2015

Studies on the causes and consequences of sex determination in turtles

Srihari Radhakrishnan
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd

Part of the Bioinformatics Commons, Developmental Biology Commons, Evolution Commons, and the Genetics Commons

Recommended Citation
http://lib.dr.iastate.edu/etd/14959

This Dissertation is brought to you for free and open access by the Graduate College at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Studies on the causes and consequences of sex determination in turtles

by

Srihari Radhakrishnan

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Nicole Valenzuela, Co-major Professor
Peng Liu, Co-major Professor
Amy L. Toth
Eric R. Henderson
Matthew Hufford

Iowa State University
Ames, Iowa
2015

Copyright © Srihari Radhakrishnan, 2015. All rights reserved.
# TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: GONADAL TRANSCRIPTOMES OF THE PAINTED AND SOFT-SHELL TURTLES
WITH TEMPERATURE-DEPENDENT AND GENOTYPIC SEX DETERMINATION ILLUMINATES
THE EVOLUTION OF THE GENE NETWORK FOR PRIMARY SEXUAL DEVELOPMENT 5
   2.1 Introduction 7
   2.2 Results 9
   2.3 Discussion 14
   2.4 Materials and Methods 31
   2.5 Tables and Figures 35

CHAPTER 3: TEMPERATURE INDUCES DIMORPHIC GENOME-WIDE DNA METHYLATION IN
TURTLES WITH ENVIRONMENTAL SEX DETERMINATION 51
   3.1 Introduction 53
   3.2 Results 55
   3.3 Discussion 61
   3.4 Materials and Methods 70
   3.5 Tables and Figures 74

CHAPTER 4: RELATIVE RATES OF SEX-LINKED AND AUTOSOMAL CODING SEQUENCE
EVOLUTION MEASURED IN VERTEBRATES 84
   4.1 Introduction 86
   4.2 Methods 89
   4.3 Results 92
   4.4 Discussion 95
   4.5 Tables and Figures 106

CHAPTER 5: CONCLUSION 116

BIBLIOGRAPHY 120

ACKNOWLEDGEMENTS 142
CHAPTER 1: INTRODUCTION

Vertebrates exhibit a variety of sex-determining mechanisms spanning two main kinds: genotypic sex determination (GSD) and temperature-dependent sex determination (TSD). In most GSD systems, offspring sex is determined by its sex chromosomes - the genotype of an individual at conception controls its gonadal development into either testes or ovaries. In TSD, the environmental temperature experienced during a period of embryonic development (thermosensitive period) directs the male and female fate (Bull 1983; Rhen and Schroeder 2010; Deeming and Ferguson 1991; Valenzuela and Lance 2004). Many vertebrates including mammals have GSD, while TSD is common in reptiles including all crocodilians, most turtles, some lizards and tuataras.

While sex chromosomes and their role in sex determination was proposed over a century ago in independent insect studies (Stevens 1905; Wilson 1905), it was not until many decades later, that Susumu Ohno laid out a theoretical framework on the progression of sex chromosomes in vertebrates (Ohno 1967). Around the same time, the curious effect of temperature on lizard sex ratios was elucidated (Charnier 1966). Over the next decade after considerable debate about whether TSD was biologically relevant, egg incubation temperature was recognized as an alternative to sex chromosomes in driving sex determination in some vertebrates (Bull and Vogt 1979; Pieau 1972). Since this pivotal finding, the evolution of the diversity of mechanisms to decide sexual fate has defied scientific explanation, and to date, many unanswered questions remain relating to the underlying causes. Is the evolution and maintenance of TSD systems adaptive? What are the key genetic/epigenetic elements that are influenced by temperature in TSD systems
and what is their mechanism of thermal transduction? How do those elements change with the evolution of sex chromosomes in GSD systems? What drives changes in sex determining mechanisms from TSD to GSD and vice versa? These conundrums persist partly due to the fact that the genetic and epigenetic architecture of sex-determining systems remain incompletely understood. Indeed, while much effort has been dedicated to studying the gene networks in the mammalian urogenital pathway (Liu et al. 2010; Wainwright and Wilhelm 2010), the full extent of the similarities or differences of the reptilian network remain unknown (Valenzuela 2008a). This gap precludes our thorough understanding of the molecular basis and evolution of vertebrate sex determination.

Further, as a consequence of sex chromosomes differentiating from ancestral autosomes (Muller 1914), their evolutionary rate can be distinct from that of autosomes (Charlesworth et al. 1987). This divergence in sequence context (sex-linked versus autosomal) has also raised multiple questions. Do genes with sex-biased fitness effects tend to accumulate on sex chromosomes? Is there a difference in mutation rate and strength of natural selection between sex chromosomes and autosomes (Vicoso and Charlesworth 2006)? Multiple studies have delved into the relative rates of evolution of sex-linked and autosomal sequences in animals (Thornton et al. 2006; Lu and Wu 2005; Hvislom et al. 2012; Betancourt et al. 2002; Ávila et al. 2014; Kousathanas et al. 2014; Halligan et al. 2013). However, the datasets employed in these studies have been limited to non-reptilian systems and thus, these studies paint a picture of sex chromosome evolution that is useful, yet incomplete.

Turtles are an excellent system to study sex determination and sex chromosome evolution because they possess both TSD and GSD – (XX/XY and ZZ/ZW) (Valenzuela and
Lance 2004), and their genomes have undergone lineage-specific reorganization resulting in a diploid chromosomal count ranging between 28 and 68 through a process that is tied to shifts in sex determination but whose functional links remain unknown (Valenzuela and Adams 2011). Further, increasing genomic resources are now available for this group (Janes et al. 2008). Among turtles, one of my focal species, the painted turtle (*Chrysemys picta*) is an emerging model to study sex determination, evolution, ecology and human health (Valenzuela 2009) as it has a recently sequenced genome (Shaffer et al. 2013) that is partially physically mapped (Badenhorst et al. 2015), a BAC library, cytogenetic resources and some transcriptomes (Shaffer et al. 2013). Genomic resources are also available for the green sea turtle and the Chinese softshell turtle (Wang et al. 2013a) providing a comparative framework to carry out genome-wide analyses.

In this dissertation, I contribute to the quest for answers to multiple questions related to the proximate causes (Chapters 2 and 3) and consequences (Chapter 4) of the evolution of sex determination. In chapter 2, I analyze a time series of embryonic gonadal transcriptomes in TSD (*Chrysemys picta*) and GSD (*Apalone spinifera*) turtles to characterize the composition of the transcriptional network that regulates male and female gonadal development, and test how it responds to temperature in TSD and GSD systems. I identify multiple candidate genes and pathways that could potentially drive temperature-dependent sex determination in TSD turtles, and contrast the results with well-characterized mammalian urogenital systems. In chapter 3, I test for the involvement of epigenetic modification in TSD and whether it might potentially mediate the differential gene expression patterns detected via transcriptomic approaches in chapter 2. I first use *in silico* methods to predict the genome-wide DNA methylation landscape in a TSD turtle
(Chrysemys picta), and then test the validity of this approach to predict true methylation status by sequencing the gonadal methylome of male and female C. picta hatchlings. With these data I also test for differential methylation by temperature in these fully differentiated male and female hatchlings, and test whether the differentially methylated regions coincide with those differentially expressed in the transcriptomes of late-stage hatchlings. In chapter 4, I tackle the consequences of the evolution of sex chromosomes by measuring the relative rates of evolution of coding sequences, which are sex-linked in some species, but autosomal in others, across eleven GSD and TSD vertebrates to test the theoretical expectation that sex chromosomes evolve faster than autosomes inevitably. Taken together, these chapters help illuminate the molecular basis underlying sex determination by temperature, and the molecular consequences of sex linkage on genome evolution.
CHAPTER 2: GONADAL TRANSCRIPTOMES OF THE PAINTED AND SOFT-SHELL TURTLES WITH TEMPERATURE-DEPENDENT AND GENOTYPIC SEX DETERMINATION ILLUMINATES THE EVOLUTION OF THE GENE NETWORK FOR PRIMARY SEXUAL DEVELOPMENT

AUTHORS: Srihari Radhakrishnan¹, ⁴, Robert Literman², ⁴, Andrew Severin³ Nicole Valenzuela⁴

AFFILIATIONS:
1: Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011
2: Ecology and Evolutionary Biology Program, Iowa State University, Ames, IA 50011
3: Genome Informatics Facility, Iowa State University, Ames, IA 50011
4: Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, IA 50011

SUMMARY: Sexual fate in vertebrates is established by multiple mechanisms ranging from genotypic sex determination (GSD) where the individual's genotype directs that decision, to temperature-dependent sex determination (TSD) where the embryonic incubation temperature during development directs sexual fate. Turtles exhibit both TSD and GSD, making them ideal to study the evolution of sex determination. Little is known about the molecular basis of TSD and how it differs from GSD. Here we analyze temperature-specific gonadal transcriptomes (RNA sequencing validated by qPCR) of painted turtles (Chrysemys picta - TSD) and soft-shell turtles (Apalone spinifera- GSD) before and during the thermosensitive period for gonadal development in C. picta. We show for the first time in turtles and reptiles that most gene homologs of the mammalian
urogenital network are active during gonadogenesis. Notably, the antagonistic transcription of $\beta$-catenin and insulin-receptor family essential for mouse female and male determination was active in both turtles but thermosensitive only in TSD early embryos. Furthermore, thermosensitive transcription was detected in genes that may help mediate the transduction of the environmental temperature signals into male and female development, such as those related to signaling and hormonal pathways, RNA-binding and heat-shock, among others. Results uncovered numerous new candidate regulators of gonadal formation in TSD turtles as well as putative genes and biological pathways that might set apart TSD and GSD vertebrate systems, thus contributing significantly to our understanding of the evolution of vertebrate sex determination.

**KEYWORDS:** Genotypic sex determination, Temperature-dependent sex determination, time series gonadal transcriptome, differential gene expression
2.1 Introduction

Organisms vary wildly in how they determine sex (Bachtrog et al. 2014). Vertebrate sex-determining mechanisms range between Genotypic Sex Determination (GSD) and Environmental Sex Determination (ESD) (Valenzuela and Lance 2004; Valenzuela et al. 2003). The most common ESD mechanism in vertebrates is Temperature-dependent sex determination (TSD). The commitment of the bipotential gonad to differentiate into testes or ovaries is triggered by the genotype in GSD, and by temperatures experienced during the thermosensitive period of embryonic development in TSD (Deeming and Ferguson 1991; Rhen and Schroeder 2010; Valenzuela and Lance 2004; Bull 1983). All studied mammals, birds and amphibians exhibit GSD, while TSD is found in some fish, tuatara, all crocodilians, most turtles and some lizards (Valenzuela and Lance 2004; Tree of Sex Consortium 2014). This diversity has defied scientific explanation (Bachtrog et al. 2014), partly because our understanding of the molecular basis of GSD and TSD is incomplete. For instance, the key genetic elements that mediate the effect of temperature in TSD systems remain elusive. Unlike GSD models such as mammals (Liu et al. 2010; Wainwright and Wilhelm 2010; Jameson et al. 2012; Munger et al. 2013) and chicken (Ayers et al. 2013) whose gonadal developmental pathways are well understood (albeit not fully) our knowledge for GSD reptiles and TSD species is incipient. This gap prevents understanding the genetic architecture of sex determination and its evolution in vertebrates. Past candidate-gene studies on GSD and TSD turtles using quantitative PCR and in situ hybridization targeted a number of genes underlying sexual development, including \textit{Wt1} (Valenzuela 2008b; Spotila et al. 1998), \textit{Sf1} (Valenzuela et al. 2006; Ramsey et al. 2007), \textit{Dax1} (Torres Maldonado et al. 2003; Valenzuela 2008a), \textit{Sox9} (Torres Maldonado et al. 2003; Barske and Capel 2010;
Valenzuela 2010; Matsumoto et al. 2013b), Aromatase (Valenzuela and Shikano 2007; Matsumoto et al. 2013a), Dmrt1 (Torres Maldonado et al. 2003; Kettlewell et al. 2000; Valenzuela 2010), Estrogen receptor (Bergeron et al. 1998; Chávez et al. 2009), Rspo1 (Matsumoto et al. 2013b) among others (Table A1). Each of those studies provide valuable but fragmentary information on the expression of one or a few genes at a time (Torres Maldonado et al. 2003; Rhen et al. 2007; Valenzuela et al. 2013; Matsumoto et al. 2013b), but deciphering the composition, environmental sensitivity, and evolution of the gonadal gene network in TSD and GSD turtles is overdue.

Here, we use a comparative approach to test for transcriptional responses to incubation temperature (or lack thereof) at several stages of embryonic development in two turtle species, the painted turtle Chrysemys picta (TSD) and the soft-shell turtle Apalone spinifera (GSD), hereafter denoted as Chrysemys and Apalone respectively. Our RNAseq approach provides the first glimpse of the full transcriptional network in closely related reptiles with contrasting sex-determining mechanisms (GSD and TSD), and between turtles and model mammalian developing gonads. Results uncovered genes whose expression in the developing gonad differs between high and low temperatures that produce males or females in TSD turtles across developmental stages before, during and after the activation of the thermosensitive period (or TSP) (Bull and Vogt 1981) and under identical incubation conditions in GSD turtles. We identify novel candidate genes whose early differential expression suggest that they may contribute to transmitting the temperature signal to the developmental pathway, potentially helping determine the sexual fate in TSD turtles. Other candidate genes were also identified in turtles that are yet to be
described in the gonadal regulatory network of mice and chicken, and may help guide research to test their potential involvement in gonadogenesis in these model vertebrates.

2.2 Results

Transcriptome assembly: RNAseq data was obtained from Chrysemys embryonic tissues at stages 9 and 12 (before the TSP), 15 (onset of the TSP), 19 and 22 (mid and late TSP) from male-producing temperature (MPT = 26°C), and female-producing temperature (FPT = 31°C) (Valenzuela et al. 2013). Identical incubation conditions and sampling scheme were followed for Apalone. De novo transcriptome assemblies constructed using Trinity (Grabherr et al. 2011) resulted in a high percentage of mapped reads (>92%), and high representation of human Core Eukaryotic Genes (CEGs) (>77%) and mammalian urogenital development pathway genes in both species (>96%) (Table 2.1). All subsequent analyses reported here are based on the de novo transcriptome assemblies. We also tested the alternative approach using reference genome-guided assemblies. However this approach was discarded because while the mapping rate of the Chrysemys libraries to the Chrysemys reference genome (Shaffer et al. 2013) was high (97%), that of Apalone was only 44% (Table 2.1) and resulted in significantly fewer gene models for Apalone (5,596 unique to Apalone, 14,661 unique to Chrysemys, and 23,465 overlapping). The problem remained when using the genome of the close relative of Apalone, Pelodiscus sinensis, as a reference (Wang et al. 2013b), because the P. sinensis assembled genome is more fragmentary than Chrysemys's, as evidenced by the lack of complete exonic sequences for several genes (such as some homologs of mammalian urogenital genes - results not shown).
Normalization schemes to identify differentially expressed genes per species:

Gene read abundance was normalized multiple ways, first to the housekeeping genes Tfr and Hprt1 which were constitutively expressed across all developmental stages in both Chrysemys and Apalone, and then by the standard upper-quartile normalization (in a scheme we call UQHK100) (Bullard et al. 2010). The order of these steps did not affect the overall assessment of gene expression. When compared to other normalization procedures described in the methods (UQ100 = upper-quartile only, UQ99 = upper-quartile after eliminating the top 1 percentile of transcripts with the highest expression, UQHK99 = upper-quartile and house-keeping gene normalization after eliminating the top 1 percentile of transcripts with highest expression), the chosen UQHK100 approach resulted in fewer differentially expressed genes than using the upper-quartile alone (Fig. 2.1a), and therefore, it is more conservative. Furthermore, UQHK100 normalization revealed differential expression patterns which were most consistent with extensive qPCR data of several candidate genes previously obtained for Chrysemys (Valenzuela et al. 2013), as determined qualitatively by visual inspection of the expression profiles over developmental time for individual genes. Therefore, we used UQHK100 to identify differentially expressed genes for further enrichment analyses to ensure unbiased comparisons between species.

Gene annotation and genes of interest: Between 26% and 28% of the de novo transcripts per species were represented in the SwissProt protein database. The Chrysemys transcriptome showed an overall higher representation of annotated genes in the Chrysemys genome (Shaffer et al. 2013) than the Apalone transcriptome (Table 2.2). Of the
transcripts not annotated in SwissProt, 252 from *Chrysemys* and 169 from *Apalone* correspond to non-coding RNA sequences (ncRNAs) as identified by BLAST (Camacho et al. 2009). To search for candidates which might potentially help transduce the temperature signals into sex-specific development and thus may have a significant role in TSD based on their differential expression pattern by temperature, we focused on known genes involved in (a) vertebrate sex determination/differentiation in model mammals and birds (Morrish and Sinclair 2002; Valenzuela and Lance 2004; Valenzuela et al. 2013), (b) epigenetic modification (Kuroki et al. 2013), (c) hormonal pathways (Carmi et al. 1998) and (d) general sensing responses (Kohno et al. 2010), out of the variety of annotated genes. These target categories included heat-shock genes, transient receptor potential genes, germ-line and histone-related genes, androgen- and estrogen related genes and genes linked to human/chicken sex chromosomes (http://www.ensembl.org/info/data/ftp/index.html). While the overall composition of the transcriptomes of the two turtle species was similar with regard to these categories, a few genes present in *Chrysemys*’s transcriptome were notably absent in *Apalone* across all stages, including some genes X-linked in human and Z-linked in chicken (Table 2.3). Further, some genes were differentially expressed by temperature in both turtles (Tables A2-A11). Interestingly, a number of genes that are involved in histone modification show low temperature bias (MPT) just before the onset of the thermosensitive period in *Chrysemys* but high-temperature bias in *Apalone*: histone H1-x like protein (H1x-like), histone chaperone protein (*Asf1B*-like), H3-Histone family 3B (*H3f3b*) and Nuclear autoantigenic sperm protein (*Nasp*). Details about the transcriptional response of these genes of interest are presented in the discussion.
**Differential expression in painted turtles (TSD):** *Chrysemys* transcription patterns measured by RNA-seq per temperature and stage recapitulated those previously detected by qPCR (Valenzuela et al. 2013) when differences in expression were large, whereas qPCR identified smaller expression differences at certain embryonic stages that passed undetected in our transcriptomes (Fig. 2.3). We then chose a highly conservative p-value cutoff value of $1 \times 10^{-10}$ to further correct for false positives in our differential expression tests. This approach revealed significant overlap of highly differentially expressed genes across developmental stages for both species (Figs. 2.1b, c). The same differentially expressed genes were also recovered with a second approach where reads from each RNAseq library were randomly subdivided into multiple representative subsamples (Liu et al. 2014) which were then used in the differential expression analysis using both DESeq and EdgeR toolkits (Anders 2012; Robinson et al. 2010). Among these, we identified 1065 genes that were differentially expressed only in *Chrysemys* across development (Table A15). Some results of particular interest are highlighted below.

