCOs for a species, as the evaluation of heterozygosity as well as the potential technical issues regarding some loci will be crucial to help planning larger scale analyses (that could use shorter reads) and therefore would decrease issues related to genotyping at the population level.

As pointed out by McCartney-Melstad et al. (2019), no single default value for clustering threshold should be expected to be accurate for all studies or species. Our approach can help define empirical thresholds for any new species analyzed, with no need for previous knowledge regarding that species genetic variability. It is important, however, to have educated assumptions about population representativeness by a given specimen. This way, the new reference set would not only be representative of the set of populations analyzed, but could also be used for a preliminary analysis of genetic distances among different populations, even with one single individual per population.

Conclusions

Coupling stepwise quality control and overlapping reads, this study presents a detailed overview of ddRAD-like methodologies up to the stage of variant calling. Our empirical tests showed several potential issues in the ddRAD laboratory protocol and bioinformatic analyses. The use of long haplotypes, for part or the totality of samples, can substantially improve the quality of the set of reference loci and therefore potentially eliminate a series of biases incorporated during analysis and that can strongly affect downstream analysis.

There is an ever-growing number of studies trying to overcome issues produced by RAD experiments and data (e.g. Barbanti et al., 2020; Chow et al., 2019; LaCava et al., 2020; Paris et al., 2017; Shafer et al., 2017). However, the issues caused by the technique are highly variable and seem to be specific to each experiment (DaCosta & Sorenson, 2014). The approach developed here proposes to use overlapping reads to gain power over the data created, helping scientists evaluate every single library created for a given species or population.

Finally, our study is a pioneer in performing a comparative study among sea turtles species using genome-wide data and confirming the very slow rate of genomic change in chelonians (Avise, Bowen, Lamb, Meylan, & Bermingham, 1992; Shaffer et al., 2013). Low yet variable levels of genetic variability combined with high levels of homology make sea turtles an ideal system among vertebrates for evaluating genomic methods that analyze vertical evolution. The very low levels of genetic diversity (~0.03%) found in this study in at least two of the sea turtle species reinforces the urgency to broaden genomic studies across the globally threatened group of sea turtle species.

Acknowledgements

This work was supported financially by the Alexander von Humboldt Foundation, the Leibniz Institute for Zoo and Wildlife Research and the ANTIDOT project (Pépinière Interdisciplinaire Guyane, Mission pour l'Interdisciplinarité, CNRS). All sequencing and computer infrastructure utilized was part of the BeGenDiv Consortium. We are grateful to the Brazilian sea turtle conservation program Projeto TAMAR for collecting and providing part of the samples. We are similarly grateful to Prof. Fabricio Santos, who provided some of the Brazilian samples. We are also grateful for the sequencing lab support of Susan Mbedi.

REFERENCES

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews. Genetics*, *17*(2), 81–92.

Auguie, B., Antonov, A., & Auguie, M. B. (2017). Package "gridExtra." *Miscellaneous Functions for "Grid" Graphics*. Retrieved from http://cran.dcc.fc.up.pt/web/packages/gridExtra/gridExtra.pdf

Avise, J. C., Bowen, B. W., Lamb, T., Meylan, A. B., & Bermingham, E. (1992). Mitochondrial DNA evolution at a turtle's pace: evidence for low genetic variability and reduced microevolutionary rate in the Testudines. *Molecular Biology and Evolution*, *9*(3), 457–473.

Barbanti, A., Torrado, H., Macpherson, E., Bargelloni, L., Franch, R., Carreras, C., & Pascual, M. (2020). Helping decision making for reliable and cost-effective 2b-RAD sequencing and genotyping analyses in non-model species. *Molecular Ecology Resources*, *10*, 555.

Beichman, A. C., Huerta-Sanchez, E., & Lohmueller, K. E. (2018). Using genomic data to infer historic population dynamics of nonmodel organisms. *Annual Review of Ecology, Evolution, and Systematics*, *49*, 433–456.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* , *30*(15), 2114–2120.

Campbell, E. O., Brunet, B. M. T., Dupuis, J. R., & Sperling, F. A. H. (2018). Would an RRS by any other name sound as RAD? *Methods in Ecology and Evolution* , *9*(9), 1920–1927.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140.

Chow, J. C., Anderson, P. E., & Shedlock, A. M. (2019). Sea Turtle Population Genomic Discovery: Global and Locus-Specific Signatures of Polymorphism, Selection, and Adaptive Potential. *Genome Biology and Evolution*, *11*(10), 2797–2806.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* , *25*(11), 1422–1423.

Dabney, J., & Meyer, M. (2012). Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques*, *52*(2), 87–94.

DaCosta, J. M., & Sorenson, M. D. (2014). Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. *PloS One*, *9*(9), e106713.

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, *9*(5-6), 416–423.

Díaz-Arce, N., & Rodríguez-Ezpeleta, N. (2019). Selecting RAD-Seq Data Analysis Parameters for Population Genetics: The More the Better? *Frontiers in Genetics*, *10*, 533.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., … Aiden, E. L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95.

Eaton, D. A. R., & Overcast, I. (2020). ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics* , 1–3.

Fitak, R. R., & Johnsen, S. (2018). Green sea turtle (Chelonia mydas) population history indicates important demographic changes near the mid-Pleistocene transition. *Marine Biology*, *165*(7), 110.

Franchini, P., Monné Parera, D., Kautt, A. F., & Meyer, A. (2017). quaddRAD: a new high-multiplexing and PCR duplicate removal ddRAD protocol produces novel evolutionary insights in a nonradiating cichlid lineage. *Molecular Ecology*, *26*(10), 2783–2795.

