Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014, **5,** 872–880 doi: 10.1111/2041-210X.12228

qPCR-based molecular sexing by copy number variation in rRNA genes and its utility for sex identification in soft-shell turtles

Robert Literman, Daleen Badenhorst and Nicole Valenzuela*

Department of Ecology, Evolution and Organismal Biology, Iowa State University, 251 Bessey Hall, Ames, IA 50011, USA

Summary

1 Sex diagnosis is important in ecology, evolution, conservation biology, medicine, and food production. However, sex diagnosis is difficult in species without conspicuous sexual dimorphism or at life stages before such differences develop. This problem is exacerbated when the diagnostic trait is a continuous (non-discrete) variable to which general analytical methods are not commonly applied.

2 Here we demonstrate the use of copy-number variation between males and females of the nucleolar organizing region (NOR) in the genome of *Apalone spinifera* softshell turtles, which we quantify by real-time PCR. We analyze these continuous data using mixture models that can be applied either in discriminant analysis when a subset of individuals of known sex is used as a training set, or in clustering procedures when all individuals are of unknown sex.

3 Using individuals of known sex, the discriminant analysis exhibited 100% accurate classification rate for both the training set and the test set. Classification rates were also 100% when using the clustering procedure to identify groups and classify individuals in the absence of sex information. Standard curves using only male DNA provided better discrimination than using mixed-sex DNA during qPCR. NOR copy number is an effective sex diagnostic for *A. spinifera* turtles. Our sexing approach using qPCR of 18S genes should prove useful for other taxa that also possess dimorphic NORs, as is known in some vertebrates and insects. While the 18S copy numbers in our dataset exhibited a non-overlapping binomial distribution, this may not always be the case in future studies of *A. spinifera* or for other taxa.

4 Importantly however, our sex-typing approach using mixture models provides an attractive alternative under overlapping distributions of these and of other continuous data such as hormone levels, gene expression levels, shape or behavior. We present an example using overlapping distributions of hormone levels in *Chelydra serpentina* turtles, to demonstrate the broader utility of mixture models for sex-typing, and obtain a high correct classification of 90%.

Key-words: animal science, conservation, ecology, evolution, mixture models and discriminant analysis, molecular sexing of animals, nucleolar-organizing region, reptiles, sex typing or sex diagnosis by quantitative real-time PCR, trionychid turtles *Apalone spinifera* and *Pelodiscus sinensis*

Introduction

Accurate and early identification of the sex of animals is imperative in fields spanning from medicine to evolutionary and conservation biology. For instance, sex assessment is required prior to embryo implantation in humans and domestic animals to diagnose diseases or to appraise embryo quality during *in vitro* fertilization (Hamilton *et al.* 2012). Likewise, fisheries and other animal industries benefit from early sex identification to select the most desirable gender for commercial purposes (Singh 2013). In conservation biology, sex information is crucial to implement sex-specific management strategies or sex ratio monitoring of endangered species (Mrosovsky 1982; Korstian *et al.* 2013). Finally, studying the ecology and evolution of sex allocation and sex-specific traits also requires reliable sex identification (Ellegren & Sheldon 1997; Griffiths 2000).

Obtaining information on individual sex is simple for species or life stages that exhibit obviously dimorphic phenotypes. However, difficulties emerge for organisms or life stages where no diagnosable external dimorphism exists that is detectable by visual inspection. Several techniques have been devised to sex individuals in such cases and applied to diverse taxonomic groups. However, some direct techniques are destructive, such as the observation of gonadal morphology or histology in dead animals (Yntema & Mrosovsky 1980), while others are invasive, such as laparoscopic inspection of gonadal morphology in live animals (Wood *et al.* 1983; Rostal *et al.* 1994). Less intrusive sex diagnosis can be accomplished by detecting the

*Correspondence author. E-mail: nvalenzu@iastate.edu

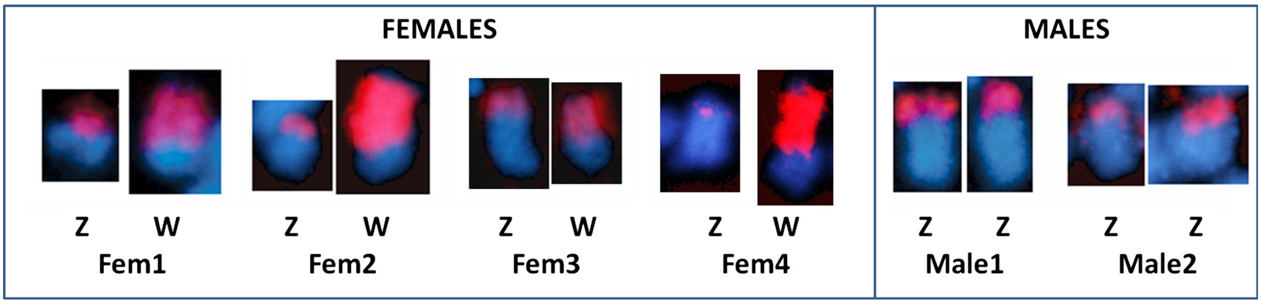

Fig. 1. ZZ/ZW sex chromosomes of *Apalone spinifera* (modified from Badenhorst *et al.* 2013). Red colour corresponds to the fluorescent *in situ* hybridization of an 18S rRNA gene probe revealing a larger block of 18S repeats in the W than in the Z chromosomes.

presence/absence of a sex-linked trait using molecular approaches, such as the cytogenetic detection of sex chromosomes (Ezaz *et al.* 2005; Badenhorst *et al.* 2013), PCR amplification of a sex-specific marker (Griffiths 2000; Morinha, Cabral & Bastos 2012; Korstian *et al.* 2013) or quantitative PCR (qPCR) of genes linked to the sex chromosomes that are present in two copies in one sex and one copy in the other (Phillips & Edmands 2012; Alasaad *et al.* 2013; Ballester *et al.* 2013). The molecular techniques mentioned above represent examples of discrete traits. Alternatively, sex assessment may rely on the indirect measurement of some continuous feature that is sexually dimorphic such as hormone levels (Owens *et al.* 1978; Akyuz *et al.* 2010), gene expression (Hamilton *et al.* 2012) or multivariate data such as shape (Valenzuela *et al.* 2004; Ceballos & Valenzuela 2011; Ceballos, Hernandez & Valenzuela 2014).