Our *Chrysemys* RNA-seq data corroborated qPCR results of known sex-determining gene homologs in *Chrysemys: Sf1, Wt1, Sox9, Aromatase* and *Dax1* (Valenzuela et al. 2013) which serve to validate our transcriptomes. RNA-seq also permitted profiling a number of candidate sex-determining genes in *Chrysemys* and *Apalone* (Figs. 2.2, A2), which were previously uncharacterized in turtles. A significant number of these genes in *Chrysemys* show MPT-bias late in the TSP (Stage 22), two (*Igf1r* and *Insr*) before the formation of the bipotential gonad (Stage 9), and only one (*β-catenin - Ctnnb1*) shows FPT-bias before the thermosensitive period (Fig. 2.4). Contrastingly, expression patterns in *Apalone* were more variable, shifting between MPT-bias and FPT-bias throughout development.
**Gene Ontology (GO) Enrichment:** At each embryonic stage the differentially expressed genes were enriched for a number of GO pathways. While no GO pathways were consistently deployed in both turtle species throughout development, GO pathways relating to translation and translational elongation were present across three out of the five stages in both taxa. Overall, we found more shared pathways by stage (including general cell functions such as mitotic cycles, mRNA processing and RNA-splicing) between *Chrysemys* and *Apalone* before stage 15 (onset of TSP in *Chrysemys*) than later in development, suggesting that temperature triggers different network modules after the onset of thermosensitive period in *Chrysemys* than in *Apalone*. Indeed, enriched pathways during stages 9 and 12 in both turtles, including intracellular transport, protein localization and protein catabolic processes remained enriched only in *Chrysemys* after stage 15 (Table A12). Contrastingly, genes upregulating protein ubiquitination and ubiquitin protein ligase were enriched only in stage 12 of *Apalone* and not in *Chrysemys*.

**Novel transcripts:** Around half of *Chrysemys* (53%, or 150195/279903) and *Apalone* transcripts (54% or 152579/279753) were absent in SwissProt or ncRNA databases. However, 87% (131,131) of *Chrysemys* transcripts were mapped to the *Chrysemys* genome, of which only 7% (10,660) were unannotated, indicating a gap between the SwissProt/ncRNA databases and the annotated *Chrysemys* genome. Among these, most differentially expressed transcripts are MPT-biased at stages 9, 19 and 22 (Fig. A1). Only 50% of *Apalone* novel transcripts could be mapped to the *Chrysemys* genome and 26% to the *Trachemys scripta* whole-body transcriptome (Kaplinsky et al. 2013), while 68% of *Chrysemys* novel transcripts mapped to *T. scripta*. Since this mapping was carried out with
low stringency, this difference is likely be due to the absence of many novel transcripts in the GSD turtle, as extensive divergence is not expected in coding sequences between Chrysemys and Apalone, given that transcripts from Chrysemys and Apalone are more often identical than not (data not shown).

**Gene clustering and coexpression:** The genes of interest (Table 2.3) clustered into modules of co-expression patterns across embryonic stages that differ between turtles. Interestingly, Chrysemys showed stronger clustering differences between temperatures, with more gene modules at 26°C (45) than at 31°C (10) (Fig. 2.5a, 2.5b, A5a, A5b). Contrastingly, Apalone did not differ significantly between temperatures in the number of gene modules (Fig. 2.5c, 2.5d, A5c, A5d). Similar species-specific clustering differences were also detected for 157 core eukaryotic genes (data not shown). The composition of the largest clusters also varied by temperature in terms of GO pathways (Table 2.4).

### 2.3 Discussion

The evolution of sex determination remains an evolutionary enigma. Turtles are an ideal vertebrate study system since TSD and GSD co-occur in this group (Valenzuela and Adams 2011). While whole-body embryonic transcriptomes were recently characterized in turtles (Kaplinsky et al. 2013) relatively little is known about the gene network controlling urogenital development. Here we characterize the full composition of this network using RNAseq and its transcriptional response to incubation temperature in developing gonads of TSD and GSD turtles. Our results illuminate the genetic architecture of vertebrate gonadogenesis, which remains incompletely known even for humans. Ours is the first
transcriptomic time series of gonadal development in turtles and the first transcriptomic contrast between sex-determining mechanisms. *Apalone* remains the only GSD reptilian genus studied for embryonic urogenital transcription (this study, Valenzuela et al. 2006; Valenzuela and Shikano 2007; Valenzuela 2008a, 2008b, 2010). Temperature does not affect sex ratios in *Apalone* (Bull and Vogt 1979) such that this species serves as a negative control for TSD responses. Also, because molecular sexing was unavailable for *Apalone* when data were collected (Literman et al. 2014) only temperature effects (and no sex effects) could be analyzed here. Overall, our data suggest that the transcriptional circuitry underlying gonadogenesis in TSD and GSD turtles is broadly the same, and that differences between these mechanisms are largely due to the differential deployment of these common elements as detailed below.

**Transcriptome assembly:** The genome-guided transcriptome assembly using *Chrysemys* as reference (Shaffer et al. 2013) worked well for *Chrysemys* but produced poor results for *Apalone* (44% mapped reads and fewer gene models; Table 2.1), underscoring the extensive divergence accrued in these turtle genomes since their lineages split >180mya (Valenzuela and Adams 2011). Using the *Pelodiscus sinensis* genome as reference (Wang et al. 2013b) did not solve this problem, and comparative approaches require common analyses for all species. However, *de novo* transcriptome assemblies had similar high quality and permitted the discovery of novel transcripts previously unidentified in public databases.

Contrastingly, the *Chrysemys* genome (Shaffer et al. 2013), was useful to map the transcripts from the *de novo* assemblies (since transcripts are longer and align better than
reads) to quantify the representation of annotated genes per library (Table 2.2). The *Chrysemys* transcriptome had an unsurprising slightly higher representation of annotated genes overall (62%) than *Apalone* (57%). The *P. sinensis* genome was excluded here because its annotation is less extensive than *Chrysemys's*. These observations highlight the need to improve current turtle genome assemblies (Badenhorst et al. 2015) and to sequence additional genomes from representative phylogenetic lineages to illuminate turtle and vertebrate genome evolution.

**Gene enrichment analysis:** Enrichment analyses of Gene Ontology (GO) categories represented in the transcriptomes using DAVID (Huang et al. 2007) revealed that species shared more pathways before stage 15 overall, except for chromatin organization and chromatin modification pathways which were enriched only at stage 9 in *Chrysemys*, and pathways linked to protein ubiquitination and ubiquitin protein ligases which were enriched only at stage 12 in *Apalone*. Ubiquitination is a post-translational modification that results in protein degradation (Roos-Mattjus and Sistonen 2004). This suggests that temperature triggers a different set of downstream pathways in *Chrysemys* potentially leading to sexual fate determination by temperature. Pathways including intracellular transport, protein localization and protein catabolic processes, were enriched in both species at stages 9 and 12, but remain enriched only in *Chrysemys* after stage 15 (Table A12). Thus, important evolutionary changes may have occurred in GSD turtles in the machinery underlying gonadogenesis preceding the thermosensitive period of TSD turtles (before stage 15) perhaps inactivating genes regulating the male- and female-specific TSD pathways, thus determining sex independent of temperature. These and earlier findings
(reviewed in Valenzuela et al 2013) underscore that key thermosensitive events for sexual development occur in early embryogenesis.

**Differentially expressed genes by temperature:** We searched for genes showing thermosensitive expression in *Chrysemys* (TSD) to uncover candidate temperature sensors or transducers that may activate TSD male and female gonadogenesis. We detected many such genes, including numerous homologs of mammalian sex determination/differentiation genes, plus previously undescribed candidates (Fig. 2.4). Salient results highlighted below reveal that the vertebrate gene network regulating primary sexual development is highly conserved in its composition and is active in turtles, but regulated differently between TSD and GSD turtles, and between turtles, mammals and birds (Fig. 2.2, Table A1).

**Known genes in the turtle gonadal network:** Multiple genes involved in turtle sex determination have been studied, including *Vasa, Dazl, Mis, Foxl2, Dmrt1, Aromatase, Androgen receptor* and *Estrogen receptors* α and β, among others (Bergeron et al. 1998; Ramsey and Crews 2007a; Smith et al. 2008; Bachvarova et al. 2009; Rhen and Schroeder 2010; Shoemaker-Daly et al. 2010). Our transcriptomes recapitulated expression patterns from qPCR for various genes (Valenzuela 2006, Valenzuela and Shikano 2007, Valenzuela 2008a, Valenzuela 2008b, Valenzuela 2010, Valenzuela et al., 2013), although subtle differences passed undetected in the transcriptomes (Fig. 2.3), perhaps because transcriptomic inferences have lower power overall than qPCR approaches (Devonshire et al. 2013; Cristino et al. 2011). To avoid false positive results from the absence of
biologically replicated transcriptomes, and from the potential biased introduced by the lower number of embryos per RNA library for Apalone compared to Chrysemys, we applied a stringent cutoff of 1e-10 to control for false discoveries, and discarded any genes with lower significant differential expression (Fig. 2.4). The following paragraphs highlight the transcriptional patterns of known vertebrate determination/differentiation genes found in our transcriptomes (full gene names are presented in table A17).

**Wt1** is a transcription factor important in the formation of the bipotential gonad, and the maintenance of Sertoli cells and seminiferous tubules in developing testis (Pelletier et al. 1991). qPCR studies detected higher Wt1 transcription at low temperature in *Chrysemys* and *Apalone mutica* (GSD) prior to the thermosensitive period (TSP), and across the TSP in *Chrysemys* (Valenzuela et al. 2013), suggesting a relic thermal sensitivity for Wt1 in GSD turtles (Valenzuela 2008b). Our transcriptomes reveal higher Wt1 expression at low temperature during stage 22 in *Chrysemys* (late-TSP) and *Apalone spinifera* (Apalone hereafter) (Fig. 2.4). Contrastingly, Wt1 expression in mice and chicken gonads is sexually monomorphic through embryogenesis (Jameson et al., 2012, Oréal et al. 2002). Finding differential expression from stage 19-22 Apalone gonads is important because previous studies in *A. mutica* used adrenal-kidney-gonad complexes (AKGs) (Valenzuela et al. 2006; Valenzuela and Shikano 2007; Valenzuela 2008b, 2008a, 2010), and expression from the adrenal-kidney can mask gonadal expression, as occurs in *Chrysemys* and other turtles (Pieau and Dorizzi 2004; Ramsey and Crews 2007b; Shoemaker et al. 2007; Valenzuela et al. 2013).

**Sf1** is required for gonadal and adrenal gland formation and steroidogenic activity (Parker and Schimmer 1997). Sf1 is directly activated by Wt1, and its expression is thermo-
insensitive in GSD turtles measured by RNAseq and qPCR (this study and Valenzuela, 2008b). Our transcriptomes detected strong male-biased expression of Sf1 during- and late-TSP in *Chrysemys* consistent with qPCR data (Valenzuela et al. 2013), as observed also in the slider turtle *Trachemys scripta* (TSD) (Crews et al. 2001), but counter to the monomorphic expression in the snapping turtle *Chelydra serpentina* (TSD) (Rhen et al. 2007). This underscores that Sf1 expression in vertebrates is evolutionarily labile, as it is male-biased in some vertebrates (rat, mouse, pig, trout), female-biased in others (chicken, frogs and fish), and monomorphic in humans (reviewed in Valenzuela et al., 2013).

**Dax1** encodes a nuclear orphan receptor important for mammalian ovarian and testicular formation (Ramkissoon and Goodfellow 1996; Meeks et al. 2003). Our RNA-seq revealed male-biased expression during *Chrysemys* TSP consistent with qPCR results (Valenzuela et al. 2013). In *Apalone*, RNA-seq also revealed Dax1 low-temperature biased expression from stage 15 onward, consistent with expression in *A. mutica* (by qPCR), which seems driven by the relic thermosensitive expression of its activator (*Wt1*) (Valenzuela 2008a). Contrastingly, Dax1 expression is female-biased in birds and monomorphic in several TSD taxa including the green sea turtle *Lepidochelys olivacea*, *C. serpentina* and *T. scripta* (reviewed in Valenzuela et al., 2013).

**Sox9**, is immediately downstream of the *Sry* gene in eutherian mammals, tipping the bipotential gonad towards the male fate. Sox9 shows male-biased expression only at stage 22 in *Chrysemys* transcriptomes, while male-biased transcription was also detected at stages 15 and 19 by qPCR (Valenzuela et al. 2013). Sox9 expression in *Apalone* shifts from high-temperature biased during stage 15 to low-temperature biased during stage 19, perhaps reflecting the evolutionary drift in GSD turtles from the ancestral thermal
response. Consistently, Sox9 expression by qPCR in *A. mutica* was monomorphic (Valenzuela 2010).

*Aromatase* encodes an enzyme that aromatizes androgens to estrogens and is key in ovarian formation and steroidogenic activity (Hughes et al. 1999; Castro et al. 2005). Our transcriptomes revealed female-biased *Aromatase* expression during *Chrysemys*’ late-TSP, consistent with qPCR results (Valenzuela et al., 2013). The timing of *Aromatase* involvement varies across vertebrates, acting late in mice, and earlier in ovarian differentiation in humans and birds (George and Wilson 1978; Smith and Sinclair 2001). The monomorphic aromatase transcription in the bipotential gonad (stages 9 and 12) in turtles observed here is consistent with expression in chicken (Ayers et al. 2013).

*Dmrt1* is a known regulator of sexual development in vertebrates (Morrish and Sinclair 2002), whose molecular evolution is associated with transitions in sex determination in reptiles (Janes et al. 2014). *Dmrt1* is sex-linked in fish (Nanda et al. 2002) and in birds, where it is the major sex-determining gene. Our RNAseq data revealed high male-biased *Dmrt1* expression during stage 22 gonads (late-TSP) in *Chrysemys*, counter to qPCR results using AKGs (Valenzuela, 2010). However, *Dmrt1* expression in *Apalone* shifted from low-temperature biased in stage 19 to high-temperature biased in stage 22 gonads, while it showed no differential expression in *A. mutica* using AKGs (Valenzuela, 2010). These results resemble Sox9 and suggest the drift of *Dmrt1* transcription during GSD evolution (Valenzuela 2010). They also suggest a critical role of *Dmrt1* in testicular development in *Chrysemys*, consistent with findings in *Trachemys scripta* (TSD) (Kettlewell et al. 2000). *Dmrt1* expression is male-biased in fish, birds and mammals (Valenzuela et al. 2013; Ayers et al. 2013)
**Genes in the vertebrate gonadal network but unknown in turtles:** RNAseq provided novel transcriptional profiles of several vertebrate genes unstudied in turtles (Table 2.4), including genes implicated in testicular differentiation in mammals \([Amh, AR, Cbln4, Dhh, Dmrt2, Fgf9, Fgfr2, Fhl2, Fog2, Pgds, Ptc1, Srd5a2 and Vnn1]\), genes involved in ovarian formation \([Ctnnb1, Esr2, Foxl2, Gata2, Rspo1 and Wnt4]\), and genes important for both testicular and ovarian function or general gonadogenesis prior to their sexual commitment \([Cbx2, Ck1, Gsk3b, Apc, Insr, Igf1r, Kdm3a, Six1, Six4, Dmrt3, Emx2, Esr1, Gata4, Lhx1, Lhx9]\) (Valenzuela 2008a; Eggers et al. 2014). A comparison of differential expression by temperature across select vertebrates is summarized in Table A1. Of these, \(Amh, Ar, Esr1, Fog2, Gata4\) and \(Lhx9\) show significant MPT bias at stage 22 in *Chrysemys* and are thus important candidate genes for turtle thermosensitive testicular differentiation in TSD vertebrates that deserve further functional research.

Finding differential expression prior to the onset of the canonical thermosensitive period in *Chrysemys* is of particularly importance as any such gene may be the key TSD master element that senses the environmental temperature signal or a key activator of the thermosensitive period. Notably, \(Ctnnb1\) (\(\beta\)-catenin, a member of the *Wnt* signaling pathway) showed female-bias at stages 9 and 12 in *Chrysemys* consistent with its involvement in early ovarian formation in mammals (Chassot et al. 2012; Liu et al. 2009). Similarly, \(Ctnnb1\) shows high-temperature bias in *Apalone* during stages 12 and 15. *Follistatin* (*Fst*) a gene activated by *Ctnnb1* in mice bipotential gonads (Eggers et al. 2014) showed slight female-bias (\(\alpha=0.05\)) in *Chrysemys* at stage 9, and slight high-temperature bias in *Apalone* at stage 15 (\(\alpha=0.05\)), suggesting that *Ctnnb1* could also activate *Fst* in
turtles. Further, this suggests that Ctnnb1 and Fst thermosensitive expression may be ancestral to cryptodiran turtles (the suborder to which Chrysemys and Apalone belong) and relic in Apalone, and underscores that downstream elements are key to rendering GSD sex immune to temperature as is the case of Wt1 and Dax1 via the loss of thermosensitivity of Sf1 (Valenzuela 2008b, 2008a). Indeed, Ctnnb1 is a repressor of Sf1 (Bernard et al. 2012), a gene with thermosensitive expression in Chrysemys but not Apalone (Valenzuela et al 2006). Also noteworthy, Chrysemys (and not Apalone) shows high male-bias of Insr and Igf1r expression during stage 9, which are indispensible for testiculogenesis in mice (Nef et al. 2003) and antagonize the Ctnnb1-Wnt signaling pathway essential for ovarian formation. This indicates that the same molecular antagonism exists in TSD turtles, is active before the canonical thermosensitive-period, and could influence growth trajectories via the insulin receptor family, inducing male determination (Mittwoch 2004) and other sexual dimorphisms with potential temperature-specific fitness consequences (Charnov and Bull 1977). These early differences between TSD and GSD systems may have functional significance for the evolution of sex determination.