Green, R. E., Braun, E. L., Armstrong, J., Earl, D., Nguyen, N., Hickey, G., … Ray, D. A. (2014). Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science*, *346*(6215), 1254449.

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *bioRxiv*. doi: 10.1101/729962

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize Implements and enhances circular visualization in R. *Bioinformatics* , *30*(19), 2811–2812.

Huson, D. H., Albrecht, B., Bağcı, C., Bessarab, I., Górska, A., Jolic, D., & Williams, R. B. H. (2018). MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct*, *13*(1), 6.

Ilut, D. C., Nydam, M. L., & Hare, M. P. (2014). Defining loci in restriction-based reduced representation genomic data from nonmodel species: sources of bias and diagnostics for optimal clustering. *BioMed Research International*, *2014*, 675158.

LaCava, M. E. F., Aikens, E. O., Megna, L. C., Randolph, G., Hubbard, C., & Buerkle, C. A. (2020). Accuracy of de novo assembly of DNA sequences from double-digest libraries varies substantially among software. *Molecular Ecology Resources*, *20*(2), 360–370.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25.

McCartney-Melstad, E., Gidiş, M., & Shaffer, H. B. (2019). An empirical pipeline for choosing the optimal clustering threshold in RADseq studies. *Molecular Ecology Resources*, *19*(5), 1195–1204.

McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, *17*(4), 656–669.

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*(6), db.prot5448.

Naro-Maciel, E., Le, M., FitzSimmons, N. N., & Amato, G. (2008). Evolutionary relationships of marine turtles: A molecular phylogeny based on nuclear and mitochondrial genes. *Molecular Phylogenetics and Evolution*, *49*(2), 659–662.

Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, *22*(11), 2841–2847.

Paris, J. R., Stevens, J. R., & Catchen, J. M. (2017). Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution* , *8*(10), 1360–1373.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS One*, *7*(5), e37135.

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., & Mayer, G. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*, *8*(1), 10950.

Phillippy, A. M. (2017). New advances in sequence assembly. *Genome Research*, *27*(5), 11–13.

Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, *2*, e431.

Reid, B. N., Naro-Maciel, E., Hahn, A. T., FitzSimmons, N. N., & Gehara, M. (2019). Geography best explains global patterns of genetic diversity and postglacial co-expansion in marine turtles. *Molecular Ecology*, *9*, 367.

Reis, E. C., Soares, L. S., Vargas, S. M., Santos, F. R., Young, R. J., Bjorndal, K. A., … Lôbo-Hajdu, G. (2010). Genetic composition, population structure and phylogeography of the loggerhead sea turtle: colonization hypothesis for the Brazilian rookeries. *Conservation Genetics* , *11*(4), 1467–1477.

Roehr, J. T., Dieterich, C., & Reinert, K. (2017). Flexbar 3.0 - SIMD and multicore parallelization. *Bioinformatics* , *33*(18), 2941–2942.

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584.

Rossum, G. (1995). *Python reference manual* [Technical Report]. Retrieved from CWI (Centre for Mathematics and Computer Science) website: https://dl.acm.org/citation.cfm?id=869369

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in Molecular Biology* , *1962*, 227–245.

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution* , *8*(8), 907–917.

Shaffer, H. B., Minx, P., Warren, D. E., Shedlock, A. M., Thomson, R. C., Valenzuela, N., … Wilson, R. K. (2013). The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biology*, *14*(3), R28.

Soares, L. S., Bjorndal, K. A., Bolten, A. B., dei Marcovaldi, M. A. G., Luz, P. B., Machado, R., … Wayne, M. L. (2018). Effects of hybridization on sea turtle fitness. *Conservation Genetics* , *19*(6), 1311–1322.

Soares, L. S., Bolten, A. B., Wayne, M. L., Vilaça, S. T., Santos, F. R., dei Marcovaldi, M. A. G., &

Bjorndal, K. A. (2017). Comparison of reproductive output of hybrid sea turtles and parental species. *Marine Biology*, *164*(1), 9.

Team, R. C., & Others. (2013). *R: A language and environment for statistical computing*. Retrieved from https://repo.bppt.go.id/cran/web/packages/dplR/vignettes/intro-dplR.pdf

Vilaca, S. T., Vargas, S. M., Lara-Ruiz, P., Molfetti, E., Reis, E. C., Lôbo-Hajdu, G., … Santos, F. R. (2012). Nuclear markers reveal a complex introgression pattern among marine turtle species on the Brazilian coast. *Molecular Ecology*, *21*(17), 4300–4312.

Wang, Z., Pascual-Anaya, J., Zadissa, A., Li, W., Niimura, Y., Huang, Z., … Irie, N. (2013). The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nature Genetics*, *45*(6), 701–706.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., … Others. (2015). *gplots: Various R programming tools for plotting data*. Retrieved from https://www.scienceopen.com/document?vid=0e5d8e31-1fe4-492f-a3d8-8cd71b2b8ad9

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* , *30*(5), 614–620.

Data Accessibility Statement

The codes in python are available through github at https://github.com/BeGenDiv/Driller_et_al_2020. The ddRAD data are available in the Sequence Read Archive through bioproject PRJNA######. The sets of R2SCOs representative sequences for each individual and comparison levels (species, Carettini, Cheloniidae, Chelonioidea) are available at Dryad doi:##.####/dryad.#####

Author Contributions

C.J.M. and S.T.V. designed the study. B.T. and D.C. provided samples and financial support. S.T.V. performed the lab work. C.J.M, M.D. and L.S.A. analyzed the data. M.D. developed the accompanying scripts and produced most of the figures. T.C.V. supported the development of the decision tree. F.H. developed the code for clustering. C.J.M. and M.D wrote the manuscript with input from all co-authors.