Turtles are a lineage exemplifying the need and difficulty of sex diagnosis. While many turtle species display sexually dimorphic characters as adults such as size or shape differences (Ceballos *et al.* 2012), hatchlings and juveniles usually lack early sexual dimorphism that is visually diagnosable. Yet, sex information of embryonic or young turtles is crucial to monitor sex ratios and to study sex-specific traits that may influence fitness [e.g. (Janzen 1993; Ceballos, Hernandez & Valenzuela 2014)]. Consequently, multiple sexing techniques have been developed for turtles, including gonadal inspection or histology (Yntema & Mrosovsky 1980), laparoscopy (Wood *et al.* 1983; Rostal *et al.* 1994), radioimmunoassay of circulating hormones in blood (Owens *et al.* 1978; Lance, Valenzuela & von Hildebrand 1992; Rostal *et al.* 1994; Valenzuela 2001) or chorioallantoic/amniotic fluid of the egg (Gross *et al.* 1995). The least invasive sexing method for juveniles utilizes geometric morphometric quantification of subtle dimorphism in the turtle carapace of several species (Valenzuela *et al.* 2004) or in the anal region of the plastron in others (Ceballos, Hernandez & Valenzuela 2014). However, because geometric morphometric quantifies shape by the relative position of carapace scutes, which serve as homologous landmarks, it cannot be applied to soft-shell turtles because their shells lack carapace scutes altogether, and their sexual size dimorphism is not evident prior to sexual maturity at 8–10 years of age (Ernst & Lovich 2009). Moreover, tests of circulating hormones are expensive and cumbersome.

Apalone spinifera soft-shell turtles exhibit a ZZ/ZW sex chromosome mechanism of genotypic sex determination (Badenhorst *et al.* 2013). Unfortunately, molecular cytogenetic techniques are costly and highly specialized, such that ZZ/ZW detection for sex-typing large numbers of individuals in population-level studies is precluded. Importantly, fluorescent *in situ* hybridization (FISH) of an 18S rRNA gene probe revealed that the nucleolar-organizing region (NOR) in *A. spinifera* is located on the sex chromosomes and exhibits a much greater copy number on the W than on the Z (Fig. 1), making it a promising dimorphic marker for sex identification (Badenhorst *et al.* 2013). The NOR contains genes for the three major ribosomal RNA subunits (18S, 5.8S and 28S) repeated in tandem to permit sufficient transcription to supply cellular demands for ribosomes (Shaw & McKeown 2011). When NORs are located in the non-recombining region of sex chromosomes, the number of repeats may become sexually dimorphic, as in *A. spinifera* turtles (Badenhorst *et al.* 2013).

When using continuous traits for sex typing, the analytical methods to assign individuals as male or female fall into two main categories. The first category uses a set of individuals of known sex to train an algorithm that is then used to assign the sex of unknown samples as male or female (Valenzuela *et al.* 2004; Ceballos & Valenzuela 2011; Ceballos, Hernandez & Valenzuela 2014). The second category relies on the bimodality of the continuous variable in the absence of any *a priori* sex information from any individual and then assignment of an individual as male or female based on how close its value is to one or the other group mean. This latter assignment, however, is usually performed in an *ad hoc* fashion rather than using standardized statistical procedures, especially for individuals with intermediate values that approach the area of overlap in the bimodal distribution [e.g. (Valenzuela 2001; Weissmann *et al.* 2013)]. Thus, while a variety of molecular sexing techniques have been widely used to assign individuals to sexes, a general approach for the use of any continuous dimorphic molecular data as a sex diagnostic tool is not commonly applied, particularly when the cut-off between males and females in the binomial distribution is not as evident. Mixture models provide such a framework (Fraley & Raftery 2002).

Here, we use the novel 18S genomic region for sexing *A. spinifera* turtles. The 18S copy number variation among individuals represents a continuous variable that can be quantified via qPCR and analysed using mixture models and univariate discrimination (Fraley & Raftery 2002) for sex typing. Our approach offers an attractive alternative for the fast, accurate and reliable sex diagnosis in soft-shell turtles. Our molecular method is applicable to broader taxa that possess sexually dimorphic NORs (Goodpasture & Bloom 1975; Hsu, Spirito & Pardue 1975; Schmid *et al.* 1983, 1993; Bickham & Rogers 1985; Born & Bertollo 2000; Kawai *et al.* 2007; Abramyan, Feng & Koopman 2009; Monti, Manicardi & Mandrioli 2011; Takehana *et al.* 2012; Badenhorst *et al.* 2013), and our analytical approach is appropriate for any other bimodal continuous variables and multivariate traits with overlapping distributions. We provide such an example using hormonal data from snapping turtles, *Chelydra serpentina*.

Materials and methods

SAMPLE COLLECTION

Apalone spinifera eggs were incubated at 26°C, 28°C or 31°C as described previously (Valenzuela 2010). Hatchlings were housed in a temperature-controlled facility and were given access to UV light, burrowing substrate, water and a dry basking area to ensure healthy growth. At *c.* 3 months of age, gonadal differentiation was advanced to the point that the sex of 89 hatchlings could accurately be determined by visual gonadal inspection. At this age, ovaries displayed clear ovarian ducts and prominent follicles, while testes exhibit substantial seminiferous tubule development and are smaller than the ovaries.

DNA EXTRACTION AND QUALITY CONTROL

DNA was extracted from muscle tissue using Gentra Puregene DNA extraction kit (Gentra) following the manufacturer's instructions and was quantified and quality checked using a NanoDrop ND-1000 Spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and gel electrophoresis (0·8% agarose). Then, a subset of 40 male and 40 female hatchlings with high molecular weight DNA was selected for further analysis. DNA was diluted to 1·25 ng μL^{-1} for the use in the quantitative PCR (qPCR) assay. This DNA concentration produced qPCR amplification profiles with similar fluorescence levels for both the 18S and GAPDH genes during a pilot test.