**Genes in other functional categories:** We explored additional functional gene categories of plausible transducers of the temperature signal to gonadal developmental, some previously known as temperature-sensitive or linked to gonadal formation in other animals (Carmi et al. 1998; Kohno et al. 2010; Kuroki et al. 2013). These include vertebrate genes involved in gonadal and germ-line differentiation, androgen- and estrogen related genes, and genes linked to sex chromosomes, heat-shock and transient receptor potential genes and histone-related genes. Many of these genes exhibited thermosensitive
expression in both turtles (Tables A2-A11), including genes involved in histone modification, several kinases, genes involved in androgen- and estrogen signaling pathways, sex-linked genes and heat shock proteins. Overall, transcriptome composition was similar between species with some noticeable differences. Namely, *Apalone*'s transcriptome exhibited slightly lower representation of kinases, ubiquitin- and histone-related genes (Table 2.3), although we confirmed that they exist in *Apalone*'s genome using BLAST. Kinases are indispensable for cell functioning and orchestrate many cellular processes. One of these, the protein kinase *Map3k4*, directly affects *Sry* and *Sox9* expression in bipotential mice gonads, inducing testicular development (Bogani et al. 2009; Warr et al. 2012). Several heat shock proteins show sexually-dimorphic expression in American alligator (TSD), potentially influencing sex determination (Kohno et al. 2010). Sex-linked genes such as *Nf2* and *Prdx4* (Kawagoshi et al. 2009; Chocu et al. 2012) are differentially expressed by temperature in *Chrysemys* at stage 9. Similarly, we found thermosensitive expression for, *Serpinh1, Hsp90ab1* and *Hspa8* across stages in both turtles (Table A6) (perhaps relic in *Apalone*) while expression is monomorphic in mouse (Jameson et al. 2012), suggesting their potential turtle-specific role in gonadogenesis. Additional genes differentially expressed in turtles but not in the mouse gonad include *Ctnnb1* (early acting at 26°C in both *Chrysemys* and *Apalone*) and *Git2* [a sex-linked gene in *Pelodiscus sinensis* (Kawagoshi et al. 2009)] among others (Fig. 2.1d, Tables A15, A16).

We also focused our attention on the temperature-specific expression of genes involved in histone modification activities, including acetylation, methylation, phosphorylation and ubiquitination and multiple types of small RNAs that are involved in epigenetic gene silencing. We elaborate on a few of these genes below.
**Acetylation:** Histone acetylation occurs mostly on the N-terminal of histone tails, which are accessible for modification. Acetylation occurs on lysine residues that weakens the bond between the histone and DNA, and is usually associated with gene activation (Kouzarides 2007). *Ncoa6*, a nuclear receptor protein is a known recruiter of the *Cbp/p300* complex that directly leads to lysine acetylation on a histones H3, H4, H2A and H2B (Bannister and Kouzarides 2011). Our dataset shows MPT upregulation of *Ncoa6, Cbp* and *p300* during stage 9 of development in *Chrysemys*. *Ncoa6* is important for gonadal development, as its knockout leads to hypofertility in male and female mice (Mahajan and Samuels 2008). Among the interactors of *Ncoa6* is the *Retinoic acid receptor, Rara* which regulates ovarian formation (Mu et al. 2013) and *Estrogen receptor* genes ERα and ERβ. More candidate studies are needed to directly test the role of *Ncoa6* in mediating the temperature signal during gonadal development in TSD turtles.

**Methylation:** Histone methylation marks can lead to transcription activation or repression depending on the context. For instance methylation on residues H3K4, H3K36 and H3K79 are marks of active transcription, while H3K9 and H3K27 methylation are usually associated with transcription repression (Kouzarides 2007). *Dnmt3b* is a known *de novo* methylator in mammals i.e. methylation of unmethylated CpG dinucleotide sequences (Okano et al. 1999; Kato et al. 2007) that is upregulated at MPT during stage 9 in *Chrysemys*, and at high temperature during stages 15 and 19 in *Apalone We* also observed stage 9 upregulation of *Setd1a* (H3K4me3) and *Nsd1* (H3K36me), both marks of gene activation (Barski et al. 2007; Morris et al. 2007) during the MPT in *Chrysemys*. The early activity of *Dnmt3B, Setd1a and Nsd1* in *Chrysemys* strongly suggests regulation of gene expression by methylation could underlie gonadal development in TSD reptiles.
**Ubiquitination and Phosphorylation:** Lysine residues on histones subject to ubiquitination could be candidates for transcription activation or repression depending on the context. We found evidence of the FPT upregulation of the ubiquitin-conjugating enzyme *Ube2a* during stage 9 in *Chrysemys* that is involved in mono-ubiquitination of H2BK120, leading to transcriptional activation (Kim et al. 2005). *Ube2a* is also an important X-linked gene whose mutation causes mental retardation in humans (Nascimento et al. 2006). Phosphorylation of residues including serine, threonine and tyrosine on histone tails can trigger other modifications including H3K9 acetylation and H3K4 methylation, which are switches for transcription activation (Rossetto et al. 2012). We found evidence of early phosphorylation activity by *Nek6*, required for cell cycle progression (Yin et al. 2003) in *Chrysemys* (FPT upregulation during stage 9 and MPT upregulation during stage 12).

**ncRNAs:** Over the years, non-coding RNAs have been unearthed to play vital roles in transcriptional regulation, translation and DNA protection from foreign molecules among a host of other functions (Cech and Steitz 2014). We identified 252 transcripts in *Chrysemys* and 169 in *Apalone* annotated as ncRNAs. Among these, some ncRNA transcripts were also differentially expressed by temperature in both species (Figure A4). Interestingly, the difference in number of differentially expressed ncRNA transcripts by temperature was lowest at the onset of the thermosensitive period (stage 15) in *Chrysemys*. At all other stages, including stages 9 and 12, there were more ncRNAs expressed higher at MPT than FPT.

**Argonaute proteins and small RNAs:** Argonaute proteins, acting in conjunction with small RNAs are involved in post-transcriptional gene silencing (Meister 2013). We
identified 4 argonaute genes expressed in both turtle transcriptomes (Ago1-4). Ago1 and Ago4 were upregulated in Apalone at 31°C during stage 22, but not in Chrysemys. We identified 257 miRNA transcripts in Chrysemys and 170 in Apalone transcriptomes, found by BLASTing to known miRNA sequences (www.mirbase.org). Interestingly, in both Chrysemys and Apalone, we identified slightly more miRNAs as differentially expressed at 26°C than 31°C, with the difference more pronounced in Apalone. We identified 66 piRNA transcript sequences in Chrysemys and 61 in Apalone by BLASTing to sequences from piRNABank (Sai lakshmi and Agrawal 2008), with a handful of them differentially expressed by temperature. Interestingly, piRNAs have been reported as sex determinants in silkworms (Kiuchi et al. 2014), and have sex specific roles in fish (Zhou et al. 2010), but evidence of any direct association with sex determination in vertebrates is still unknown. Our results showing differential expression by temperature of many piRNAs warrants further investigation into their putative role in turtle sex determination.

**Discovery of new elements in the vertebrate gonadal network:** Our dataset also revealed genes expressed in turtle gonads but unreported in mice, thus possibly unique to reptilian gonadogenesis. Among these are Calr (female-biased at stage 9) and Dcn, a component of the extracellular matrix uncharacterized in the mouse gonad (Miqueloto and Zorn 2007), and which showed female-bias in Chrysemys stage 15 onwards. Ribosomal proteins are involved in gonadal development (Bhavsar et al. 2010) but their role is obscure. We found 49 ribosomal proteins (including 28s, 40s, and 60s) showing female-bias in stage 9 and male-bias in stage 12 in Chrysemys that warrant further investigation and may be involved in the differentially expression of temperature sensitive gene
networks. Interestingly, genes enriched in hypoxia tolerance and mitochondrial functions that mediate the adaptation to sub-zero temperatures were differentially expressed in Chrysemys across all stages (Costanzo et al. 2001), including translocases that function as chaperones across the mitochondrial membranes, the SLC25 family of mitochondrial transporters (Palmieri 2013), Ep300, Casp1 and Thioredoxin family involved in hypoxia signaling (Zhou et al. 2007; Zhang et al. 2003). Our data indicate that these genes, which underlie thermal adaptation, are also involved in early development.

**Thermosensitive response of signaling pathways:** Distinct cell types may derive from a handful of cell signaling pathways (Pires-daSilva and Sommer 2003). We found evidence that numerous signaling pathways are differentially regulated by temperature. For instance, Jak-Stat signaling, involved in cell proliferation and hematopoiesis (Hou et al. 2002) exhibits male-bias Egfr at stage 22 in Chrysemys. Nf-kB signaling plays a role in immune and stress response (Hayden et al. 2006), and involves members of the hypoxia-induced *Tumour necrosis family (Tnf)* (Chandel et al. 2000) and *Breakpoint cluster region (Bcr)* (Korus et al. 2002) which showed female-bias in Chrysemys at stage 22 and male-bias at stage 9, respectively. The receptor gene *Vegf*, which regulates sex-specific gonadal vasculogenesis (Bott et al. 2008) was also female-biased at stage 15 in Chrysemys and high-temperature biased during stages 19-22 in Apalone. Further, retinoic acid has been identified to induce meiosis in mice germ cells regulating ovarian formation (Mu et al. 2013). Two retinoic acid binding proteins were differentially transcribed in Chrysemys, *Crabp1* (male-biased during Chrysemys TSP) and *Crabp2* (female-biased pre-TSP and TSP) in Chrysemys and high-temperature biased at stage 15 onwards in Apalone). Among
signaling pathways implicated in vertebrate sex determination, *Foxl2* and members of the *Wnt* signaling pathway regulate ovarian formation (Ricken et al. 2002; Nef and Vassalli 2009). *Wnt* activates *Ctnnb1*, which inhibits *Srf* from activating *Sox9* and inducing testiculogenesis (Bernard et al. 2012). The canonical *Wnt* machinery including *Ck1, Apc*, and *Gsk3* show male-bias during stage 9 in *Chrysemys*, and monomorphic expression in *Apalona*, indicating that *Wnt* signaling is active in TSD and GSD turtles, but deployed differentially by temperature. Members of the *Mapk* signaling family, required for *Sry* activation testiculogenesis in mice (Warr et al. 2012) were also low-temperature biased in turtle bipotential gonads despite the absence of *Sry* (*Map3k3* at stage 9 in *Chrysemys*; *Map3k7* at stage 12 in *Apalona*) (Table A14), rendering them additional candidates for functional tests. *Akt* signaling is directly activated by *Fgf9* in mice, promoting steroidogenesis (Lai et al. 2014), but neither gene showed thermosensitive expression. Finally, *Ras*-mediated signaling is implicated in sex myoblast migration in nematodes (Sundaram et al. 1996), and a subtle thermosensitive expression was detected in *Chrysemys* (Table A14). Thus, *Jak-Stat*, *Nf-κB*, retinoic acid, *Wnt*, and *Mapk* signaling are potentially involved in TSD gonadogenesis, while this process appears independent of *Akt* and *Ras*-mediated signaling.

**Detection of temporally co-expressed gene clusters:** Genes of interest (described in Table 2.3) clustered during embryogenesis by their coexpression patterns in both turtles. *Chrysemys* differ more in the number of coexpressed modules (45 modules at 26°C; 10 at 31°C) than *Apalona* (16 modules at 26°C, 21 at 31°C) (Fig. 2.5), a pattern similar to core eukaryotic genes used as negative control. This suggests that temperature
differentially orchestrates gene co-expression in TSD versus GSD, such that *Chrysemys'* response is more compartmentalized, and *Apalone'*s is broader. Some vertebrate sex determination/differentiation genes were clustered, such as *Cbx2* and *Dmrt2* in *Chrysemys* at both temperatures, and *Ar/Lhx9* and *Insr/Srd5a1* in *Apalone*, whereas no associations among these genes are known in mammals or birds. Future functional assays on these candidates are warranted. Cluster composition differed by temperature and species, as clusters were enriched in different biological pathways, reflecting temperature effects on gene co-expression and the existence of modules in the urogenital network. *Chrysemys* male transcriptomes were enriched for pathways regulating transcription, cell proliferation, reproductive development and amino acid phosphorylation (a kinase activity that has been linked to *Sry* regulation in mice (Warr et al. 2012; Bogani et al. 2009)). Immune response functions like lymphocyte and leukocyte activation were female-bias concordant with humans (Fish 2008). Cell proliferation, which showed thermosensitive responses here (Table 2.4, A13), is linked to mammal sexual development as it is affected by *Sry* and MAPK signaling (Mittwoch 2013).

**Novel Transcripts:** A high percentage of novel transcripts in *Chrysemys* (53%) and *Apalone* (54%) are currently uncharacterized in SwissProt (which contains manually curated, non-redundant eukaryotic protein sequences) and the ncRNA databases, corroborating genes that were annotated as “predicted” in the *Chrysemys* genome (Shaffer et al. 2013). Many novel transcripts are male-biased at stages 9, 19 and 22 of *Chrysemys* (Fig. A1). In conjunction with the greater number of co-expressed clusters discovered at 26°C, this discovery of higher number of novel *Chrysemys* transcripts at 26°C is indeed
curious, and merits further investigation the putative roles of these transcripts. Do they produce functional proteins? Why does the production of male gonads at 26°C require a greater investment of transcripts? Could a knockout of any or a combination of these transcripts contribute to a sex reversal from male to female? Could the temperature signal at 31°C somehow inactivate these transcripts to guide ovarian genesis? If so, what could mediate such a temperature signal? High-quality protein sequence data and experimental assays are required to elucidate their function and help answer the above questions.

**Conclusion:** Ours is the first transcriptomic analysis of TSD vertebrates and GSD turtles. The strengths of the transcriptomic time series through embryogenesis permitted characterizing the full gonadogenesis network, its response to environmental temperature, and the discovery of previously unknown gonadal regulators in vertebrates. Our data underscore that differences between TSD and GSD in turtles are less likely due to unique elements in this network but instead, to the differential deployment of common elements and modules. Our work thus contributes to the evolutionary puzzle of vertebrate sex determination.
2.4 Materials and Methods

**Sample collection:** Total RNA was extracted using RNeasy Kits (Qiagen) (Valenzuela 2008a) from *Chrysemys* (TSD) and *Apalone* (GSD). Embryos were collected at stages 9, 12, 15, 19 and 22 from 26°C and 31°C which are male- (MPT) and female-producing (FPT) in *Chrysemys* and fall within the optimal thermal range for *Apalone* (Bull and Vogt 1979). Egg incubation followed (Valenzuela 2009; Valenzuela et al. 2013), using boxes containing moistened sand. Total RNA was extracted from trunks (stage 9), adrenal-kidney-gonadal (AKG) complex (stages 12, 15) and gonads alone (stages 19, 22). RNA-seq libraries were generated using pooled samples from ten embryos per temperature per stage for *Chrysemys* and five for *Apalone*. The libraries were sequenced using Illumina’s HiSeq 2000, and ~35 million 100-bp paired-end reads were obtained per library.

**Transcriptome assembly:** Reads were splice-mapped across exon boundaries to the *Chrysemys* genome version 3.0.1 (Shaffer et al. 2013) using GSNAP (version 2012-03-23), with the novel-splicing feature turned on (Wu and Nacu 2010). Independently, reads were quality filtered and adapter sequences were removed with Trimmomatic (Bolger et al. 2014) and assembled into species-specific *de novo* transcriptomes using the Trinity package (release 2013-02-25) (Grabherr et al. 2011) and their quality compared with the genome-guided assemblies. *de novo* transcriptomes were annotated using the Trinotate pipeline (Haas et al. 2013), mapping the longest open reading frame from each transcript/isoform to the SwissProt protein database (Boeckmann et al. 2003). *de novo* transcripts were mapped to 22,380 genes from the annotated *Chrysemys* genome (Shaffer et al. 2013) using GMAP (version 2012-03-23) (Wu and Watanabe 2005). Unannotated
transcripts were mapped to *Trachemys scripta* embryonic transcriptome (Kaplinsky et al. 2013) for comparison using GMAP (version 2012-03-23) (Wu and Watanabe 2005) with default settings. To quantify gene expression levels, the reads from each of the libraries were mapped back to these genes using GSNAP (Wu and Nacu 2010). Then, read-counts for each gene were calculated using HTSeq with the –s (strand-specificity) parameter set to no (Anders et al. 2014).

**Gene expression normalization:** We implemented a novel normalization procedure for read-counts using R version 2.15.2 (R Development Core Team 2012) and employing a mixed approach that combined normalization by the upper-quartile expression levels (Bullard et al. 2010) with normalization to the housekeeping genes *Transferrin receptor* (*Tfr*) and *hypoxanthine phosphoribosyl transferase 1* (*Hprt1*) (which were constitutively expressed across all stages in both species). This approach permitted validation of the transcriptomic expression levels by comparison to extensive expression data from *Chrysemys* obtained by qPCR of candidate genes from individual embryos and which were normalized to housekeeping genes (Valenzuela et al. 2013). To test the effect of the normalization procedure on the number of differentially expressed genes, we conducted Fisher exact tests between transcript expression levels that were normalized by (1) upper-quartile only (procedure 1 - UQ100), (2) upper-quartile after eliminating the top 1 percentile of transcripts with the highest expression (procedure 2 - UQ99), (3) upper-quartile and house-keeping gene normalization (procedure 3 - UQHK100), and (4) upper-quartile and house-keeping gene normalization after eliminating the top 1 percentile of transcripts with highest expression (procedure 4 - UKHK99).
**Differential expression tests:** Differential expression tests were performed per developmental stage between the MPT and FPT for *Chrysemys* (TSD), which correspond to high/low temperature for *Apalone* (GSD). The resulting p-values were corrected for false discovery (Benjamini and Hochberg 1995). Then, we concentrated on the highly differentially expressed genes after controlling the false discovery rate at a stringent cutoff of $1e^{-10}$. Differentially expressed genes were annotated using the KEGG database (Kanehisa 2002). Enrichment analyses of Gene Ontology (GO) categories against their respective species-specific transcriptomes were conducted using the DAVID Bioinformatics knowledgebase (Huang et al. 2007). Additionally, turtle transcriptomes were tested for enrichment against the mouse gonadal transcriptomes (Jameson et al. 2012) using DAVID (Huang et al. 2007) to test for transcriptional divergence between the turtle and mammalian lineages. In a complementary approach, we randomly subdivided each read library into 2 and 3 subsets (or “subsamples”) (Liu et al. 2014), identified the GSNAP alignments corresponding to these subsamples, and regenerated read counts per gene. We then used DESeq (Anders 2012) and EdgeR (Robinson et al. 2010) to independently determine the differentially expressed genes by leveraging the multiple subsamples while controlling false discoveries at 1%. Finally, the R package WGCNA (Langfelder and Horvath 2008) was used to identify modules of genes co-expressed across turtle embryonic stages in the original set of libraries as well as the subsamples.

**Author Contributions:** N.V. conceived of the project, helped with sample collection, guided data analyses and helped write the manuscript. S.R. performed all bioinformatics analyses, contributed to the interpretation of results and wrote the manuscript. A.S.
contributed the wrapper script for differential expression analyses and helped design the RNA-seq pipeline. R.L. collected samples and data, and contributed to the interpretation of results. All authors edited the manuscript. The authors declare no competing interest.
### 2.5 Tables and Figures

Table 2.1: Genome-guided and de novo transcriptome assembly results for *C. picta* (CPI) and *A. spinifera* (ASP)

<table>
<thead>
<tr>
<th>Species</th>
<th>Total Reads (Million)</th>
<th>Million Reads mapped (%)</th>
<th>Genomes-guided</th>
<th>De novo (Genome-independent)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Annotated Transcripts (Total)</td>
<td>Longest Transcript (bp)</td>
<td>Percent Mapped Reads</td>
<td>Human CEGs mapped out of 456 (%)</td>
<td>Mammalian sex genes mapped out of 27 (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CPI</strong></td>
<td>305.9</td>
<td>297.5 (97%)</td>
<td>72615 (279903)</td>
<td>18,723</td>
<td>92</td>
<td>377 (83%)</td>
<td>27 (100%)</td>
</tr>
<tr>
<td><strong>ASP</strong></td>
<td>373.4</td>
<td>163.6 (44%)</td>
<td>76761 (279753)</td>
<td>28,965</td>
<td>93</td>
<td>352 (77%)</td>
<td>26 (96%)</td>
</tr>
</tbody>
</table>
Table 2.2: Representation of 22,380 annotated genes in the *C. picta* genome in the *C. picta* (CPI) and *A. spinifera* (ASP) transcriptomes by treatment (26°C and 31°C) and developmental stage (9, 12, 15, 19, 22).

<table>
<thead>
<tr>
<th>Developmental Stage</th>
<th>Species</th>
<th>CPI</th>
<th>ASP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Genes per treatment</td>
<td>Genes per stage</td>
</tr>
<tr>
<td>Stage 9</td>
<td>26°C</td>
<td>13,362</td>
<td>13,687</td>
</tr>
<tr>
<td></td>
<td>31°C</td>
<td>13,474</td>
<td>13,714</td>
</tr>
<tr>
<td>Stage 12</td>
<td>26°C</td>
<td>13,308</td>
<td>13,714</td>
</tr>
<tr>
<td></td>
<td>31°C</td>
<td>13,557</td>
<td>13,714</td>
</tr>
<tr>
<td>Stage 15</td>
<td>26°C</td>
<td>13,215</td>
<td>13,590</td>
</tr>
<tr>
<td></td>
<td>31°C</td>
<td>13,313</td>
<td>13,590</td>
</tr>
<tr>
<td>Stage 19</td>
<td>26°C</td>
<td>13,598</td>
<td>13,742</td>
</tr>
<tr>
<td></td>
<td>31°C</td>
<td>13,328</td>
<td>13,742</td>
</tr>
<tr>
<td>Stage 22</td>
<td>26°C</td>
<td>13,182</td>
<td>13,629</td>
</tr>
<tr>
<td></td>
<td>31°C</td>
<td>13,398</td>
<td>13,629</td>
</tr>
<tr>
<td>TOTAL across stages</td>
<td></td>
<td>13,929</td>
<td>12,667</td>
</tr>
</tbody>
</table>
Table 2.3: Number of genes per category of interest present in the overall transcriptomes of *C. picta* and *A. spinifera*. Gray cells denote categories with genes absent in the *A. spinifera* transcriptome. Many of these genes are also differentially expressed by temperature in both species (see Tables sheets A2-A11).

<table>
<thead>
<tr>
<th>Gene category</th>
<th>CPI</th>
<th>ASP</th>
<th>Differences (Gene IDs): Genes absent in <em>A. spinifera</em> transcriptome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heat shock</td>
<td>27</td>
<td>27</td>
<td>None</td>
</tr>
<tr>
<td>Transient receptor potential</td>
<td>16</td>
<td>16</td>
<td>None</td>
</tr>
<tr>
<td>Germ cell-related</td>
<td>85</td>
<td>83</td>
<td>ANXA9, INCA1</td>
</tr>
<tr>
<td>Ubiquitin-related</td>
<td>200</td>
<td>192</td>
<td>E3-ubiquitin protein ligase TRIM41-like HERC6,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E3 ubiquitin/ISG15 ligase TRIM25-like, E3 ubiquitin-protein ligase TRIM39-like,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ubiquitin carboxyl-terminal hydrolase 8-like,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ubiquitin-60S ribosomal protein L40-like</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E3 ubiquitin-protein ligase TRIM39-like,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ubiquitin-like protein ISG15-like</td>
</tr>
<tr>
<td>Kinases</td>
<td>524</td>
<td>513</td>
<td>HIPK4, serine/threonine-protein kinase SBK2-like, myosin light chain kinase smooth muscle-like, pseudopodium-enriched atypical kinase 1-like, ANKK1, adenylate kinase 8-like, cyclin-dependent kinase 4 inhibitor B-like, ITK, proline-rich receptor-like protein kinase PERK9-like, c-Jun-amino-terminal kinase-interacting protein 3-like, putative uncharacterized serine/threonine-protein kinase SgK110-like</td>
</tr>
</tbody>
</table>
Table 2.3 continued

<table>
<thead>
<tr>
<th>Histone-related</th>
<th>43</th>
<th>37</th>
<th>histone H2A type 1-F-like, histone H2A.x-like, histone H2B 8-like-1, histone H2B 8-like-2, histone H2A.J-like, histone H2B 7-like</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human X (Y-linked within brackets)</td>
<td>146(1)</td>
<td>144(1)</td>
<td>FRMD7, SPRY3</td>
</tr>
<tr>
<td>Chicken Z (W-linked within brackets)</td>
<td>349(1)</td>
<td>345(1)</td>
<td>TMEM174, CER1, GZMA, SIGLEC15</td>
</tr>
<tr>
<td>Cell proliferation</td>
<td>10</td>
<td>10</td>
<td>None</td>
</tr>
<tr>
<td>Androgen/Estrogen related</td>
<td>19</td>
<td>19</td>
<td>None</td>
</tr>
<tr>
<td>Homologs of epigenetic genes</td>
<td>56</td>
<td>56</td>
<td>None</td>
</tr>
</tbody>
</table>
Table 2.4: Categories of GO pathways enriched (at p=0.05) in the largest coexpressed clusters in the *Chrysemys* and *Apalone* embryonic transcriptomes

<table>
<thead>
<tr>
<th>Cluster</th>
<th><em>Chrysemys picta</em></th>
<th><em>Apalone spinifera</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>26°C</td>
<td>31°C</td>
</tr>
<tr>
<td><strong>Cluster 1</strong></td>
<td>Cell proliferation (2)</td>
<td>Immune response (2)</td>
</tr>
<tr>
<td><strong>Cluster 2</strong></td>
<td>Post translational modification (1)</td>
<td>Protein translation (1)</td>
</tr>
<tr>
<td><strong>Cluster 3</strong></td>
<td>Post-translational modification (3)</td>
<td>Cell proliferation (2)</td>
</tr>
</tbody>
</table>
Figure 2.1: (a-c): Differentially expressed genes overlap (controlling FDR at $1 \times 10^{-10}$) across various categories: (a) between male- and female-producing temperatures in Stage 22 of C. picta based on expression levels normalized by (1) upper-quartile (UQ100); (2) upper-quartile excluding the top 1 percentile of transcripts with the highest expression (UQ99); (3) upper-quartile and house-keeping genes (UQHK100); and (4) upper-quartile and house-keeping genes excluding the top 1 percentile of transcripts with highest expression (UQHK99). (b) and (c): across stages in Chrysemys picta and Apalone spinifera. (d): Overlap of genes of interest present in turtles with the mouse gonadal genes described in Jameson et al., (2012). DE= differentially expressed, GoI= genes of interest described in Table 2.3.
Figure 2.2: Differential expression in turtles based on RNAseq of a subset of genes involved in the mammalian urogenital pathway (modified from Valenzuela 2008b, Liu et al. 2009, Chassot et al., 2012, Eggers et al., 2014, Lai et al., 2014) and other genes of interest for turtle gonadogenesis. Approximate equivalency is provided between mice and turtle developmental stages of gonadal development. [not sig. diff. exp. = Not significant differential expression].
Figure 2.3: Average qPCR expression at 26°C (blue) and 31°C (red) across developmental stages from qPCR experiments (top panels; modified from Valenzuela et al., 2013) and RNAseq (bottom panels; this study). Stars denote significant differential expression by temperature. Boxed stages fall within the thermosensitive period.
Figure 2.4: P-values of differentially expressed genes (after applying Benjamini-Hochberg correction) linked to mammalian urogenital pathways (Valenzuela, 2008b), showing higher expression at 26°C (blue) and 31°C (red). Highly differentially expressed genes (while controlling false discoveries at 1e-10) identified in dark blue and dark red. Light colored cells denote significance at a standard α=0.05. GA= gene absent
Figure 2.5: Panels a-d illustrate the Eigengene networks and dendrograms in (a) *Chrysemys*, 26°C (b) *Chrysemys*, 31°C, (c) *Apalone*, 26°C and (d) *Apalone*, 31°C for 981 genes of interest described in Table 2.3. Colors along the X and Y-axes represent clusters of genes showing similar expression. Gene order varies by plot along the X and Y-axes.
Figure A1: Novel transcripts in *Chrysemys picta* that are highly differentially expressed (while controlling false discoveries at 1e-10). Blue: upregulated at 26°C, red: upregulated at 31°C.
Figure A2: RNA-seq expression patterns for multiple genes previously implicated in mammalian urogenital development across developmental stages 9-22 at 26°C (blue) and 31°C (red) in Chrysemys picta and Apalone spinifera. Statistically significant differences are indicated with an asterisk (*). Thermosensitive period is indicated by a box.
Figure A3: RNA-seq expression patterns for multiple histone modifying across developmental stages 9-22 at 26°C (blue) and 31°C (red) in *Chrysemys picta* and *Apalone spinifera*. Statistically significant differences are indicated with an asterisk (*). Thermosensitive period is indicated by a box.
Figure A4: Non-coding and small RNAs in *Chrysemys* and *Apalone* that are highly differentially expressed (while controlling false discoveries at $1 \times 10^{-10}$). Blue: upregulated at 26°C, red: upregulated at 31°C.
Figure A5: Gene co-expression patterns by temperature for each turtle species from the RNA-seq data. Panels a-d illustrate modules of high (red) and low (yellow) co-expression for 981 genes of interest (described in Table 2.3) profiled across five developmental stages in *Chrysemys picta* [(a) and (b)] and *Apalone spinifera* [(c) and (d)] at 26°C and 31°C respectively.
List of supplementary tables:

Table A1: Genes involved in sex determination – a comparison across reptiles and mammals

Table A2: List of differentially expressed genes by category – sex-linked genes in vertebrates

Table A3: List of differentially expressed genes by category - Histone related

Table A4: List of differentially expressed genes by category - Ubiquitin related

Table A5: List of differentially expressed genes by category - Androgen/Estrogen related

Table A6: List of differentially expressed genes by category - Heat-shock related

Table A7: List of differentially expressed genes by category - Transient receptor potential related

Table A8: List of differentially expressed genes by category - germ cell related

Table A9: List of differentially expressed genes by category - cell proliferation related

Table A10: List of differentially expressed genes by category - kinases

Table A11: List of differentially expressed genes by category - RNA Binding Proteins

Table A12: List of GO pathways enriched in differentially expressed genes in the turtles

Table A13: Gene enrichment in the 3 biggest clusters at p=0.05 according to expression levels by species and temperature. The three biggest modules at each condition are indicated by turquoise, blue and brown colors in that order in Figure A5.

Table A14: List of genes differentially expressed by signaling pathway

Table A15: Comparisons of genes of interest (described in Table 2.3) between turtle and mouse gonad (Jameson et al., 2012): Column A: Differentially expressed genes of interest shared between both mouse and turtle; Column B: Differentially expressed genes of interest in turtle but not present in mouse gonad, Column C: Differentially expressed in mouse but not differentially expressed in turtle, Column D: Differentially expressed in turtle but not differentially expressed in mouse, Column E: Differentially expressed in Chrysemys but not in Apalone, Column F: Differentially expressed in Apalone but not in Chrysemys.

Table A16: Differential expression pattern of genes not characterized in the mouse gonad (Jameson et al., 2012)

Table A17: List of gene names (expanded) explored in this study
CHAPTER 3: TEMPERATURE INDUCES DIMORPHIC GENOME-WIDE DNA METHYLATION IN TURTLES WITH ENVIRONMENTAL SEX DETERMINATION

AUTHORS: Srihari Radhakrishnan¹,³, Robert Literman²,³ & Nicole Valenzuela³

AFFILIATIONS:
1: Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA, 50011
2: Ecology and Evolutionary Biology Program, Iowa State University, Ames, IA 50011
3: Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, IA 50011

SUMMARY: DNA methylation alters gene expression but not DNA sequence, and helps regulate phenotypic plasticity. Temperature-dependent sex determination (TSD) epitomizes phenotypic plasticity where environmental temperature drives embryonic sexual fate, as occurs commonly in turtles. Importantly, only two genes were previously identified in TSD fish and reptiles whose differential methylation induces their temperature-specific transcription, contributing to ovary and testis differentiation. Here we estimated genome-wide methylation levels in silico in the TSD turtle Chrysemys picta using a normalized CpG content, and validated its utility as a proxy by Methyl DNA Immunoprecipitation sequencing (MeDIP-seq) in male and female hatchling gonads. 57% of the genome was methylated (including 78% of all CpG dinucleotides). Repeats were highly methylated, some proportionally to their genomic abundance, whereas CR1-like repeats appear to escape silencing somewhat. Numerous genes regulating turtle gonadogenesis exhibited sex-specific methylation and were proximal to methylated
repeats. Our is the first demonstration that sexually dimorphic DNA methylation is pervasive in turtle gonads, supporting the hypothesis that DNA methylation could regulate thermosensitive transcription in TSD vertebrates more extensively than previously known. Further research of novel candidates identified here will help test this hypothesis and the prevalence of DNA methylation in governing the sexual outcome in TSD species.

**KEYWORDS:** DNA methylation, normalized CpG content, MeDIP-seq, Temperature-dependent and genotypic sex determination, Gonadal embryonic development, reptile vertebrate.
3.1 Introduction

Epigenetic modifications are heritable changes to the DNA that do not change the nucleotide sequence. Among them, DNA methylation is a biochemical process that adds methyl groups to cytosine or adenine nucleotides. Methylated DNA alters gene expression by preventing transcription factor binding (Watt and Molloy 1988) or by sometimes favoring the binding of repressors (Boyes and Bird 1991; Hendrich and Bird 1998). The regulatory role of methylation is widespread across eukaryotes (Barski et al. 2007; Benevolenskaya 2007). The addition of methyl groups occurs on CpG dinucleotides (cytosine linked to a guanine by a phosphate group) within genes in invertebrates (Suzuki and Bird 2008), and across genic and intergenic regions in vertebrates (Zemach et al. 2010). Importantly, changes in DNA methylation levels are linked to the regulation of phenotypic plasticity (Szyf et al. 2007; Kucharski et al. 2008). Temperature-dependent sex determination (TSD) represents a textbook example of phenotypic plasticity (a thermal polyphenism), where individuals with identical genotypes can develop alternative phenotypes (male or female) based on environmental cues (Valenzuela and Lance 2004; Tree of Sex Consortium 2014). Differential methylation of some genes in the sex-determining pathway has been experimentally observed in few TSD vertebrates (Kuroki et al. 2013; Navarro-Martin et al. 2011; Parrott et al. 2014; Matsumoto et al. 2013a), but the extent to which TSD plasticity is mediated by DNA methylation remains unknown.

DNA methylation patterns in animals have been estimated using in silico techniques by measuring the normalized CpG content (or nCpG), i.e., the ratio of the CpG dinucleotide abundance observed at particular genomic regions compared to that expected at random
based on the frequency of cytosines and guanines present in the genome [CpG observed/expected] (Elango et al. 2009). This value of nCpG is used as a proxy for DNA methylation since (a) DNA methylation is almost entirely targeted to CpG dinucleotides in animals (Jabbari and Bernardi 2004), and (b) 5-methylcytosine has the tendency to undergo spontaneous deamination which converts it to Thymine (Coulondre et al. 1978; Shen et al. 1994). Therefore, nCpG negatively correlates with the extent of DNA methylation such that in hypermethylated regions (where Cytosines within methylated CpGs have been converted to Thymine), the nCpG will be less than one. On the other hand, an nCpG ratio equal to 1 is indicative of no deviation from random expectation, while a value greater than one indicates hypomethylated regions. In hymenopteran insects, particularly multiple ant species, the distribution of nCpG values across the genome is unimodal and centered around a ratio equal to 1 (Simola et al. 2013), suggesting that DNA methylation does not play a strong role in gene regulation in those species. In contrast, the honeybee and pea aphid show a bimodal genic nCpG distribution with one peak centered around 0.5 and the other at 1 (Glastad et al. 2011), suggesting that some genomic regions are strongly affected by methylation while others are not. Among vertebrates, human promoters also show a bimodal pattern, such that genes with lower CpG content undergo higher methylation in somatic and germline cells (Weber et al. 2007; Yang et al. 2014). However, the pattern of CpG distribution in TSD vertebrates remains unknown.

Here, we use a two-pronged approach to test for the potential involvement of large-scale epigenetic modification in sexual development by TSD. First, we predict the genome-wide DNA methylation landscape in TSD turtles via \textit{in silico} analysis of the genome of the painted turtle (\textit{Chrysemys picta}) (Shaffer et al. 2013), a TSD reptile. Second, we validate the
*in silico* predictions from the normalized CpG content analysis by sequencing the methylome of hatching gonads from males and females obtained at temperatures that produce a single sex, and then test for differential methylation between the sexes. Thus, our study provides the first insight into the association between CpG content and differential methylation in any TSD vertebrate. Additionally, we focus on differentially methylated genes in the turtle urogenadal regulatory network, and identify important candidates for a putative role as mediators of sex determination via epigenetic modification in TSD vertebrates to guide further research. Finally, we analyze the repeat content of the methylome and explore the association between methylated repeat sequences and proximal methylated genes.