QUANTIFICATION OF 18S rRNA REPEAT COPY NUMBER

Copy number of the 18S rRNA repeats was quantified in each individual using qPCR and normalized against GAPDH, a single copy gene used as endogenous control. Using data from an *A. spinifera* transcriptome (S. Radhakrishnan and N. Valenzuela, unpublished data), qPCR primers were designed to amplify a 144-bp fragment of 18S rRNA (forward 5′-GAGTATGGTTGCAAAGCTGAAA-3′; reverse 5′-CGA GAAAGAGCTATCAATCTGT-3′) and a 129 bp fragment of GAPDH (forward 5′-GGAGTGAGTATGACTCTTCCT'-3′; reverse 5′-CAGCATCTCCCCACTTGA-3′). Standard curves were generated by pooling equimolar amounts of five high-quality genomic DNA (gDNA) samples. Pooled DNA was diluted to 100 ng μL^{-1} and then serially diluted from 1 : 10 to 1 : 640 for final concentrations of 10, 5, 2·5, 1·25, 0·625, 0·3125 and 0·1563 ng μL^{-1}. Two different standard curves were tested in this study: (1) a mixed-sex standard curve containing DNA from three male and two female samples chosen at random (to simulate conditions where individual sex is unknown) and (2) a male-only standard curve made by pooling the DNA of five known males (to test whether a standard curve made with DNA from the sex that has smaller 18S blocks provides better discrimination of 18S copy number between males and females). qPCR was performed using Brilliant II SYBR Green qPCR Master Mix (Agilent) in an Mx3000P thermocycler (Agilent, Santa Clara, CA, USA), with ROX as the reference dye for background correction. qPCR was performed in 25 μL reactions containing 2 μL of sample DNA (2·5 ng) or standard DNA and a final primer concentration of 400 nM. qPCR cycling conditions were as follows: 1 cycle at 95°C for 10 min; 45 cycles of 95°C for 30 s, 58°C for 1 min, 72°C for 1 min; and a dissociation curve cycle of 95°C for 1 min, 55°C for 30 s, taking readings at 0·5°C increments until reaching 95°C for 1 min, to test for unspecific amplification. Samples and standards were run in duplicate in each qPCR plate. Threshold fluorescent values for each qPCR plate were first automatically assigned by the MXPRO software (Agilent), and an overall average threshold value was manually chosen, which was appropriate for all genes and plates. Any samples whose replicates exhibited non-specific amplification or a C_T deviation greater than 0·5 between duplicates were excluded from further analysis. Negative, no-template controls were also run in duplicate to test for primer dimers or contamination. The efficiency of each qPCR reaction was calculated from the standards as follows:

$$\text{Eff} = 10^{-(1/\text{slope})}$$

Copy number of the 18S gene was normalized against GAPDH using the comparative C_T method of normalization (Livak & Schmittgen 2001):

$$\text{Ratio}\,(\frac{18\text{S}}{\text{GAPDH}}) = 2^{-\Delta C_T} = 2^{C_{T\,\text{GAPDH}} - C_{T\,18\text{S}}}$$

Other normalization methods are compared in Appendix A.

GENERAL ANALYTICAL METHOD FOR SEX IDENTIFICATION

The goal of any sexing technique is to assign individuals to groups (males and females). Using a single continuous trait, the first step in this process is to visualize a histogram of the data, which should be bimodal with respect to sex (Appendix B). A test is then carried out to validate the sexual dimorphism of the trait in question and its efficacy for accurate sex typing of individuals as described below. Here, we use mixture models which consider the data as containing combinations of two or more distributions, with each mixture component corresponding to a group whose parameters can then be estimated (Baudry *et al.* 2010). The most common component is typically a combination of multiple normal distributions. Analytically, parameter estimates of mixture models may be calculated using an expectation maximization (EM) procedure in a likelihood framework [see (Fraley & Raftery 2002)]. To implement the procedure described above, two conceptual approaches are possible, which depend on the data available (Appendix B). R-code and data for an implementation example are found in Appendices C and D.

Procedure 1 – discriminant analysis

If the sex of a subsample of individuals is known (determined by other techniques such as gonadal inspection), this subsample is first used as a training set to find the parameters for each group's distribution (means and standard deviations for males and females). The conditional probabilities of each sample belonging to each of the groups given the parameters of the data (z) are calculated, and individuals are assigned to the group that minimizes the uncertainty (1–z). When applied to the training set, the training classification rates measure the fit of the model to the data. Second, conditional probabilities are calculated for each unknown sample in the test data set, and individuals are assigned to groups in the same fashion using the parameters calculated from the training set. An additional test can be carried out by dividing the subsample of individuals of known sex into two groups, one to be used as a smaller training set and the other to be used as a test set by ignoring the known sex information. In this case, the parameters of the male and female groups are calculated as described above using the smaller training set and then used to classify the test set individuals as male or female. Thus, the classification for the test set serves as cross-validation for the sex-typing approach (because the true sex of individuals in the test set is actually known). The classification error rate for the test set provides the level of confidence that can be expected for the sex typing of unknowns using this approach (Fraley & Raftery 2002).

Procedure 2 – clustering analysis

If the sex of all individuals is unknown, mixture models are first used to find the distributions (groups) that best fit the data and to estimate the means and standard deviations of each group thus identified. The conditional probabilities of each sample belonging to each of the groups given the parameters of the data (z) are calculated, and individuals are assigned to the group that minimizes the uncertainty (1–z), in the same manner as for Procedure 1. The uncertainty provides a measure of the quality of the classification by subtracting from 1 the probability of the most likely group for each individual (Fraley, Raftery & Scrucca 2012).

POTENTIAL DATA COMPLICATIONS

The method described above is straightforward when the variation for both sexes is similar (standard deviations are comparable) and when there are no outlier values. To ensure this, some additional steps should be followed. First, deviant values are identified using the EM algorithm in the mixture model, where outliers are classified into their own cluster [see (Fraley & Raftery 2002)]. This classification is inspected visually to determine the sex of the group(s) that corresponds to the outlier values. For instance, if the biological expectation is that males have low values, while females have high values, samples with deviant high numbers will denote females with extreme values at the upper tail of their distribution, while deviant low values would correspond to males at the low tail of their distribution. After classification, assignments could be visually inspected with respect to the distribution. Second, if the variation is not uniform between the sexes, the mixture model procedure will favour a model with unequal variance, as this provides the best fit to the data. That model is then implemented for parameter estimation and classification.

Another complication emerges when the distributions of male and female values overlap. In the case that sex information is available for a subset of individuals, an estimate of the overlap and the classification error that it generates can be obtained using Procedure 1. Additionally, for this case and for the case when sex information is unavailable for a subsample of individuals, the uncertainty levels calculated as described for both procedures can be used to remove from the data set individuals that cannot be classified with an acceptable confidence level as defined by the researcher (e.g. >80%, >90%, >95%) (Appendix F).

Here, we tested both analytical methods (Procedures 1 and 2) in *A. spinifera*. First, because sex information was available for all our samples, we tested the classification rates employing Procedure 1 (discriminant analysis) when two-thirds of the samples (46 hatchlings) were used as a training set to generate the discriminant model, and the other third (22 hatchlings) as a testing set for cross-validation. Second, we tested the classification rates employing Procedure 2 (clustering analysis) by treating all the samples as if they were unknowns (ignoring the gonadal sex information available) and allowing the mixture model to identify the groups in the absence of any prior sex information. We then examined the concordance of the estimated and true sex information to assess the performance of Procedure 2. Statistical analyses were carried out using the MCLUST v4.2 package (Fraley, Raftery & Scrucca 2012) in R using the MclustDA function for Procedure 1 and the Mclust function for Procedure 2.