### 3.2 Results

**In silico prediction of genome-wide methylation:** Genome-wide normalized CpG content (nCpG) revealed consistent unimodal profiles across gene bodies (exons plus introns), exons alone, introns alone, promoters and intergenic sequences (Fig. 3.1). In almost all genomic regions, the nCpG was much lower than the expected ratio of 1, predicting that a significant fraction of the turtle genome is methylated. Notably, for the exons the distribution of nCpG was centered at 0.35 (indicated by the red line in Fig. 3.1a) (range = 0 to 1.39), whereas the distributions for the rest of the profiled regions were centered at 0.25 (Fig. 3.1b-f) and varied in range by region. The range of nCpG values for gene bodies was 0-1.7, an interval much smaller than that of either the upstream or intergenic regions [range = 0-5] or introns alone [range = 0-4].
**Methylome analysis:** Over 98% of the MeDIP-seq reads from the male and female hatchling gonads mapped to the *C. picta* genome (Shaffer et al. 2013). The methylome analysis uncovered ~2.95 million methylated 500bp-windows, totaling 1.48 gigabases in size, or ~57% of the genome, and overlapping with 17,646 genes. This corresponds to 78% of the CpG nucleotides in the genome. A total of 40% of the methylated windows fall within gene bodies which is significantly lesser than the 46% located within 50kb-upstream sequences that have potential regulatory functions (permutation test p-value=0.001), while the remaining 14% of methylated windows fall outside of gene bodies and sequences 50kb upstream of all genes.

**In silico predictions:** 94% of all methylated gene bodies detected experimentally via MeDIP-seq had an *in silico* nCpG value of 0.5 or lower, thereby showing a strong association between CpG depletion and actual methylation. Further, of all those genes that have an nCpG of 0.5 or lower (17,104 genes), 98.7% (16,884 genes) are methylated. Only 75 (out of 17646) methylated genes show an nCpG content of 0.8 or higher, with 36 (out of 90) methylated tRNA genes falling under this group. A Kolmogorov-Smirnov test revealed no significant difference between the number of genes predicted to be methylated using the nCpG index and the number identified as methylated by MeDIP, suggesting that *in silico* predictions are fairly accurate.

**Differentially methylated regions and their normalized CpG content:** The differential methylation analysis revealed 5647 differentially methylated windows between the sexes. Of these, 3076 windows were upmethylated in females (in 2414 genes)
and 2571 windows in males (in 2086 genes) (Fig. 3.2a). Interestingly, the nCpG content of the differentially methylated genes was significantly lower than for all methylated genes identified by MeDIP (p-value of resampling test = 0.001) (Fig. 3.2b).

Among these empirically-detected differentially methylated genes there were a number of kinases, androgen/estrogen related genes, histone- and ubiquitin related genes, heat shock and transient potential receptor genes; which by the nature of their biological functions are good candidates for a role to help transduce the temperature signal during development leading to alternate sexual fates in TSD species (Tables B1 to B8). The log-fold change in methylation between the sexes was enriched the most in introns, followed by promoter sequences and finally exons (Fig. 3.2c). Differential methylation of exons was around half that of promoters (there were 536 differentially methylated windows in promoter sequences versus 281 in exons). Some genes categories were somewhat enriched among the differentially methylated genes between the sexes, such as members of the Wnt signaling pathway and genes involved in transcriptional regulation which tended to be upmethylated in males relative to females, whereas genes involved in cell and neuron differentiation tended to be upmethylated in female hatchlings. However, at the level of GO terms the difference between sexes was not statistically significant after controlling for false discovery (Tables B9, B10). Among the differentially methylated genes, 541 genes showed distinct sex-specific methylation windows such that some windows were upmethylated in male hatchlings and other windows in the same gene were upmethylated in females (Table B11, Fig. 3.2d). Relative to all methylated genes, whose nCpG in the methylated window ranged from 0 to 12 (mean=0.306), the differentially methylated windows had a lower nCpG content ranging from 0 to 1.27 (mean=0.282) (Fig. 3.2e). This
pattern was observed in gene bodies comprising of exons and introns as well as promoters (Fig. 3.2f), suggesting that differential methylation of hatchling gonads was restricted to the regions of greater CpG depletion.

**Biology of differentially methylated genes:** We observed significant differential methylation between the sexes for genes involved in a number of functions, including phosphorylation (kinases), heat shock, transient receptor potential potential and histone/ubiquitin modification. We focused our attention first on reptilian homologs of genes involved in mammalian gonadogenesis (Valenzuela 2008a; Eggers et al. 2014). For instance, genes important for testicular formation, some of which are highly expressed at male producing temperature in TSD turtles, including Amh, Ar, Gata4, Lhx1, Lhx9 and Sf1 (Valenzuela et al. 2013; Yeh et al. 2002; Manuylov et al. 2011; Oréal et al. 2002) were significantly upmethylated in female hatchlings. In contrast, genes important in ovarian formation in mammals, such as Wnt4 and Emx2 (Pellegrini et al. 1997; Tripathi and Raman 2010) were upmethylated in males. While some of these genes exhibited upmethylation in the promoter regions near the 5’ end, others were upmethylated in their gene bodies, mostly in their intronic sequences. Yet others, such as Lhx1 and Gata4 show female upmethylation in both promoter (1 window each) and intronic sequences (2 and 3 windows, respectively). The Wilm’s tumor protein Wt1, a gene important for the formation of the bipotential gonad and later testicular development (Wilhelm and Englert 2002), and which is differentially expressed in C. picta turtles (Valenzuela 2008b), exhibited three upmethylated intronic windows in male hatchlings (Table 2.3).
A comparison of the methylomic signatures and differential transcription patterns in gonads of late stage embryos (stage 22) we obtained from male- (MPT) and female- (FPT) producing temperatures during another study (Radhakrishnan et al, unpublished), revealed a partial association between embryonic gene regulation and hatchling DNA methylation. Namely, 58 genes over-transcribed in male embryos were upmethylated in female hatchlings and 40 genes over-transcribed in female embryos were upmethylated in male hatchlings. However, in contrast, 714 genes in female embryos and 336 genes in male embryos were upregulated but showed no differential methylation in hatchlings (Table B12). Thus, evidence was detected consistent with a gene-specific influence of DNA methylation on sexually dimorphic transcription rather than a global effect (Fisher exact test, p = 1). Further, a significantly higher number of differentially methylated genes that were also differentially expressed at stage 22 are located in the vicinity of methylated repeats, than those that were not differentially expressed (p-value of permutation test=0.001).

Repeat elements: Because repetitive DNA sequences such as transposable elements can be subject to silencing by DNA methylation (Weisenberger et al. 2005) which could affect nearby genes, we analyzed the repeat content of the methylome. Repeatmasker analyses revealed that around 40% of the methylome consists of repeats, with significant representation (~45% of the methylome repeats) from the CR1 and HAT repeat categories. CR1s repeats were also the most abundant in the C. picta gonadal transcriptome (Fig. 3.3b). Pairwise Kruskal-Wallis tests suggested that the repeat abundances came from the same populations in all three pairwise comparisons between the genome, methylome and
transcriptome (p-value in each comparison: 0.453). Further, methylated repeats were located at significantly higher concentration (p-value of resampling test = 0.001) in 95% of all methylated genes, (the HAT category repeats being the most common), while being relatively scarce around non-methylated genes. Methylated repeats were also common (although significantly less so) in the vicinity of genes that were differentially methylated between males and females (~80% instead of 95%; p-value of permutation test = 0.001), with the DIRS repeat category being the most common. Interestingly, over 70% of those methylated repeats located nearby differentially methylated genes exhibited a sex-specific methylation bias concordant with the sex-specific methylation bias of the gene (Table 3.3).

Additionally, a significantly higher number of differentially methylated genes that were also differentially expressed at embryonic stage 22 are located in the vicinity of methylated repeats, than those that were not differentially expressed (p-value of permutation test=0.001). The distribution in category of methylated repeats did not differ between the differentially- and non-differentially expressed genes. Using regression, we evaluated the effect of methylation status on repeat silencing, by assessing the covariation between repeat transcription level and repeat abundance in both the genome and in the methylome. In both cases, the relationship was highly significant, and explained a significant proportion of variation in repeat transcription level (repeat abundance in genome: b = 0.0239, p = 0.0001, r² = 0.59; repeat abundance in methylome: b = 0.0569, p = 0.0007, r² = 0.49), although the variation in repeat expression was itself small (Fig. 3.3c, 3.3d). Further, multiple regression analyses with all variables did not significantly improve the explanatory power, implying that repeat transcription, repeat methylation status, and repeat genomic abundance are tightly linked.
3.3 Discussion

Genomic approaches are advancing our understanding of phenotypic plasticity at unprecedented rates, including the role that DNA methylation plays in mediating plastic responses to environmental inputs (Szyf et al. 2007; Kucharski et al. 2008; Bonasio 2014; Elango et al. 2009). Here we predicted *in silico* the methylation profile of the genome of *Chrysemys picta*, the first such analysis in any reptile and in any vertebrate with temperature-dependent sex determination (TSD) using the normalized CpG content (nCpG) as a proxy, and tested this prognosis experimentally using MeDIP-sequencing. Our results validate nCpG as a reasonable overall indicator of true methylation status, as evidenced by the observation that 98.7% of all genes with an nCpG of 0.5 or lower showed actual methylation in the gonads of *C. picta* hatchlings. We identified differentially methylated windows corresponding to over 4000 genes and found evidence consistent with the notion that DNA methylation might mediate the temperature-specific embryonic transcription of critical genes involved in gonadal formation, as detailed below.

*In silico analysis of genome-wide methylation:* Our data revealed that nCpG values follow a unimodal distribution at promoter regions of genes in painted turtles. This result contrasts with human promoters, which show a bimodal nCpG distribution with two distinct classes (Weber et al. 2007; Yang et al. 2014) that differ in methylation levels. Low CpG promoters in humans are hypermethylated in both somatic and germline cells, such that genomic methylation marks could be heritable (Bird 2002). Whether methylation marks in turtles are heritable remains to be determined. Human studies indicate that 65% of all promoters are high CpG promoters, out of which only 25% are hypermethylated
(Weber et al. 2007), while only 23% of are low CpG promoters, out of which 42% are hypermethylated. By comparison, the absence of a sizeable proportion of high CpG promoters in C. picta (Fig. 3.1d, e) suggests that DNA methylation is more extensive in turtle than in human promoters, further suggesting a potentially larger role for DNA methylation as a regulator of gene expression in reptiles than previously anticipated.

The distribution of nCpG values for exons in the C. picta genome is centered at 0.35, and at 0.25 for the rest of the genome, including introns, promoters and intergenic sequences. This remarkably high CpG depletion in C. picta is in stark contrast to in silico methylation profiles in ants (Simola et al. 2013) and honey bees (Glastad et al. 2011), where the values for exons are centered around an nCpG of 1. Patterns in C. picta by genomic region also contrasts with ants and honeybees where insect exons show a greater CpG depletion than the rest of the genome (Simola et al. 2013). Besides CpG content in promoters, CpG depletion in exons can also affect transcription as occurs in humans where it reduces gene expression (Bauer et al. 2010; Krinner et al. 2014). Thus, we hypothesize that the higher nCpG ratio in exons in C. picta relative to introns/intergenic/promoter regions (i.e. lower CpG depletion in turtle exons relative to overall depletion) may be the result of natural selection acting to preserve gene expression, particularly given that the turtle genome displays a generally greater CpG depletion compared to other animals. Additionally, we hypothesize that natural selection could also favor lower methylation in exons to prevent the accumulation of mutations via the methylation-deamination process where methylated CpG mutate to TpG, changing the coding sequence in ways that could produce non-functional proteins.
**Methylome Analysis:** The boundaries of all 500bp methylated windows in hatchling gonads encompassed 57% of *C. picta's* genome and include 78% of all CpG dinucleotides. This fraction is consistent with findings for most mammals, where 60% to 90% of all CpG dinucleotides undergo methylation (Glastad et al. 2011). Upstream sequences, which should have potential regulatory roles, were methylated at significantly higher levels (46%) than gene bodies (40%) in hatchling gonads. Of these upstream windows, 16% fall within the boundaries of an upstream gene (falling within an upstream intron 94% of the time), suggesting that perhaps these methylated genes possess alternative upstream promoters as previously suggested in humans (Kimura et al. 2006), where intragenic methylation of an upstream element may regulate alternative promoters for downstream genes (Maunakea et al. 2010).

**Validation of in silico predictions:** Out of all the methylated genes revealed by MeDIP, 94% had a nCpG content of 0.5 or lower, revealing an association between methylation and CpG depletion as seen in insects (Glastad et al. 2011) and humans (Weber et al. 2007). Further, 98.7% of all genes having an nCpG of 0.5 or lower were identified as methylated by MeDIP. Thus, our study underscores that CpG content is a reasonable indicator of methylation status. Of the few exceptions identified, most correspond to tRNAs. Namely, while 90 out of 182 annotated tRNAs in the *C. picta* genome were methylated, 40% (36/90) of methylated tRNAs have nCpG content of 0.8 or higher, suggesting that CpG depletion has been suppressed in many tRNAs from what would be expected based on CpG content alone, perhaps to prevent the methylation-deamination process (methylated CpG mutating to TpG) and thus, preserving the DNA sequence itself. This is consistent with the
idea that tRNA sequences have been largely conserved throughout evolution in all three domains of the tree of life (Widmann et al. 2010).

**Differentially methylated genes:** Out of all methylated genes (17,646), 541 genes (3%) contained multiple windows that exhibited sex-specific methylation, with some window(s) showing upmethylation in females while other window(s) of the same gene showed upmethylation in males (Fig. 3.2d). Because the males and females studied here were produced at contrasting temperatures that produce a single sex (males at 26°C and females at 31°C), the sex-specific methylation observed in hatchlings is also temperature-specific. This thermosensitive methylation of distinct windows within the same gene could potentially lead to alternative splicing of sex-specific transcripts, as shown in humans (Maunakea et al. 2013). The presence of male- and female-specific methylated windows (that in the case of C. picta are also temperature-specific) within single genes also suggests that the epigenetic machinery itself must be temperature-sensitive. However, further research is needed to test these hypotheses. Among these differentially methylated genes in painted turtles we detected important candidates known to regulate sexual development in other organisms. These include the *Epidermal growth factor receptor* (*Egfr*) which has been previously implicated in sexual dimorphism in *Drosophila* (Foronda et al. 2012), and *Mafb*, a gene with sexually dimorphic expression responsible for masculinization of male genitalia in mice (Suzuki et al. 2014).

We now discuss a number of gene categories that could potentially transmit environmental signals to the developmental pathways and thereby help determine the sexual fate in TSD taxa (Morrish and Sinclair 2002; Valenzuela and Lance 2004; Valenzuela
et al. 2013; Kohno et al. 2010) (Tables B1-B8). These included heat shock genes, transient receptor potential genes, a number of kinases, androgen/estrogen-related genes and histone-related genes. **Heat shock proteins** show differential expression by temperature in alligators and could play a role in TSD (Kohno et al. 2010). We found a number of heat shock genes, including *Hspa4, Hspa12a* and *Hspa12b*, all of which were upmethylated in male hatchlings. Additionally, the **Cold-Inducible RNA-Binding protein Cirbp** was proposed to regulate the expression of sex-determining genes and is differentially expressed during the early sex-determining period in the snapping turtle *Chelydra serpentina*, also a TSD species (Rhen and Schroeder 2010). Our data revealed upmethylation of Cirbp in female hatchlings, suggesting that Cirbp's thermosensitive transcription could be mediated by methylation. **Transient receptor potential genes** including *Trpm1, Trpm2, Trpm3, Trpm7* and *Trpm8* can respond to temperature stimuli (Dhaka et al. 2006) and were differentially methylated in *C. picta* hatchlings. Many **kinases**, including members of the *Mapk* signaling family, which are required for the activation of the sex-determining *Sry* gene in mice (Bogani et al. 2009; Warr et al. 2012) (which is absent in turtles as in all non-Therian vertebrates (Wallis et al. 2008)), are also differentially methylated, as are many **androgen and estrogen signaling** genes. Genes involved in **histone modification** directly regulate transcription and can act in a sex-specific manner. For instance, transcription of gonadal *aromatase* in slider turtle embryos (*T. scripta*) increases by demethylation (Matsumoto et al. 2013a), an activity that can be directed by local histone acetylation in mammals (Cervoni and Szyf 2001). Our results show upmethylation in male hatchlings for the **histone acetyltransferases Kat2a** and **Kat6a** genes, while **deacetylases** including *Hdac4, Hdac7* and *Hdac8* are upmethylated in female
hatchlings. Whether differential methylation of histone modifiers is linked to sexually dimorphic transcription in turtles remains to be tested. Because differential methylation by temperature could result in differential regulation, the genes listed above represent candidates with potential roles in governing sex determination by temperature.

We also investigated several genes in the mammalian urogenital regulatory network, of which several reptilian homologs display differential expression in TSD turtles, including \textit{Wt1} (Spotila et al. 1998; Valenzuela 2008b), \textit{Sf1} (Valenzuela et al. 2006; Ramsey et al. 2007), \textit{Dax1} (Torres Maldonado et al. 2003; Valenzuela 2008a), \textit{Sox9} (Torres Maldonado et al. 2003; Barske and Capel 2010; Valenzuela 2010; Matsumoto et al. 2013b), \textit{Aromatase} (Valenzuela and Shikano 2007; Matsumoto et al. 2013a), \textit{Dmrt1} (Torres Maldonado et al. 2003; Kettlewell et al. 2000; Valenzuela 2010), \textit{Estrogen receptor} (Bergeron et al. 1998; Chávez et al. 2009), \textit{Rspo1} (Matsumoto et al. 2013b) among others. Of these, only \textit{aromatase} and \textit{Sox9} methylation have been studied and demonstrated to influence gene transcription in developing TSD reptiles [the slider turtle (Matsumoto et al. 2013a), and the American alligator (Parrott et al. 2014)], and in TSD fish (Navarro-Martin et al. 2011). Here we provide evidence that many more genes in this regulatory network are differentially methylated between male and female gonads (Table 3.4). Our data indicate that the dimorphism in DNA methylation persists post-hatching, and that dimorphic methylation occurs in some genes that are differentially expressed during gonadal development. Namely, testicular formation in TSD turtles relies on the orderly, spatio-temporal expression of a number of genes in this network. Genes such as \textit{Amh}, \textit{Ar}, \textit{Gata4}, \textit{Lhx9} and \textit{Sf1}, are transcribed at higher levels at male-producing temperature (MPT) during stage 22 of embryonic development in \textit{C. picta} (Radhakrishnan et al, unpublished), a
stage late in the thermosensitive period for sex determination in painted turtles (Valenzuela et al. 2013). These five genes were upmethylated in female hatchlings, consistent with a repressive role of DNA methylation on their expression. The differentially methylated regions in these genes fall within the upstream promoter sequences (with the exception of Sf1, where differential methylation is intronic; see Table 3.4), underscoring the potential importance of DNA methylation of promoting regions in regulating transcription.

In C. picta embryos, the male-bias in Wt1 transcription at 26°C early in the thermosensitive period, decreases to monomorphic expression by stage 22 (Valenzuela et al. 2013). The presence of three upmethylated windows in Wt1 intronic sequence in male hatchlings is consistent with this reduction of Wt1 expression in males. Our results are consistent with the notion that the upregulation of some genes involved in testicular formation might be the result of the repressive effect that DNA methylation of the promoter region has at female-producing temperatures, as with Amh, Ar, Gata4, Lhx9 and Sf1. The converse is true for genes Wnt4 and Emx2, which are involved in ovarian formation (Tripathi and Raman 2010; Pellegrini et al. 1997), as they were upmethylated in the male hatchlings.

Importantly, our findings show that methylation marks are stable enough to persist post-hatching, perhaps even for long periods of time as has been described in mammals (Cedar and Bergman 2009; Lande-Diner and Cedar 2005). But notably, we did not observe differential methylation in hatchlings for aromatase, even in the 500bp window overlapping with its transcription start site as reported previously in the slider turtle embryos (Matsumoto et al. 2013a). Assuming that embryonic methylation patterns between slider and painted turtles are as similar as the patterns of aromatase transcription
are between these closely related turtles (Ramsey et al. 2007; Valenzuela et al. 2013), the lack of aromatase differential methylation in hatchlings detected here suggests that the differential methylation of aromatase during embryogenesis may be transient. Further methylation and transcription analyses in embryos are needed to tests this hypothesis directly.

**Normalized CpG content of differentially methylated genes:** Intriguingly, differential methylation in *C. picta* hatchlings was restricted to regions that showed significantly lower normalized CpG content relative to the genome-wide methylation levels, a pattern observed in both gene bodies and promoters (Fig. 3.2e, f). This indicates that CpG depletion is a precursor for differential methylation, but that not all regions with higher CpG depletion are differentially methylated. In phenotypically plastic insects, low CpG regions are enriched in basic biological processes such as metabolism and nucleotide processing (Elango et al. 2009), but our results indicate that turtle hypermethylated regions are enriched for cell morphogenesis and adhesion (Tables B9, B10), which are enriched in high CpG regions in insects. While these enrichment results are not statistically significant when controlling for false discoveries, our results suggest that turtles and insects differ fundamentally in their genomic CpG content distributions and its consequences, with turtles undergoing a higher level of genomic methylation.

**Repeat element analysis:** While 10% of the *C. picta* genome is composed of transposable elements (Shaffer et al. 2013), our analysis showed 40% of all methylated regions were composed of a combination of repeats, indicating that repeat elements in
general are the target of DNA methylation at a disproportionately higher rate than given simply by their abundance in the genome. In contrast, repeats comprise 41% of the rat genome and about 53% of the methylome (Sati et al. 2012). In turtles however, the relative proportion of particular repeat categories in the methylome was concordant with their relative abundance in the turtle genome, albeit the actual percentage of a particular repeat in the methylome was higher or lower than its abundance in the genome. Namely, on the one hand, CR1 elements were the most commonly represented repeat in the methylome followed by elements such as HATs, DIRS, Gypsy and Harbinger repeats, and this order of abundance reflects their order of abundance in the turtle genome (Shaffer et al. 2013). On the other hand, ~25% of the methylated repeats were CR1 elements, which is lower than their abundance in the painted turtle genome (~53%), whereas HATs, DIRS, Gypsy and Harbinger, were more abundant in the methylome than in the genome (Fig. 3.3a). However, the overall repeat abundances in the genome, methylome and transcriptome didn’t differ in their means. This indicates that the observed difference in abundance of specific repeat categories is not statistically significant to suggest that the painted turtle cell machinery targets repeat elements for silencing by DNA methylation as occurs in humans (Si et al. 2009). The transcriptomkal repeat abundances in stage 22 embryos is slightly better explained by their genomic abundance than by their methylomic abundance, as evidenced by our regression analyses (Fig. 3.3c, 3.3d). Further, over 95% of methylated genes were in the vicinity of a methylated repeat sequence, a pattern not observed in non-methylated genes. Additionally, we discovered an association between the methylation pattern of repeats and the sexually dimorphic methylation pattern of genes they border. Indeed, over 70% of repeat windows in the vicinity of differentially methylated genes were methylated
in the same direction as the genes (Table 3.3). This suggests that the presence of nearby methylated transposable elements could promote DNA methylation in nearby genes perhaps helping mediate gene expression patterns in C. picta as reported for Drosophila and human (Garrison et al. 2007; Cridland et al. 2015).

**Conclusion:** Ours is the first genome-wide assessment of DNA methylation in reptiles, and the first study of sexually dimorphic methylation levels in a TSD vertebrate. As such, this study sheds light on the epigenetic modifications that may play a role in mediating phenotypic plasticity in vertebrates. Our MeDIP-seq data provide empirical validation of *in silico* predictions obtained from nCpG content for the first time in any reptile, and show that nCpG content is a reasonable predictor of gene body methylation status. We find that painted turtles possess a unique pattern with nCpG values well below those observed in other animals and below those expected from the genome content of C and G nucleotides. In contrast, actual methylation levels given the genome CpG content agree with those in several other animals. Our data helped us identify several candidate genes whose methylation status putatively regulate transcription levels in a thermosensitive manner, perhaps playing a key role in driving the ultimate sexual fate in TSD reptiles.

### 3.4 Materials and Methods

**DNA isolation and sequencing:** DNA was extracted from the gonads of 3 month-old *C. picta* hatchlings (2 males and 2 females incubated at 26°C and 31°C respectively) using the Gentra Puregene DNA extraction kit (Gentra) following the manufacturer’s instructions.
Extracted DNA was fragmented and denatured, followed by immunoprecipitation of methylated DNA using the MeDIP protocol (Weber et al. 2005), and sequenced using the Illumina HiSeq paired-end protocol. We obtained between 126 million and 163 million 50bp reads per library amounting to a total of ~564 million reads (Table 3.1).

**Methylome construction and differential methylation analysis:** Sequencing reads were mapped to *C. picta* genome version 3.0.1 (Shaffer et al. 2013) using Bowtie2 version 2.2.5 (Langmead and Salzberg 2012). Unmapped reads were filtered out using Samtools (Li et al. 2009). The MEDIPS package (Lienhard et al. 2014) was used to (a) build an index for the *C. picta* genome, to ensure fast querying of the alignment files, (b) model read counts under a negative binomial distribution, (c) quantify mapped read counts per 500bp-windows then in RPKM (Reads Per Kilobase of Million mapped reads), (d) merge methylated windows and compute differential methylation by sex, while controlling for false discoveries (Benjamini and Hochberg 1995) at a level of 5%. Only windows with a total count of 50 or more uniquely mapped reads across all four libraries were considered for all subsequent analyses. Kolmogorov-Smirnov tests were used to determine if the nCpG content distribution of all annotated genes was comparable to those identified by the MEDIPS package. To test if the nCpG of the differentially methylated genes differed significantly from that of all the methylated genes a resampling test was performed (Crowley 1992) by iteratively drawing a random subset of genes (equal to the number of differentially methylated genes) from the entire set of methylated genes. Repeatmasker v3.3.0 (Smit et al. 1996) was used to identify repeats in the *C. picta* genome. Bedtools v2.17.0 (Quinlan and Hall 2010) was used to compute repeats overlapping with methylated
regions identified by MEDIPS. Permutation tests were used to test the significance of methylated repeats occurring in vicinity of methylated genes. In the absence of transcriptional data from hatchlings, we use an available transcriptomic dataset from late-developing embryos (stage 22) from another study (Radhakrishnan et al. unpublished) to test for an association between methylation patterns in hatchlings and transcription patterns using the Fisher’s exact test. Regression tests to model transcriptome abundance using genomic abundance of repeats and methylated repeats were performed in R (R Development Core Team 2012). The DAVID Bioinformatics knowledgebase (Huang et al. 2007) was used to assess the enrichment of functional categories to which the differentially methylated genes belong.

**Analysis of normalized CpG content:** The normalized CpG content \((nCpG)\) is calculated as:

\[
nCpG = \frac{(cg)}{\left(\frac{c}{l}\right) \times \left(\frac{g}{l}\right)}
\]

for a sequence of length \(l\), where \(c\) is the number of occurrences of Cytosine, \(g\) is the number of occurrences of Guanine, and \(cg\) is the number of times Cytosine is bordered by Guanine linked by a phosphate group (CpG) (Elango et al. 2009). In theory, nCpG values can range from 0 to +infinity for an infinitely long sequence, with a value of 1 when the number of CpG dinucleotides observed is equal to the expected based on the sequence length and abundance C and G. Values <1 denote CpG depletion from what is expected by chance and values >1 represent overabundance of CpGs from random expectation. In DNA methylation,
of CpG dinucleotides initiates the deamination of the cytosine, transforming it to thymine, thus lowering the nCpG to less than 1. Studies conducted in vertebrate and invertebrate animals reveal an upper limit between 2 and 2.5 for various genomic regions (Elango and Yi 2008; Elango et al. 2009; Glastad et al. 2011; Park et al. 2011; Simola et al. 2013; Yang et al. 2014)

Bedtools (Quinlan and Hall 2010) was used to parse the exon, intron, promoter and intergenic coordinates from the *Chrysemys picta* genome (Shaffer et al. 2013), given the annotations in gff3 format. In-house perl scripts were used to compute the CpG contents by genomic region.

**Author contributions:** N.V. conceived of the project and helped edit the manuscript. S.R. performed all bioinformatics analysis and wrote the paper. R.A.L extracted DNA from turtle hatchlings. All authors edited the manuscript.
## 3.5 Tables and Figures

Table 3.1: Illumina library statistics for *Chrysemys picta* hatchlings

<table>
<thead>
<tr>
<th>Sex (Incubation temp)</th>
<th>Library Size</th>
<th>% Mapped Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male (26°C)</td>
<td>137,159,464</td>
<td>98.85</td>
</tr>
<tr>
<td>Male (26°C)</td>
<td>138,462,674</td>
<td>98.80</td>
</tr>
<tr>
<td>Female (31°C)</td>
<td>126,682,102</td>
<td>98.83</td>
</tr>
<tr>
<td>Female (31°C)</td>
<td>163,323,313</td>
<td>98.87</td>
</tr>
</tbody>
</table>
Table 3.2: Summary of studies exploring the diversity of nCpG distributions in vertebrates and invertebrates

<table>
<thead>
<tr>
<th>Group</th>
<th>Species</th>
<th>Region</th>
<th>nCpG profiled</th>
<th>nCpG distribution</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammals</td>
<td><em>Homo sapiens</em></td>
<td>Promoters</td>
<td>Bimodal</td>
<td></td>
<td>1,2,7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Introns</td>
<td>Unimodal</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td><em>Pan troglodytes</em></td>
<td>Promoters</td>
<td>Bimodal</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Mammals</td>
<td><em>Gorilla gorilla</em></td>
<td>Promoters</td>
<td>Bimodal</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Mammals</td>
<td><em>Pongo abelii</em></td>
<td>Promoters</td>
<td>Bimodal</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Mammals</td>
<td><em>Macaca mulatta</em></td>
<td>Promoters</td>
<td>Bimodal</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Mammals</td>
<td><em>Monodelphis domestica</em></td>
<td>Promoters</td>
<td>Bimodal</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Mammals</td>
<td><em>Mus musculus</em></td>
<td>Promoters</td>
<td>Bimodal</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Mammals</td>
<td><em>Ornithorhynchus anatinus</em></td>
<td>Promoters</td>
<td>Bimodal</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Birds</td>
<td><em>Gallus gallus</em></td>
<td>Promoters</td>
<td>Bimodal</td>
<td></td>
<td>2,7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Introns</td>
<td>Unimodal</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Amphibians</td>
<td><em>Xenopus tropicalis</em></td>
<td>Promoters</td>
<td>Bimodal</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Introns</td>
<td>Unimodal</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Fish</td>
<td><em>Danio rerio</em></td>
<td>Promoters</td>
<td>Bimodal</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Introns</td>
<td>Unimodal</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Tunicates</td>
<td><em>Ciona intestinalis</em></td>
<td>Promoters</td>
<td>Bimodal</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Introns</td>
<td>Unimodal</td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>
Table 3.2 continued

<table>
<thead>
<tr>
<th>Insects</th>
<th>Gene body</th>
<th>CDS</th>
<th>Exons</th>
<th>Introns</th>
<th>Promoters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anopheles gambiae</td>
<td>Unimodal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apis mellifera</td>
<td>Bimodal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene body</td>
<td>Bimodal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tribolium castaneum</td>
<td>Unimodal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acrthosiphon pisum</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasonia vitripennis</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solenopsis invicta</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bombyx mori</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pediculus humanus</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daphnia pulex</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reticulitermes flavipes</td>
<td>ESTs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solenopsis invicta</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tricholium castaneum</td>
<td>Gene body</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drosophila</td>
<td>Gene body</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acyrthosiphon pisum</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasonia vitripennis</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solenopsis invicta</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bombyx mori</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pediculus humanus</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daphnia pulex</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reticulitermes flavipes</td>
<td>ESTs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solenopsis invicta</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bombyx mori</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pediculus humanus</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daphnia pulex</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reticulitermes flavipes</td>
<td>ESTs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solenopsis invicta</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bombyx mori</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pediculus humanus</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daphnia pulex</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reticulitermes flavipes</td>
<td>ESTs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solenopsis invicta</td>
<td>CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3.2 continued

<table>
<thead>
<tr>
<th>Genus</th>
<th>Exons</th>
<th>Introns</th>
<th>Promoters</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Atta cephalotes,</em></td>
<td>Unimodal</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><em>Acromyrmex echinatior</em></td>
<td>Unimodal</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><em>Pogonomyrmex barbatus</em></td>
<td>Unimodal</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><em>Camponotus floridanus</em></td>
<td>Unimodal</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><em>Harpegnathos saltator</em></td>
<td>Unimodal</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><em>Linepithema humile</em></td>
<td>Unimodal</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

Sources: 1 = Weber et al., 2007; 2 = Elango et al., 2008; 3 = Elango et al., 2009;
4 = Glastad et al., 2011; 5 = Glastad et al., 2012; 6 = Simola et al., 2013;
7 = Yang et al., 2014
Table 3.3: Presence of methylated repeats in the genic context in *Chrysemys picta* hatchlings

<table>
<thead>
<tr>
<th>Number (#) and percentage (%) of methylated repeats</th>
<th>Number of gene bodies with methylated repeats within x kb of start codon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 kb</td>
</tr>
<tr>
<td># Among all 17,646 methylated genes (%)</td>
<td>16,791</td>
</tr>
<tr>
<td></td>
<td>(95.1%)</td>
</tr>
<tr>
<td># Among 2086 upmethylated genes (male)</td>
<td>1650</td>
</tr>
<tr>
<td></td>
<td>(79%)</td>
</tr>
<tr>
<td># Among 2414 upmethylated genes (female)</td>
<td>1949</td>
</tr>
<tr>
<td></td>
<td>(80.7%)</td>
</tr>
<tr>
<td># Among 840 methylated genes of interest (Tables B1-B8) (%)</td>
<td>822</td>
</tr>
<tr>
<td></td>
<td>(97.8%)</td>
</tr>
<tr>
<td># Among all 433 non-methylated genes (%)</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>(6.7%)</td>
</tr>
<tr>
<td># Methylated repeat windows that are male-upmethylated near male-upmethylated genes</td>
<td>681/946 (72%)</td>
</tr>
<tr>
<td># Methylated repeat windows that are female-upmethylated near female-upmethylated genes</td>
<td>801/1032 (77.6%)</td>
</tr>
</tbody>
</table>
Table 3.4: Summary of differentially methylated (FDR cutoff: 0.05) genes in hatchlings putatively involved in reptilian gonadogenesis, with genes upregulated at the male producing temperature (26°C) (shaded cells) and female producing temperature (31°C) (unshaded cells) during stage 22 of embryonic development (Radhakrishnan et al., unpublished).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sex of hatchlings showing Upmethylation</th>
<th>Upmethylated Region</th>
<th>Sex of stage 22 embryos showing over-transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Amh</em></td>
<td>Female</td>
<td>P, I</td>
<td>Male</td>
</tr>
<tr>
<td><em>Ar</em></td>
<td>Female</td>
<td>P, D</td>
<td>Male</td>
</tr>
<tr>
<td><em>Gata4</em></td>
<td>Female</td>
<td>P, I (3)</td>
<td>Male</td>
</tr>
<tr>
<td><em>Lhx1</em></td>
<td>Female</td>
<td>P, I (2)</td>
<td>Male</td>
</tr>
<tr>
<td><em>Lhx9</em></td>
<td>Female</td>
<td>P</td>
<td>Male</td>
</tr>
<tr>
<td><em>Sf1</em></td>
<td>Female</td>
<td>I</td>
<td>Male</td>
</tr>
<tr>
<td><em>Emx2</em></td>
<td>Male</td>
<td>E</td>
<td>Female</td>
</tr>
<tr>
<td><em>Insr</em></td>
<td>Male</td>
<td>I</td>
<td>Female</td>
</tr>
<tr>
<td><em>Wnt4</em></td>
<td>Male</td>
<td>I</td>
<td>Female</td>
</tr>
<tr>
<td><em>Wt1</em></td>
<td>Male</td>
<td>I (3)</td>
<td>Female</td>
</tr>
</tbody>
</table>