Results

GONADAL INSPECTION, QPCR QUALITY CONTROL AND 18S NORMALIZATION

The sex ratio of the 89 hatchlings was 46 males : 43 females (Table 1), which did not differ from 1 : 1 and was not influenced by temperature (chi-square, $P = 0{\cdot}75$), as expected for a species with genotypic sex determination (Bull & Vogt 1979).

Of these 89 individuals, 40 males and 40 females with high-quality DNA were used for qPCR. Four male and eight female samples were removed from the analysis as their C_T values differ by >0·5 cycles between replicates. The final data set contained 68 individuals. Dissociation curves after qPCR detected no secondary products or primer dimers, and negative controls were clean. Plate efficiencies and quality of the standard curve as determined from the coefficient of determination (R^2) are summarized in Table 2. The qPCR efficiencies (Eff) per plate

Table 1. Sex ratios of *Apalone spinifera* hatchlings determined by visual sexing of gonads. *P*-values represent results of chi-square analyses that test whether the sex ratio is skewed from 1 : 1. 'Unknown' corresponds to hatchlings from 28°C or 31°C whose incubation information was lost

Egg incubation temperature	26°C	28°C	31°C	Unknown	Overall
No. of males	8	13	20	5	46
No. of females	7	13	17	6	43
Chi-square *P*-value	0·80	1·0	0·62	0·76	0·75

Table 2. Efficiency of the qPCR and quality of the standard curve for all plates run in this study. Two 96-well plates were needed for each gene given our sample size

Plate	Standard curve	qPCR efficiency	Standard curve R^2
18S 1	Male-only	1·9966	0·996
18S 2	Male-only	2·0058	0·993
GAPDH 1	Male-only	1·9864	0·997
GAPDH 2	Male-only	1·9731	0·996
18S 1	Mixed sex	2·0336	0·994
18S 2	Mixed sex	2·0080	0·996
GAPDH 1	Mixed sex	1·9975	0·993
GAPDH 2	Mixed sex	2·0126	0·994

ranged between 1·973 and 2·034 (97·3%–103·4%), and the R^2 values are all >0·99. Thus, amplification reactions for GAPDH and 18S were highly efficient, comparable and appropriate to predict the sample C_T values by linear regression. Alternative methods of normalization were also tested, and our results were robust across all methods (Appendix A).

To assess the similarity of the qPCR efficiencies between the 18S and GAPDH amplification, we ran a regression analysis of the ΔC_T values ($C_{T,GADPH} - C_{T,18S}$) of the standard curve samples against the Log$_2$(DNA amount) for each standard curve dilution, following Livak & Schmittgen (2001). The slope of the regression of ΔC_T vs. Log$_2$(DNA template amount) was <0·1 for both standard curve types (Fig. 2), indicating that the qPCR efficiencies were similar for the amplification of 18S and GAPDH, and thus, that the use of the comparative C_T method of normalization was appropriate for our data (Livak & Schmittgen 2001). This is important because the C_T method of normalization is only applicable if the qPCR efficiencies for the gene of interest and endogenous control are both around 100% (Eff≈2) and comparable between genes.

Values of the ratio of 18S rRNA to GAPDH copy number exhibited a bimodal distribution with no overlapping values between males and females (Fig. 3). The highest male 18S/GAPDH ratio was 298, while the lowest female 18S/GAPDH ratio was 429 (Table 3, Fig. 3). On average females had approximately four times as many copies of the 18S rRNA gene than males (Table 3). This result is concordant with the cytogenetic evidence, which shows that the W chromosome in female *A. spinifera* contains an extended NOR region that houses many more copies of 18S rRNA than males (Badenhorst *et al.* 2013) (Fig. 1). Additionally, the variance in 18S copy number among females (coefficient of variation = 42·3–45·5%) was greater than the variance among males (coefficient of variation = 26·3–29%) (Fig. 4, Table 3). These differences were caused mainly by a subset of females that had relatively higher 18S/GAPDH values than the rest, which are identified as an outlier group using mixture models (Fig. 3g). When those outlier females are removed, the coefficient of variation is similar for males and females. Our results were robust to using two alternative methods for normalization of 18S copy number, using the same samples run with the male-only and mixed-sex standard curves (Appendix A). When using the CT normalization, an ANOVA did not detect any effect of the standard curve type (mixed-sex vs. only-male standards) on the mean 18S/GAPDH ratio within-sex (Fig. 4, Table 3).

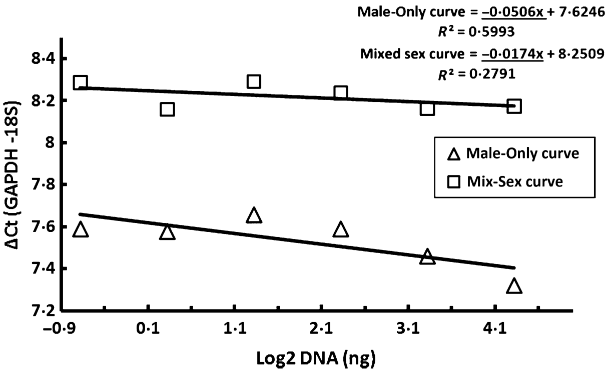

Fig. 2. Assessment of the qPCR efficiencies for the gene of interest (GOI) and endogenous control (EC). The regression of ΔC_T against template amount (Log$_2$ DNA) revealed a slope close to zero (less than 0·1), indicating that the qPCR efficiencies for the GOI and EC are similar enough to use the comparative C_T method (Livak & Schmittgen 2001). Both standard curve types employed in this study meet this requirement. Slope values are underlined.

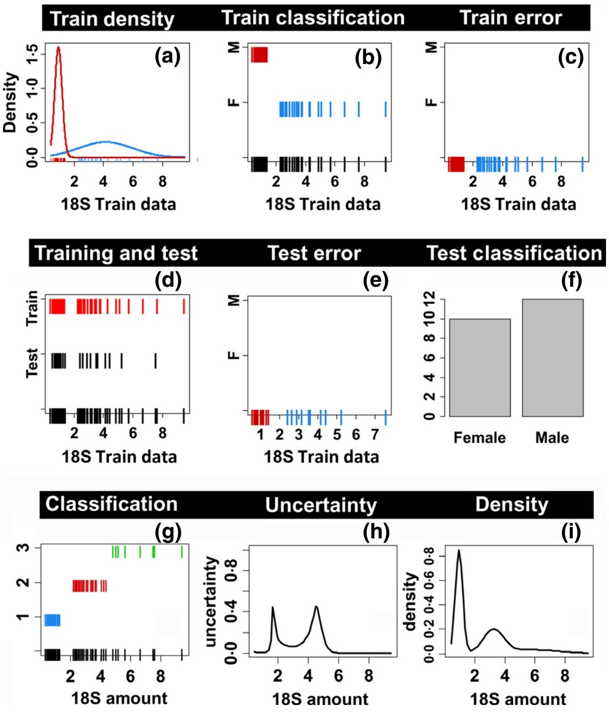

Fig. 3. Results of the use of mixture models for discriminant analysis (a–f panels) for sex typing of *Apalone spinifera* turtles when sex information is available [including distribution density, classification and error rates for the training and test data sets] and for clustering (g–i panels) in the absence of *a priori* sex information. In panel g, note the identification of three groups: typical males (blue), typical females (red) and 'outlier' high-value females (green).

INDIVIDUAL SEX ASSESSMENT USING MIXTURE MODELS OF CLUSTERING

Results from the analytical mixture models using the discriminant analysis (Procedure 1), and treating 46 individuals as a training set and 22 individuals as a test set resulted in a classification error rate of 0% for the training set (Fig. 3b,c) and for the test set during cross-validation (Fig. 3d,e). Using the mixture models and treating all individuals as unknowns identified three clusters corresponding to the male, female and female outlier groups from the bimodal distribution (Fig. 3g). As expected, uncertainty values increased in the areas between two groups (Fig. 3h). However, the classification rate treating all individuals as unknown was 100% accurate. Namely, all individuals in the lowest group (group 1) were males, and all individuals in groups 2 and 3 were females.

Table 3. Normalized 18S copy number (ratio of 18S rRNA to GAPDH) in the *Apalone spinifera* genome as measured by qPCR. All samples were run with both a male-only and mixed-sex standard curve, and 18S values normalized with three alternative methods (equations 2–4) as detailed in the Appendix A

	Relative standard curve method		Pfaffl calibrator method		Comparative C_T method ($2^{-\Delta CT}$)	
Normalized 18S	Male-only standards	Mixed-sex standards	Male-only standards	Mixed-sex standards	Male-only standards	Mixed-sex standards
Male minimum	0·420	0·295	0·306	0·295	66·028	85·036
Male average	0·955	0·667	0·832	0·670	177·698	192·544
Male maximum	1·423	1·039	1·243	1·039	266·871	298·172
Female minimum	2·320	1·361	2·247	1·236	429·049	435·039
Female average	4·095	2·431	3·973	2·193	760·853	740·707
Female maximum	9·494	5·481	9·184	4·917	1764·447	1640·591
Avg. female : male	4·288	3·645	4·775	3·273	4·282	3·847
C.V. male	26·364	28·95	26·963	28·921	26·974	28·751
C.V. female	42·355	45·482	42·372	45·365	42·563	44·419
Gap between sexes (MinFem–MaxMale)	0·897	0·322	1·004	0·197	162·178	136·867
Total range (MaxFem–MinMale)	9·075	5·186	8·878	4·622	1698·419	1555·555
ANOVA test of effect of standards on male average	F = 29·57 d.f. = 71 **P < 0·05**		F = 10·69 d.f. = 71 **P < 0·05**		F = 1·48 d.f. = 71 P > 0·05	
ANOVA test of effect of standards on female average	F = 20·94 d.f. = 63 **P < 0·05**		F = 26·52 d.f. = 63 **P < 0·05**		F = 0·0609 d.f. = 63 P > 0·05	

C.V., coefficient of variation; Avg, average; MaxFem, maximum female value; MinMale, minimum male value; *F*, *F* statistic; d.f., degrees of freedom. Significant *P*-values are denoted in bold.

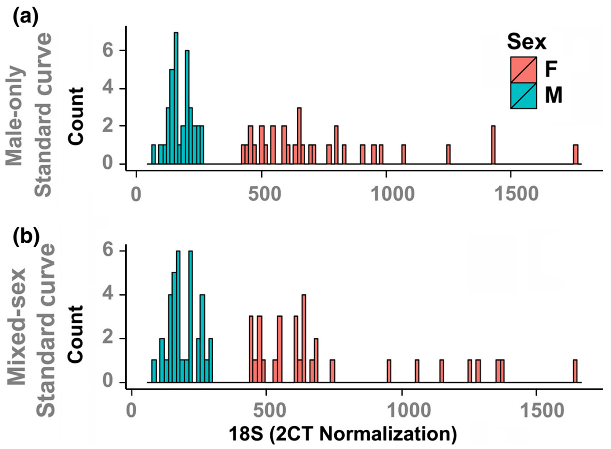

Fig. 4. Histograms of the distribution of 18S normalized to GAPDH in *Apalone spinifera* using the comparative $2C_T$ normalization method and two types of standards (male-only or mixed-sex DNA).

SEX ASSESSMENT USING MIXTURE MODELS OF CLUSTERING WHEN DISTRIBUTIONS OVERLAP

To test our approach for cases where the distribution of the values for males and females overlap, we carried out an additional analysis (Appendices E and F) using a testosterone radioimmunoassay data set of snapping turtles for which gonadal sex information was available from laparoscopic examination (Ceballos & Valenzuela 2011). Procedure 1 (discriminant analysis), dividing the data set of 136 individuals into a training set of 46 turtles and a test set of 90 turtles, resulted in a classification error of 11% for the training set and of 13% for the test set (Appendix F). Procedure 2 (clustering analysis), treating individuals as if their sex information was unknown, resulted in the misclassification of 21 of the 136 individuals (classification error = 15%). Removing individuals from the data set whose classification uncertainty exceeded 0·05 (for whom, the sex typing was less than 95% certain according to the mixture model) improved the classification rate to 90% (error rate = 10%).