Legend: P: Promoter, I: Intron, E: Exon, D: Downstream of last exon; () indicates # of methylated windows if > 1
Figure 3.1: Chrysemys picta genome: Normalized CpG (nCpG) content profiles measured in silico in (a) Exons only (CDS) (b) Introns only (c) Exons & introns (d) 10,000 bases upstream of exon 1 (e) 50,000 bases upstream of exon 1 and (f) intergenic sequences. Red lines indicate the peak value of nCpG in the respective regions.
Figure 3.2: (a) RPKM heatmap of differentially methylated genes (rows) clustered by mean methylation level per gene. Methylation levels were scaled to [-1.5, 1.5] to indicate genes undergoing high (green) and low (red) relative methylation. (b) Normalized CpG content of all annotated Chrysemys picta genes (red), experimentally verified to be methylated using MeDIP-seq (yellow) and differentially methylated (purple). (c) Fold change (Red: upmethylated in female; green: upmethylated in male) as seen in Chrysemys picta gene bodies (exons + introns), exons only and promoters. Gene bodies (particularly introns) show the highest level of fold change between sexes, followed by promoters and exons. (d) Examples of windows within the same gene undergoing sex-specific methylation in Chrysemys picta. (e) Scatterplot of normalized CpG content (nCpG) in methylated windows occurring in (e) gene bodies relative to nCpG of gene bodies and (f) promoters relative to nCpG of the complete promoter sequence (~5kb upstream). Upmethylated windows in male (blue) and female (red) hatchlings of Chrysemys picta are overlaid.
Figure 3.3: (a) Percentage of repeat categories in total repeats found identified in the *Chrysemys picta* genome (Shaffer et al, 2013) and methylome (this study) (b) Percentage of repeat categories overall identified in the *Chrysemys picta* genome (Shaffer et al, 2013), hatchling gonadal methylome (this study) and embryonic gonadal transcriptome (Radhakrishnan et al, unpublished). Plots of transcriptomal repeat abundance regressed onto repeat abundance in the genome (c) \(p=0.0001\) and methylome (d) \(p=0.0007\). Transcriptomal abundance of repeats correlates better with the genomic abundance \(R^2 = 0.59\) than with the methylomic abundance \(R^2 = 0.49\).
List of supplementary tables:

Tables B1 through B8 - Differentially methylated genes in *Chrysemys picta* hatchling methylomes along with RPKM, fold change and edgeR p-value:

Table B1: Heat shock genes
Table B2: Androgen/Estrogen related genes
Table B3: Kinases
Table B4: Histone-related genes
Table B5: Ubiquitin related genes
Table B6: Transient receptor potential genes
Table B7: Genes involved in cell proliferation
Table B8: Germline-related genes
Table B9: Categories enriched (p=0.05) in upmethylated genes (26°C)
Table B10: Categories enriched (p=0.05) in upmethylated genes (31°C)
Table B11: Genes with distinct upmethylated sex-specific windows within the same gene.
Table B12: List of genes differentially expressed in the *Chrysemys picta* transcriptome (stage 22) and differentially methylated in the methylome (stage 26 - hatchlings)
CHAPTER 4: RELATIVE RATES OF SEX-LINKED AND AUTOSOMAL CODING SEQUENCE EVOLUTION MEASURED IN VERTEBRATES

AUTHORS: Srihari Radhakrishnan¹,², Jeffrey W. Tamplin³, Arthur Georges⁴ & Nicole Valenzuela²

AFFILIATIONS:
1: Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011
2: Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, IA 50011
3: Department of Biology, University of Northern Iowa, Cedar Falls, IA 50614
4: Institute for Applied Ecology, University of Canberra, Canberra, Australian Capital Territory 2601, Australia

SUMMARY: Sex chromosomes carry master genes that decide the sexual fate of individuals and genes that contribute to sexual dimorphisms, among others. Sex chromosomes are predicted to evolve faster than autosomes because natural selection acts distinctly on them due to their reduced recombination and smaller population size (Fast-X and Fast-Z hypotheses), potentially fixing beneficial mutations quickly. A few sequence divergence studies of sex-linked genes compared to different autosomal genes support the Fast-X and Fast-Z hypotheses in some taxa. Here, we leverage the diversity of sex-determining mechanisms found in reptiles to test whether the rate of evolution of sex-linked genes in reptiles and select vertebrates with XX/XY or ZZ/ZW chromosomes differs from that of the same genes located in autosomes in closely related and distant taxa (six turtles, American alligator, anole lizard, chicken, human and mouse) We report for the first
time, the faster evolution of some Z-linked genes in turtles (the Chinese soft-shell turtle *Pelodiscus sinensis*) relative to autosomal orthologs in other taxa, a rate not matched by its close relative, the spiny soft-shell turtle (*Apalone spinifera*), and only surpassed by X-linked genes in mammals. On the contrary, we found slower evolution of X-linked genes in the musk turtle *Staurotypus triporcatus* (XX/XY), a pattern also observed in the Z-linked chicken genes whose ZW chromosomes are homologous to *Staurotypus* XY. We detected faster evolution of sequences in TSD lineages as a group relative to GSD lineages. Genome level analyses are warranted to test the generality and prevalence of the patterns found here, some of which reveal that important evolutionary changes at the gene level, mediated by chromosomal context, may be at odds with the rate of genomic evolution overall. This study represents, to our knowledge, the first analysis of evolutionary rates of sex-linked sequences relative to autosomal counterparts across amniotes, and illuminates sex chromosome evolution by providing an empirical test of their divergence expected by theoretical evolutionary models among closely and distantly related vertebrates.

**KEYWORDS:** Sex chromosome, Fast-X effect, Fast-Z effect, molecular evolution
4.1 Introduction

Sex chromosomes have evolved independently from autosomes in multiple branches of the tree of life (Charlesworth 1991; Bachtrog et al. 2014). Sex chromosomes can be homomorphic as observed in some snakes and ratite birds (Ogawa et al. 1998; Ellegren 2000) or heteromorphic as commonly observed in many other animals (Tree of Sex Consortium 2014). An important feature of sex chromosome evolution is the reduction of homologous recombination in the heterogametic sex (XY males and ZW females), driven by multiple factors (Bachtrog et al. 2014) and which is accentuated the more heteromorphic the sex chromosomes become over evolutionary time. Consequently, the heterogametic chromosome (Y in XY males and W in ZW females) may degenerate owing to an irreversible accumulation of deleterious mutations, a phenomenon that could be abetted by other factors including background selection (removal of neutrally evolving regions in linkage to deleterious sites) (Charlesworth 2012) and genetic hitch-hiking (where fixation of novel alleles could also fix linked deleterious alleles) (Singh and Petrov 2007). Reduced recombination in sex chromosomes is also driven by selection favoring the accumulation of antagonistic genes that increase fitness in one sex and are harmful to the other (Van Doorn 2009).

This differentiation of sex chromosomes from autosomes was proposed over a century ago (Muller 1914) has generated many important debates. Some of the recalcitrant questions delve into whether or not there exists a differential mutation rate and a bias to accumulate genes bestowing sex-specific fitness between sex chromosomes and autosomes (Vicoso and Charlesworth 2006), the extent of global dosage compensation prevalence in sex chromosomes as postulated by Ohno (Ohno 1967), and the extent to which sex
chromosomes affect the early stages of speciation (Payseur 2014). Indeed, certain sex-linked genes in mice have been disproportionately associated with hybrid sterility in interspecific crosses (White et al. 2012). This observed hybrid sterility is associated with increasing inter-species divergence in sex-linked sequences (Wang 2003), a phenomenon that is spurred on by the Fast-X and Fast-Z effects hypotheses (Charlesworth et al. 1987; Mank et al. 2007), which posit the faster evolution of sex chromosomes relative to autosomes due to various factors. Any new partially or fully recessive sex-linked mutations in the hemizygous sex chromosome are directly exposed to selection, whereas they are masked by the ancestral allele in autosomes if they are recessive (Mank et al. 2007). Consequently, beneficial mutations are expected to accumulate and deleterious ones to be removed faster on sex chromosomes relative to autosomes, since recombination among sex chromosomes is reduced in the heterogametic sex (Jablonka and Lamb 1990). Further, the population size of the sex chromosomes is lower than that of autosomes (¾ for X and Z, and ¼ for Y and W chromosomes compared to autosomes) (Bachtrog et al. 2011). At very low population sizes, genetic drift may be relatively stronger than natural selection in fixing or wiping alleles away.

The Fast-X and Fast-Z hypotheses have been tested in some invertebrates and vertebrates. Initial studies in *Drosophila* found no evidence of Fast-X (Thornton et al. 2006), but later work combining intraspecific polymorphism and species divergence data detected rapid evolution of X-linked sequences within the *Drosophila* lineage (Mackay et al. 2012; Campos et al. 2014). Fast-X evolution has also been observed in vertebrates including mice (Kousathanas et al. 2014), humans and chimpanzees (Lu and Wu 2005) driven by both positive (Nielsen et al. 2005) and purifying selection (Hvilsom et al. 2012).
Fast-Z, which is the mirror hypothesis to the Fast-X in species with female heterogamety, appears to occur in moths, driven by positive selection (Sackton et al. 2014) and in birds (Mank et al. 2007). While invaluable, these studies examined closely related species exclusively such that phylogenetically broader analyses of sex-linked sequences are overdue, ideally including a group where multiple sex-determining mechanisms coexist as in reptiles. Specifically, turtles afford a unique opportunity to address the evolution of sex-linked sequences, because they possess both male- and female- heterogamety, as well as temperature-dependent sex determination without sex chromosomes (Valenzuela and Lance 2004, Valenzuela et al. 2014), yet the relative rate of sex chromosome and autosome sequence evolution in turtles remains virtually unexplored.

As a first step to fill this gap here we test the relative rate of evolution in eleven amniotes including six turtles (with XX/XY, ZZ/ZW and temperature-dependent sex determination or TSD), anole (XX/XY), the alligator (TSD), chicken (ZZ/ZW) and two mammals (XX/XY), by measuring the rate of coding sequence evolution of multiple genes that are present in sex chromosomes in some species, but are autosomal in others. We observe for the first time faster evolution driven by positive selection of sex-linked sequences in the Chinese softshell turtle *Pelodiscus sinensis*, while the musk turtle *Staurotypus triporcatus* exhibits slower evolution of sex-linked sequences, a pattern also seen in the chicken. We also describe patterns of evolution of sequences between focal taxa (e.g. reptiles versus others), and of TSD species relative to GSD species.
4.2 Methods

Data Collection, and multiple sequence alignment of coding sequences: Complete coding sequences were extracted from the Ensembl genome browser (www.ensembl.org) corresponding to 20 genes (in sets of 4 genes that are sex-linked in some species while being autosomal in others) and 4 control genes described in Table 4.1 for human (Homo sapiens - HSA), mouse (Mus musculus - MMU), chicken (Gallus gallus - GGA) and anole lizard (Anolis carolinensis - ACA). Chicken-specific transcripts were mapped against available genomes of the Chinese softshell turtle Pelodiscus sinensis (PSI) (Wang et al. 2013b), the painted turtle Chrysemys picta (CPI) (Shaffer et al. 2013) and the American alligator Alligator mississippiensis (AMI) (St John et al. 2012), and against newly sequenced genomes we obtained of the musk turtle Staurotypus triporcatus (STR), the spiny softshell turtle Apalone spinifera (ASP), the wood turtle Glyptemys inscripta (GIN), and the Murray river turtle Emydura macquarii (EMA) using the discontiguous megablast feature in Geneious version R7.1.7, using default parameters (Kearse et al. 2012). Hereafter species will be referred to by their genus names or three-letter acronym. For each top BLAST hit, the coding sequence from the corresponding gene was annotated in the newly sequenced turtle genomes (STR, ASP, GIN and EMA) after ensuring that the alignments were in frame, and extracted. Stop codons were removed from all coding sequences. To test if sex-linked sequences evolved at a different rate from autosomal sequences, coding sequences from all sex-linked genes of a given species were then concatenated, and a multiple-sequence translational alignment was performed using Geneious version R7.1.7 (Kearse et al. 2012). The corresponding concatenated orthologous coding sequences in other taxa were
autosomal, except for sequences in PSI-Z which is homologous to ASP-Z (Badenhorst et al. 2013), and in STR-X which is homologous to GGA-Z (Kawagoshi et al. 2014). To test if reptilian and sauropsid (reptilian + bird) sequences evolved at a different rate than mammals, and TSD species a different rate than GSD species, the alignments were prepared such that they were composed of concatenated sequences from all 20 genes in the vertebrate dataset. All multiple sequence alignments were visually inspected prior to further analysis to ensure they had in frame codons and no premature stop codons. Conserved blocks from the multiple sequence analysis were identified for further analysis using Gblocks (Castresana 2000). Autosomal genes were selected from all the species to serve as controls as explained below, their coding sequences were extracted, concatenated and multiple-aligned as described above. The size of the raw concatenated multiple coding sequence alignments obtained ranged from 3.6kb (for genes that are sex-linked in Anolis and autosomal in others) to 11kb (for genes that are sex linked-genes in Apalone/Pelodiscus and autosomal in others), and 3.98kb for the control genes. After filtering the alignments for conserved blocks (Gblocks) and non-gapped regions (PAML), the final alignments ranged from 2.4kb (ACA) to 8.5kb (Apalone/Pelodiscus) for the sex-linked genes, and 3.61kb for the control autosomal genes.

Molecular evolution analysis: The CODEML package in PAML version 4.7 (Yang 2007) was used for all molecular evolution analyses. To test if the rate of coding sequence evolution, measured as dN/dS (ω), was significantly different in a given sex chromosome “foreground” relative to the autosomal “background”, we compared the likelihood of two models using a chi-square test (Pearson 1900) to identify the better-fit model. The first was
the branch model (Yang and Nielsen 1998) with the following parameters: Model=2, NSsites=0; where \( \omega \) is estimated from the alignment and varies across branches. The alternative was the one-ratio model with the following parameters: Model=0, NSsites=0; where \( \omega \) is invariant across branches. Under positive selection \( \omega \) is expected to be greater than 1 (higher non-synonymous rate of mutations than synonymous), while under purifying selection the opposite holds true, and thus \( \omega \) is less than 1. A \( \omega \) of 1 implies that there is no difference between the rate of non-synonymous and synonymous mutations, and thus it is a signature of neutral evolution. 100 bootstraps were used to generate confidence intervals for \( \omega \). The relative rate of evolution of sex-linked sequences relative to autosomal sequences was computed as \( \omega_{SL}: \omega_{A} \). We also tested for evidence of any difference in the rate of evolution of sauropsid (\( \omega_{S} \)) and reptilian (\( \omega_{R} \)) sequences relative to mammals (\( \omega_{M} \)), and of species with temperature-dependent sex determination relative to those with genotypic sex determination (\( \omega_{TSD} : \omega_{GSD} \)), using a concatenated alignment of all the 20 genes in our dataset that represent both sex-linked and autosomal contexts. For the autosomal genes serving as negative control, the species with the focal sex-linked sequences under analysis was chosen as foreground while the remaining species constituted the background (\( \omega_{FHK} : \omega_{OHK} \)). To test whether specific sites were undergoing positive selection, the branch-site model (Yang et al. 2005) (Model=2, NSsites=2), was used to test the alternative hypotheses that (H1) \( \omega \) is variable versus (H0) that \( \omega \) is invariant, using the following PAML parameters: (H1) omega=1 (initial guess); fix_omega=0 (estimates optimal \( \omega \) for the branches; (H0) omega=1 (initial guess); fix_omega=1 (\( \omega \) not allowed to vary). The BEB score (P>0.9) was used to identify sites under positive selection.
(Yang et al. 2005). These analyses were repeated separately for the control genes. The following phylogenetic tree was used in all molecular evolution-based analysis:

\[(\text{MMU, HSA), (ACA, ((GGA, AMI), (EMA, ((PSI, ASP), (STR, (GIN, CPI)))))})]\] (Valenzuela et al. 2013) (Fig. 4.1).

**Protein sequence and structure analysis:** Domain identification on the NF2 and TSPAN7 protein sequences was performed using the BindN (Wang and Brown 2006) and the PROSITE web servers (Sigrist et al. 2013). *De novo* 3D structures for protein sequences which exhibited particularly interesting results in the above analyses (for NF2 in ASP and PSI; and TSPAN7 in CPI and GIN) were predicted using the i-Tasser web server (Zhang 2008). Pymol (Schrödinger, LLC 2010) was used for structural alignment of protein structures.

### 4.3 Results

**Molecular evolution:** Overall, the branch-site model determined the proportion of sites undergoing purifying selection to be 93%, consistent with the dataset comprising of only coding sequences. The relative rate of evolution (or the \(\omega_{\text{SL}}:\omega_{\text{A}}\) ratio) measured using bootstrapped branch model for human/mouse was greater than 1 (\(\omega_{\text{SL}}:\omega_{\text{A}} = 5.74\)), indicating that these X-linked sequences in mammals undergo faster evolution than the same sequences in an autosomal context in other vertebrates, while the opposite was true in chicken (\(\omega_{\text{SL}}:\omega_{\text{A}} = 0.61\)) (Fig. 4.2). This ratio was equal to 1 in the anole (*Anolis*), indicating no difference between the mean \(\omega\) in the sex-linked genes relative to their
autosomal counterparts. Human and mouse sex-linked sequences evolve at the same rate \( (\chi^2 \text{ test statistic: } 0.04; \ p\text{-value: } 0.84). \) Among turtles, the branch model predicted a significantly different rate of evolution for the sex-linked genes relative to the autosomal orthologs for *Pelodiscus* \( (\chi^2 \text{ test statistic: } 101.66; \ p\text{-value: } 0) \) and *Staurotypus* \( (\chi^2 \text{ test statistic: } 13.88; \ p\text{-value: } 0; \) see Table C1 for the complete list of branch-model comparisons), indicating that sex-linked genes in these two turtles were under selection but in opposite directions, whereas sex-linked genes in *Apalone* showed no difference in \( \omega \) relative to the autosomal background. Namely, the relative rate of evolution was \( \omega_{SL}: \omega_A = 3.76 \) for *Pelodiscus* (indicating that sex-linked genes were evolving under positive selection), and \( \omega_{SL}: \omega_A = 0.54 \) for *Staurotypus* (indicating that sex-linked genes were evolving under purifying selection), and \( \omega_{SL}: \omega_A = 1 \) for *Apalone* (indicating neutral evolution) (Table 4.2). However, when each gene was tested separately, the sex-linked genes in *Staurotypus* show no slower evolution than their autosomal orthologs. Because the XY of *Staurotypus* and ZW of chicken arose from the same ancestral pair, we tested them in combination as focal species with the remaining taxa in the background, and did not observe a significantly different rate of evolution of their sex-linked sequences \( [\chi^2 \text{ test statistic comparing the null branch model (no difference in foreground and background } \omega)] \) against the alternative model (significant difference in foreground and background \( \omega \)): \( 0.018, p=0.89 \). Similarly, we found no difference in the rate of evolution of sex-linked genes when we combined the two Trionychid turtles in our dataset (*Apalone* and *Pelodicus*) as focal species relative to the autosomal orthologs in the other taxa \( (\chi^2 \text{ test statistic: } 0.46; \ p\text{-value: } 0.49). \) The control genes showed no difference in \( \omega \) between foreground (focal species) and background (other taxa) values in all species except for the chicken, where
they seem to evolve under purifying selection (Gallus) (Table 4.2). Further, the branch-site model identified multiple amino acid residues undergoing positive selection in sex-linked genes of Staurotypus and Pelodiscus respectively, relative to autosomal sequences in the remaining species, most of them concentrated in the Nf2 gene in PSI (Fig. C1).