Discussion

QPCR QUANTIFICATION

If genomic DNA is to be used to create the qPCR standard curve, our results indicate that the comparative C_T method ($2^{-\Delta C_T}$) is the simplest method to apply and perhaps preferable to alternative methods of normalization (see Appendix A for a comparison of the merits and results of alternative normalization methods) for *A. spinifera*, because once the qPCR is optimized, it requires no pre-knowledge of the sex of any individual. However, using samples of known sex would still be beneficial for validation. Additionally, using standard curves is important to evaluate if qPCR conditions are similar and optimal for the gene of interest and the endogenous control gene. Our approach can distinguish between male and female *A. spinifera* with as little as 5 ng of high-quality genomic DNA (for duplicate reactions of 2·5 ng each), which could be extracted from a blood draw or a small tissue clip, and would permit the sexing of embryos, hatchlings or juveniles in a variety of studies. For instance, sexing *A. spinifera* embryos would enable tests of the effect of temperature, sex and their interaction in developmental studies of gene expression, which were precluded in previous studies of sex determination in this species (Valenzuela, LeClere & Shikano 2006; Valenzuela & Shikano 2007; Valenzuela 2008a,b; Valenzuela, Neuwald & Literman 2013). Although not tested directly, template quality (DNA degradation) should only have a minor effect given that the amplicon from the qPCR is a small product of only ~150 bp for both genes and should amplify even if the DNA is degraded. However, a quality check should be carried out after DNA extraction to test the integrity of the DNA for qPCR. The presence of PCR inhibitors would affect both the 18S and housekeeping genes, but it would be expected that their ratio (and thus our method) should remain unaffected. The 18S primers used in this study were designed in a highly conserved region such that they should work across a wide gamut of animals from insects to vertebrates. However, GAPDH DNA sequences are more variable among taxa such that species-specific primers need to be designed for other studies.

ANALYTICAL METHOD FOR SEX IDENTIFICATION

Our test using *A. spinifera* soft-shell turtles demonstrates the utility of our analytical approach to sex-type individuals under two possible scenarios: (1) when sex information is available for a subset of individuals and (2) when all individuals are of unknown sex. Results indicated that when applied to the sexually dimorphic NOR region of the *A. spinifera* genome, the use of mixture models and univariate discrimination exhibited high classification rates (100%), low error rates during cross-validation (0%) and high discrimination power even when individuals were treated as unknowns (100%). While this is not surprising because our data set contained values with a non-overlapping bimodal distribution, our findings corroborate that the distributions estimated by the mixture models did not create an artificial overlap of values between males and females where none existed.

The error rate was higher when testing the *C. serpentina* data set, whose male and female hormonal values overlap (Appendices E and F), than our results for *A. spinifera*. However, the *C. serpentina* results using our approach compare well with those of previous studies in other turtles using continuous traits with overlapping distributions, such as to those in *Podocnemis expansa* using geometric morphometrics [e.g. 75–90% correct cross-validation (Valenzuela *et al.* 2004; Ceballos, Hernandez & Valenzuela 2014)]. These findings are important because there is no guarantee that further sampling, or data generated by other researchers from soft-shell turtles or from other species with sexually dimorphic NORs, will not contain overlapping values of 18S copy number between males and females. Thus, it is important to have in place an analytical method that is flexible in its application for all possible potential circumstances. Additionally, the level of overlap of the male and female distributions is also affected by the normalization method and composition of the standard curve (whether containing DNA from both sexes or only the sex with the lower values of the continuous trait) (Appendix A). Our results in *C. serpentina* demonstrate that our analytical method is efficient for sex typing when distributions overlap.

Conclusions

As the ZZ/ZW sex chromosome system present in *A. spinifera* has remained virtually unchanged because the split of *Apalone* and *Pelodiscus* ~95 million years ago (mya) (Kawai *et al.* 2007; Badenhorst *et al.* 2013), our sexing technique should be widely applicable to other *Apalone* and *Pelodiscus* species. Furthermore, our approach should also apply to a wide variety of species that exhibit sexually dimorphic NORs. Among those, there are species where the NORs also differ in size between the two sex chromosomes [*Hoplias malabaricus* fish (Born & Bertollo 2000); medaka fish *Oryzias hubbsi* and *O. javanicus* (Takehana *et al.* 2012); and *Bufo marinus* toads (Abramyan *et al.* 2009)]. In some other taxa, the NOR is present in the X chromosome and absent in the Y [*Staurotypus salvini* turtles (Bickham & Rogers 1985), *Gastrotheca riobambae* frogs (Schmid *et al.* 1983), long-nosed potoroo *Potorous tridactylia* and *Carollia perspicillata* bats (Goodpasture & Bloom 1975; Hsu, Spirito & Pardue 1975)] or present in both X in diploid females and in the single X of haploid males [potato aphids *Macrosiphum euphorbiae* (Monti, Manicardi & Mandrioli 2011)]. In yet others, the NOR is present in the Z, but not in the W [*Buergeria buergeri* frogs (Schmid *et al.* 1993)]. Additionally, the use of digital PCR (dPCR) (Vogelstein & Kinzler 1999) is likely to make our approach even more powerful.

Notably, the use of mixture models is an alternative to identify individual sex based on any continuous variable such as circulating hormone levels which have been used to identify sex in chickens (Weissmann *et al.* 2013) and reptiles with temperature-dependent sex determination such as *Chelonia mydas* sea turtles (Owens *et al.* 1978), *Gopherus agassizii* desert tortoises (Rostal *et al.* 1994), *C. serpentina* (Ceballos & Valenzuela 2011) and Amazonian giant river turtles *P. expansa* (Lance, Valenzuela & von Hildebrand 1992; Valenzuela 2001). Our additional example analysis on *C. serpentina* demonstrates its utility for sex typing using hormonal data and under overlapping distributions (Appendices E and F). Similarly, gene expression levels are also a continuous trait amenable to analysis by this approach and have been used to sex-type bovine blastocysts (Hamilton *et al.* 2012). Even behaviour, such as the *fee* glissando components of songs in male black-capped chickadee *Poecile atricapillus,* provides a continuous trait for sex typing (Hahn, Krysler & Sturdy 2013) that could be analysed by mixture models. Importantly, mixture models are not restricted to univariate discrimination but can be applied equally to multivariate data such as shape which can be quantified by geometric morphometrics as carried out to sex type the giant Amazonian river turtle *P. expansa* and painted turtle *Chrysemys picta* hatchlings (Valenzuela *et al.* 2004; Ceballos & Valenzuela 2011; Ceballos, Hernandez & Valenzuela 2014) or by linear measurements as in *Lepidochelys olivacea* sea turtles (Michel-Morfin, Gómez Muñoz & Navarro Rodríguez 2001).

Thus, both the genomic region and the analytical approach proposed here should be broadly applicable for sex typing beyond soft-shell turtles.

Acknowledgements

All procedures were approved by the IACUC of Iowa State University. We thank D.C. Adams and the members of the D.C. Adams and J. Serb labs at Iowa State University for their comments. This work was partially funded by Grant MCB 1244355 to NV.

Data accessibility

All data used in this manuscript are present in the manuscript and its supporting information.

References

Abramyan, J., Feng, C.-W. & Koopman, P. (2009) Cloning and expression of candidate sexual development genes in the cane toad (*Bufo marinus*). *Developmental Dynamics*, **238**, 2430–2441.

Abramyan, J., Ezaz, T., Graves, J.A.M. & Koopman, P. (2009) Z and W sex chromosomes in the cane toad (*Bufo marinus*). *Chromosome Research*, **17**, 1015–1024.

Akyuz, B., Ertugrul, O., Kaymaz, M., Macun, H.C. & Bayram, D. (2010) The effectiveness of gender determination using polymerase chain reaction and radioimmunoassay methods in cattle. *Theriogenology*, **73**, 261–266.

Alasaad, S., Fickel, J., Soriguer, R.C., Sushitsky, Y.P. & Chelomina, G. (2013) Quantitative sexing (Q-sexing) technique for animal sex-determination based on X chromosome-linked loci: empirical evidence from the Siberian tiger. *African Journal of Biotechnology*, **12**, 14–18.

Badenhorst, D., Stanyon, R., Engstrom, T. & Valenzuela, N. (2013) A ZZ/ZW microchromosome system in the spiny softshell turtle, *Apalone spinifera*, reveals an intriguing sex chromosome conservation in Trionychidae. *Chromosome Research*, **21**, 137–147.

Ballester, M., Castello, A., Ramayo-Caldas, Y. & Folch, J.M. (2013) A quantitative real-time PCR method using an X-linked gene for sex typing in pigs. *Molecular Biotechnology*, **54**, 493–496.

Baudry, J.P., Raftery, A.E., Celeux, G., Lo, K. & Gottardo, R. (2010) Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, **19**, 332–353.

Bickham, J.W. & Rogers, D.S. (1985) Structure and variation of the nucleolus organizer region in turtles. *Genetica*, **67**, 171–184.

Born, G.G. & Bertollo, L.A.C. (2000) An XX/XY sex chromosome system in a fish species, *Hoplias malabaricus*, with a polymorphic NOR-bearing X chromosome. *Chromosome Research*, **8**, 111–118.

Bull, J.J. & Vogt, R.C. (1979) Temperature-dependent sex determination in turtles. *Science*, **206**, 1186–1188.

Ceballos, C.P., Hernandez, O.E. & Valenzuela, N. (2014) Divergent sex-specific plasticity and the evolution of sexual dimorphism in long-lived vertebrates. *Evolutionary Biology*, **41**, 81–98.

Ceballos, C.P. & Valenzuela, N. (2011) The role of sex-specific plasticity in shaping sexual dimorphism in a long-lived vertebrate, the snapping turtle *Chelydra serpentina*. *Evolutionary Biology*, **38**, 163–181.

Ceballos, C.P., Adams, D.C., Iverson, J.B. & Valenzuela, N. (2012) Phylogenetic patterns of sexual size dimorphism in turtles and their implications for Rensch's rule. *Evolutionary Biology*, **40**, 194–208.

Ellegren, H. & Sheldon, B.C. (1997) New tools for sex identification and the study of sex allocation in birds. *Trends in Ecology & Evolution*, **12**, 255–259.

Ernst, C.H. & Lovich, J.E. (2009) *Turtles of the United States and Canada*, 2nd edn. John Hopkins University Press, Baltimore, MD.

Ezaz, T., Quinn, A.E., Miura, I., Sarre, S.D., Georges, A. & Graves, J.A.M. (2005) The dragon lizard *Pogona vitticeps* has ZZ/ZW micro-sex chromosomes. *Chromosome Research*, **13**, 763–776.

Fraley, C. & Raftery, A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611–631.

Fraley, C., Raftery, A. & Scrucca, L. (2012) mclust version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Technical Report No. 597, Department of Statistics, University of Washington.

Goodpasture, C. & Bloom, S.E. (1975) Visualization of nucleolar organizer regions in mammalian chromosomes using silver staining. *Chromosoma*, **53**, 37–50.

Griffiths, R. (2000) Sex identification in birds. *Seminars in Avian and Exotic Pet Medicine*, **9**, 14–26.

Gross, T.S., Crain, D.A., Bjorndal, K.A., Bolten, A.B. & Carthy, R.R. (1995) Identification of sex in hatchling loggerhead turtles (*Caretta caretta*) by analysis of steroid concentrations in chorioallantoic/amniotic fluid. *General and Comparative Endocrinology*, **99**, 204–210.

Hahn, A.H., Krysler, A. & Sturdy, C.B. (2013) Female song in black-capped chickadees (*Poecile atricapillus*): Acoustic song features that contain individual identity information and sex differences. *Behavioural Processes*, **98**, 98–105.

Hamilton, C.K., Combe, A., Caudle, J., Ashkar, F.A., Macaulay, A.D., Blondin, P. & King, W.A. (2012) A novel approach to sexing bovine blastocysts using male-specific: gene expression. *Theriogenology*, **77**, 1587–1596.

Hsu, T.C., Spirito, S.E. & Pardue, M.L. (1975) Distribution of 18 + 28S ribosomal genes in mammalian genomes. *Chromosoma*, **53**, 25–36.

Janzen, F.J. (1993) The influence of incubation temperature and family on eggs, embryos, and hatchlings of the smooth softshell turtle (*Apalone mutica*). *Physiological Zoology*, **66**, 349–373.

Kawai, A., Nishida-Umehara, C., Ishijima, J., Tsuda, Y., Ota, H. & Matsuda, Y. (2007) Different origins of bird and reptile sex chromosomes inferred from comparative mapping of chicken Z-linked genes. *Cytogenetic and Genome Research*, **117**, 92–102.

Korstian, J.M., Hale, A.M., Bennett, V.J. & Williams, D.A. (2013) Advances in sex determination in bats and its utility in wind-wildlife studies. *Molecular Ecology Resources*, **13**, 776–780.

Lance, V.A., Valenzuela, N. & von Hildebrand, P. (1992) A hormonal method to determine sex of hatchling giant river turtles, *Podocnemis expansa*: application to endangered species. *The Journal of Experimental Zoology*, **270**, 16A.

Livak, K.J. & Schmittgen, T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method. *Methods*, **25**, 402–408.

Michel-Morfin, J.E., Gómez Muñoz, V.M. & Navarro Rodríguez, C. (2001) Morphometric model for sex assessment in hatchling olive ridley sea turtles. *Chelonian Conservation and Biology*, **4**, 53–58.

Monti, V., Manicardi, G.C. & Mandrioli, M. (2011) Cytogenetic and molecular analysis of the holocentric chromosomes of the potato aphid *Macrosiphum euphorbiae* (Thomas, 1878). *Comparative Cytogenetics*, **5**, 163–172.

Morinha, F., Cabral, J.A. & Bastos, E. (2012) Molecular sexing of birds: a comparative review of polymerase chain reaction (PCR)-based methods. *Theriogenology*, **78**, 703–714.

Mrosovsky, N. (1982) Sex-ratio bias in hatchling sea turtles from artificially incubated eggs. *Biological Conservation*, **23**, 309–314.

Owens, D.W., Hendrickson, J.R., Lance, V. & Callard, I.P. (1978) Technique for determining sex of immature *Chelonia mydas* using a radioimmunoassay. *Herpetologica*, **34**, 270–273.

Phillips, B.C. & Edmands, S. (2012) Does the speciation clock tick more slowly in the absence of heteromorphic sex chromosomes? *BioEssays*, **34**, 166–169.

Rostal, D.C., Lance, V.A., Grumbles, J.S. & Alberts, A.C. (1994) Seasonal reproductive cycle of the desert tortoise (*Gopherus agassizii*) in the eastern Mojave Desert. *Herpetological Monographs*, **8**, 72–82.

Schmid, M., Haaf, T., Geile, B. & Sims, S. (1983) Chromosome-banding in Amphibia. VIII. An unusual XY/XX-sex chromosome system in *Gastrotheca riobambae* (Anura, Hylidae). *Chromosoma*, **88**, 69–82.

Schmid, M., Ohta, S., Steinlein, C. & Guttenbach, M. (1993) Chromosome-banding in Amphibia. 19. Primitive ZW/ZZ sex-chromosomes in *Buergeria buergeri* (Anura, Rhacophoridae). *Cytogenetics and Cell Genetics*, **62**, 238–246.

Shaw, P.J. & McKeown, P.C. (2011) The Structure of rDNA chromatin. *The Nucleolus* (ed. M.O.J. Olson), pp. 43–55. Springer, New York.

Singh, A.K. (2013) Introduction of modern endocrine techniques for the production of monosex population of fishes. *General and Comparative Endocrinology*, **181**, 146–155.

Takehana, Y., Naruse, K., Asada, Y., Matsuda, Y., Shin-I, T., Kohara, Y., Fujiyama, A., Hamaguchi, S. & Sakaizumi, M. (2012) Molecular cloning and characterization of the repetitive DNA sequences that comprise the constitutive heterochromatin of the W chromosomes of medaka fishes. *Chromosome Research*, **20**, 71–81.

Valenzuela, N. (2001) Constant, shift and natural temperature effects on sex determination in *Podocnemis expansa* turtles. *Ecology*, **82**, 3010–3024.

Valenzuela, N. (2008a) Evolution of the gene network underlying gonadogenesis in turtles with temperature-dependent and genotypic sex determination. *Integrative and Comparative Biology*, **48**, 476–485.

Valenzuela, N. (2008b) Relic thermosensitive gene expression in genotypically-sex-determined turtles. *Evolution*, **62**, 234–240.

Valenzuela, N. (2010) Multivariate expression analysis of the gene network underlying sexual development in turtle embryos with temperature-dependent and genotypic sex determination. *Sexual Development*, **4**, 39–49.

Valenzuela, N., LeClere, A. & Shikano, T. (2006) Comparative gene expression of steroidogenic factor 1 in *Chrysemys picta* and *Apalone mutica* turtles with temperature-dependent and genotypic sex determination. *Evolution and Development*, **8**, 424–432.

Valenzuela, N., Neuwald, J.L. & Literman, R. (2013) Transcriptional evolution underlying vertebrate sexual development. *Developmental Dynamics*, **242**, 307–319.

Valenzuela, N. & Shikano, T. (2007) Embryological ontogeny of Aromatase gene expression in *Chrysemys picta* and *Apalone mutica* turtles: comparative patterns within and across temperature-dependent and genotypic sex-determining mechanisms. *Development, Genes and Evolution*, **217**, 55–62.

Valenzuela, N., Adams, D.C., Bowden, R.M. & Gauger, A.C. (2004) Geometric morphometric sex estimation for hatchling turtles: a powerful alternative for detecting subtle sexual shape dimorphism. *Copeia*, **2004**, 735–742.

Vogelstein, B. & Kinzler, K.W. (1999) Digital PCR. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 9236–9241.

Weissmann, A., Reitemeier, S., Hahn, A., Gottschalk, J. & Einspanier, A. (2013) Sexing domestic chicken before hatch: a new method for in ovo gender identification. *Theriogenology*, **80**, 199–205.

Wood, J.R., Wood, F.E., Critchley, K.H., Wildt, D.E. & Bush, M. (1983) Laparoscopy of the green sea turtle, *Chelonia mydas*. *British Journal of Herpetology*, **6**, 323–327.

Yntema, C.L. & Mrosovsky, N. (1980) Sexual-differentiation in hatchling loggerheads (*Caretta caretta*) incubated at different controlled temperatures. *Herpetologica*, **36**, 33–36.

Received 15 May 2014; accepted 3 July 2014
Handling Editor: Michael Bunce

Supporting Information

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

Appendix A. Copy Number Quantification by real-time qPCR.

Appendix B. Flow chart illustrating the analytical method proposed here to sex-type individuals using mixture models applied to continuous data.

Appendix C. R code example using *Apalone spinifera* 18S copy number quantified by QPCR using GAPDH as normalizer, a male-only DNA standard curve, and the standard curve normalization method of 18S quantification.

Appendix D. Example dataset of *Apalone spinifera* 18S copy number (“*mydata*” in Appendix C) quantified by QPCR using GAPDH as normalizer, a male-only DNA standard curve, and the standard curve normalization method of 18S quantification.

Appendix E. Example of discriminant and clustering analyses using a dataset of *Chelydra serpentina* circulating testosterone levels measured by radioimmunoassay in individuals with reliable information of gonadal sex determined by laparoscopy from Ceballos and Valenzuela (2011).

Appendix F. Classification of *Chelydra serpentina* individuals as male or female using mixture models on a dataset of circulating testosterone whose values overlap between the sexes (Ceballos and Valenzuela 2011).