TSD vertebrates showed a high rate of molecular evolution relative to GSD taxa. Namely, results using the branch model on the concatenated alignment of all 20 genes across 11 species revealed that the painted turtle Chrysemys (a turtle with temperature-dependent sex determination or TSD) exhibits a higher rate of evolution relative to all other vertebrates with genotypic sex determination ($\chi^2$ test statistic: 86.26; p-value: 0). The same result held when the alligator (also TSD) was included in the background ($\chi^2$ test statistic: 100.04; p-value: 0). This pattern of higher rate of evolution was also seen in the alligator, albeit less strongly (with Chrysemys in the background: $\chi^2$ test statistic = 4.47; p-value = 0.03, without Chrysemys in the background: $\chi^2$ test statistic = 7.81; p-value = 0.005) (Fig. 4.3). In contrast, the control genes in the alligator and Chrysemys showed no difference in their rate of evolution. The branch-site model predicted 49 positively selected residues (BEB score > 0.9) in the Tspan7 gene in Chrysemys.

When comparing rates of evolution among vertebrate groups for the genes in our dataset that are sex-linked in any of our study species, we observed faster evolution in turtle sequences relative to all other taxa ($\chi^2$ test statistic: 6.25; p-value: 0.01). However, reptilian genes evolved at the same rate than in mammals ($\chi^2$ test statistic: -0.002; p-value=1), and genes in archosaurs (birds + crocodilians) evolved at the same rate as in turtles ($\chi^2$ test statistic: 0.016; p-value: 0.89). In contrast, control genes our dataset (which
were autosomal in all taxa) did not differ in their evolutionary rate in the focal species relative to the background in any of these group level comparisons (Fig. 4.4).

**Protein structure prediction:** We examined the predicted 3D structure of the NF2 and TSPAN7 proteins more closely to test if these sites correspond to potentially important functional regions of the proteins. Our *de novo* predictions of protein structures for NF2 in the Trionychid softshell turtles *Pelodiscus* and *Apalone* using i-Tasser revealed almost identical structures for the amino acid sequences between species. However, a structural alignment revealed differences in root-mean-square deviation (rmsd) in the alpha-helices (Fig. 4.5a) towards the C-terminal of the protein. Most of the positively selected residues belonged to the FERM domain of the protein, involved in localizing proteins to the plasma membrane (Pearson et al. 2000) (Fig. C1). We observed a much greater difference in predicted structures of the TSPAN7 protein between *Chrysemys* and its closest relative in our dataset, *Glyptemys* (Fig. 4.5b), although only 2 out of 39 residues undergoing positive selection that differ between these two turtles are located in the transmembrane family domain (Fig. C2).

### 4.4 Discussion

Sex chromosomes carry the master sex-determining genes responsible for committing the bipotential gonad to the testicular or ovarian differentiation fate either by the presence or absence of sex-linked genes [e.g. Y-linked *Sry* in mammals (Koopman 1999)], or by dosage of sex-linked genes [e.g. Z-linked *Dmrt1* in birds (Smith et al. 2009)] that trigger the sex-specific developmental cascades. Genes in linkage disequilibrium with the sex-determining region that are hostile to one sex and beneficial to the other
accumulate in sex chromosomes via sexually antagonistic selection (Van Doorn 2009) which favors a reduction in recombination around the sex-determining region to maintain favorable gene combinations from breaking apart, and may eventually lead to the degeneration of the heterogametic sex chromosome (Y or W). Thus, selection favors beneficial alleles on sex chromosomes to evolve at a faster rate than in autosomes leading to the patterns known as Fast-X and Fast-Z evolution (Charlesworth et al. 1987; Mank et al. 2007). This context-dependent evolution of sex-linked gene sequences has been tested in vertebrates and invertebrates with reasonable support (Lu and Wu 2005; Mank et al. 2007; Sackton et al. 2014; Ávila et al. 2014; Hu et al. 2013; Betancourt et al. 2002) but the evolution of sex-linked genes in the context of these hypotheses has never been studied in turtles or any other reptile. Here we contribute to filling this gap by testing for the first time whether or not the molecular evolution of a number of coding sequences linked to sex chromosomes in focal reptiles, birds and mammals differs from the evolutionary rates of the same sequences in autosomal orthologs in other amniotes, using a set of selected turtles, lizards, alligator, chicken, mouse and human for which information is available.

Overall, we found that most nucleotide sites in all the alignments (comprising genes in sex chromosomes and autosomes) in our dataset appear to be under purifying selection, which is expected since these are coding sequences and as such, should be under selective pressures to avoid mutations that could disrupt the production of functional proteins. However, we also detected neutral evolution and positive selection in some residues, genes, and species, as well as differences in the rates of molecular evolution.

For instance, we detected evidence of an accelerated rate of evolution for the sex-linked genes in mammals (*Homo/Mus*) relative to their autosomal orthologs in other
species. Bootstrapped alignments revealed a surprisingly wide range for the 95% confidence intervals of dN/dS of sex-linked sequences in mammals relative to their autosomal orthologs in the other species (0.23 to 0.69, mean = 0.46), indicating that in general, mammals accumulate disproportionately more changes at the aminoacid level in their sex-linked sequences (and therefore, could likely result in altered protein functionality) compared to birds and reptiles (no differences were detected between human and mouse). This pattern is concordant with previous observations (Janes et al. 2010; Shedlock et al. 2007). In contrast, Anolis sex-linked genes exhibited a dN/dS equal to one, suggesting that sex chromosomes in this lizard may not be under strong positive nor purifying selection but are evolving neutrally. Interestingly, chicken (Gallus) exhibited slower evolution of sex-linked genes compared to their autosomal orthologs. While faster evolution of Z-linked sequences in birds have been reported (Mank et al. 2007; Wright et al. 2015), those studies compared sequences among species that were sex-linked against different sequences that were autosomal in the same species, such that results could not disentangle gene-specific effects from chromosome-specific effects. In contrast, our approach compares the evolution of the same genes in different contexts (i.e., sex-linked versus autosomal) across various amniote groups, allowing us to decouple the “gene effect” from the chromosomal-context effect, and to identify lineage effects as well. We hypothesize that the purifying selection we detect here for chicken sex-linked genes may be explained by the fact that chicken Z carries the master sex-determining gene which acts via a dosage effect (Chue and Smith 2011) such that selection may oppose changes that could disrupt avian gonadal formation, while the mammalian X carries just a handful of genes directly involved in gonadal formation (Karkanaki et al. 2007).
Turtles exhibited contrasting patterns of molecular evolution even among closely related species with homologous and identical sex chromosome systems, revealing that species-specific effects can have a major impact in sex chromosome evolution. Namely, two of the four sex-linked genes examined in Pelodiscus (Nf2 and Sf3a1; Fig. C1) showed evidence of positive selection compared to their autosomal counterparts in other amniotes, whereas they did not differ in the relative rate of evolution in Apalone, another Trionychid turtle (results were robust to including or excluding the other in the background). Apalone and Pelodiscus share a remarkably conserved morphology of their homologous ZZ/ZW sex chromosomes despite having diverged from each other for ~95 million years (Badenhorst et al. 2013). Thus, our results revealed extensive sequence divergence in the sex chromosomes of these two turtles in at least some genes, in the absence of any noticeable morphological divergence at the cytogenetic level. We detected 24 amino acid residues in Nf2 under positive selection in Pelodiscus, many of them occurring in the FERM domain which is involved in localizing proteins to the plasma membrane, a characteristic of proteins with cytoskeletal roles (Pearson et al. 2000). Interestingly, Nf2 is also part of the hippo signaling pathway (Cockburn et al. 2013) which crosstalks with other genes including Beta-catenin (Ctnnb1) and members of the Wnt signaling pathway (Imajo et al. 2012) that help orchestrate vertebrate gonadogenesis (Liu et al. 2009; Kim et al. 2006). However, a direct role for Nf2 in gonadogenesis is yet to be determined. In contrast, only 3 residues in Sf3a1 were under positive selection and introduce a change in both polarity and hydrophobicity relative to their orthologs in Apalone. Sf3a1 has roles in transcription, including spliceosome assembly and pre-mRNA splicing, and has been implicated in colorectal cancer in humans (Sharma et al. 2014; Chen et al. 2015). The potential functional
importance of such small number of residues should not be discarded since gene function can be altered by changes in just a handful of amino acids with important fitness consequences (Nei 2005; Nachman et al. 2003), although these three *Sf3a1* residues are not within any known domains. It is unclear whether positive selection acts directly on *Nf2* or on other neighboring genes on *Pelodiscus Z* that affect *Nf2* by linkage. In other words, one or more beneficial mutations may have arisen directly in *Nf2/Sf3a1* or instead, in neighboring genes causing a selective sweep through the region containing *Nf2/Sf3a1*. Alternatively, mutations may have accumulated in *Pelodiscus Nf2* (and *Sf3a1*) due to relaxed purifying selection, which would permit the production of a different protein isoform without suffering a fitness cost. To test these alternatives we examined other genes in the region surrounding *Nf2* and *Sf3a1*, including *Zmat5*, *Cabp7*, *Nipsnap1* and *Ccdc157*, which are located within a ~212Kb stretch in *Pelodiscus-Z* (Kawagoshi et al. 2009) and found that they also show significant differences in coding sequence between *Pelodiscus* and *Apalone* whereas the regions flanking this block do not (Table 4.3). This observation supports the notion that a selective sweep may have affected this region, albeit the original target of selection within it remains unknown. Our de novo 3D structural alignments of the NF2 protein predicted for these species revealed that most differences are localized in the C-terminal portion of the protein, potentially altering the ligand binding and/or protein function (Echols et al. 2003). Interestingly, some of the genes in this positively-selected region, *Cabp7* and *Nipsnap1*, are linked to calcium ion binding and inhibition of the transient receptor potential (*Trp*) activity (Schoeber et al. 2008). These are promising candidates for putative roles in gonadal development, because the calcium ion-mobilized transcription factor machinery is required for proper male gonad formation.
in mammals (Hanover et al. 2009). Furthermore, Trp channels can have sex specific roles (San Wong et al. 2015) and are temperature-sensitive in eukaryotes (Shen et al. 2011), thus plausibly playing a role in mediating temperature-dependent sex determination. However, direct links of these genes to sex determination or differentiation in turtles are yet to be determined. While it would be tempting to attribute the positive selection in this region of *Pelodiscus* to the evolution of GSD in this lineage, this would leave unexplained the lack of selection in the same gene block in *Apalone* which shares a homologous ZZ/ZW system with *Pelodiscus*. The faster evolution of the *Pelodiscus*-Z genes (Kawagoshi et al. 2009 and this study) could also be explained by lack of recombination between the Z and W, as may result from a deletion or inversion on the W, and which would prevent mutations from being repaired. The deletion hypothesis is supported by the previously reported failure of *Nf2* and *Sf3a1* probes to hybridize on *Pelodiscus*-W while they do so on *Pelodiscus*-Z (Kawagoshi et al. 2009). The random nature of such mutation accumulation process could explain the divergence of this block in *Pelodiscus*-Z away from *Apalone*-Z.

Results in *Staurotypus* turtles also showed a pattern of molecular evolution that differs from that expected by chance, but in opposite direction to the pattern observed in *Pelodiscus*. Namely, coding sequences in *Staurotypus* X evolved slower than their autosomal counterparts, suggesting that purifying selection was acting on sex-linked genes. Alternatively, the lower molecular evolution of *Staurotypus* X could also be due to the absence of recent novel mutations, since theory predicts that selection acting on standing genetic variation rather than on novel mutations can cause sex chromosomes to evolve slower than autosomes (Orr and Betancourt 2001). The lack of novel mutations may be due to the relatively young age of *Staurotypus* XX/XY system (~50 million years) which are
younger than the ~180My old ZZ/ZY system in Pelodiscus and Apalone (Badenhorst et al. 2013). Another potential explanation for the conservation of Staurotypus X is that, the genes examined were located in the pseudoautosomal regions (PAR) where recombination with the Y is still prevalent. A PAR chromosomal location would decrease the strength of genetic drift as the effective population size of this region would not differ from that of autosomes, thus subjecting sequences in such contexts to a similar recombination rate as autosomes (Otto et al. 2011). Further, the current lack of identification of X-specific genes in Staurotypus precludes the analyses of sex-linked genes that do not undergo homologous recombination, which was possible in Pelodiscus Z. The slower evolution of the sex-linked genes observed in both chicken and Staurotypus is particularly interesting, because both arose from the same ancestral pair of autosomes (Kawagoshi et al. 2014). Whether this slower evolution of the set of 4 genes each in chicken and Staurotypus is traceable back to the ancestral autosomal pair, or is a separate independent occurrence after their differentiation into sex chromosomes in each lineage remains to be seen. It is unknown if Staurotypus relies on dosage of sex-specific genes on the X chromosome to orchestrate sex determination as seen in the chicken (Ayers et al. 2013).

The control genes in our dataset were chosen because they are autosomal in all target species, and thus no difference in rate of evolution is expected between any given foreground and background species. This expectation held true in almost all species in our dataset except for the chicken control genes, which mimicked the slow evolution pattern seen in chicken sex-linked genes (Table 4.2).

Noteworthy, there was a significant difference in the rate of molecular evolution in the TSD species (the turtle Chrysemys and the American alligator) relative to GSD species,
with the *Chrysemys* and alligator orthologs of amniote sex-linked genes evolving faster relative to their GSD counterparts (results were robust to including or excluding the other TSD species in the background). Specifically, the gene *Tspan7* showed 49 (22%) positively selected residues in *Chrysemys* out of a total 218 in the alignment. This strongly suggests that *Tspan7* could be adapted for a potentially different function in *Chrysemys* that remains to be determined. Intriguingly, while the protein encoded by *Tspan7* is transmembranous and is involved in cell development, motility, proliferation among a other functions (Huang et al. 2010), mutations in *Tspan7* have been linked to a number of diseases, including x-linked mental retardation in mammals. Thus, we hypothesize that perhaps the canonical role of *Tspan7* in *Chrysemys* was taken over by any of its 34 known paralogs (Mitchell et al. 2014), allowing the neofunctionalization of *Tspan7*. This hypothesis is supported by the observation that members of the tetraspanin family have undergone massive gene loss in 80% of all vertebrate lineages which could facilitate positive selection by triggering adaptive evolution in the remaining paralogs (Huang et al. 2010). Further support for this hypothesis derives from our predicted 3D structures of the TSPAN7 protein in *Chrysemys* and *Glyptemys* (the other turtle in our dataset from the Emydidae family), which differed greatly from one another (Fig. 4.5b). Curiously, *Tspan7* is also implicated in misregulation of transcription in cancer-related pathways, as its expression is induced by a chimera of two genes, *Ews-Wt1*. *Wt1* is a well-known gene whose role in gonadal formation has been described extensively in many vertebrates (Pelletier et al., 1991; Spotila et al., 1998; Western et al., 2000; Wilhelm and Englert, 2002) and is a candidate TSD master gene in *Chrysemys* (Valenzuela 2008b). However, similar to *Nf2* in *Pelodiscus*, the direct involvement of *Tspan7* in sex determination in *Chrysemys* is unknown. Interestingly,
amniote sex-linked gene sequences exhibited faster evolution in turtles relative to all other taxa, similar to the previously observed pattern of genes involved in immune function and musculoskeletal patterning, despite the documented slower molecular evolution of turtle genomes in general relative to other amniotes (Shaffer et al. 2013). This observation underscores how global genomic generalizations may obscure important evolutionary phenomena at the gene level, and highlight the significance of the chromosomal context in mediating these patterns.

In the broader context, understanding the roles and nature of evolution of the sex-linked genes under different selective pressures would shed light on a number of unanswered questions. For instance, in comparison to autosomal or pseudoautosomal (PAR) orthologs where recombination is prevalent, is there stronger selective pressure driven by gene dosage, to preserve coding sequences of genes that are purely hemizygous? How does selection on regulatory regions compare to that on coding sequences, which produce functional proteins that could potentially be pleiotropic? Could regulatory sequences be under greater adaptive evolution than coding sequences, as observed in rodents (Halligan et al. 2013)? Importantly, hybrid sterility in inter-specific mice crosses maps disproportionately to sex-linked genes, implicating them as players in speciation events (Payseur 2014; White et al. 2012). Studies like ours that thus provide new avenues in the form of reptilian evidences of differential rate of evolution of sex-linked- and TSD-specific sequences to study reptilian speciation and other consequences of independent sex chromosome incipience in reptiles. This includes testing the efficacy of selection on sex chromosomes relative to autosomes, the amassing of genes with sex-biased fitness effects
on sex chromosomes and the relationship between coding sequence integrity and gene dosage, where little is known in the realm of reptiles (Vicoso et al. 2013).

**Conclusion:** Sequence divergence studies on amniote datasets chosen such that the same sequences are sex-linked in some species while being autosomal in others are missing in the literature and are much needed to decouple gene-specific effects from chromosomal context effects on the rates of sex-linked and autosomal evolution. Upon analyzing such a dataset across 11 amniotes including 6 turtles, we present hypotheses involving the occurrence of a possible selective sweep and neofunctionalization to explain the observed faster evolution with signatures of strong positive selection in two turtle genes (*Nf2*, sex-linked in *Pelodiscus*, and *Tspan7* in *Chrysemys*, TSD) occurring in important functional domains. Both of these genes are associated with vital orchestrators of sex determination in vertebrates by virtue of genetic linkage or direct interaction. Our finding that the slower evolution of sex-linked sequences in the musk turtle *Staurotypus* and the chicken is an intriguing result given that their respective sex chromosomes arose from the same ancestral pair (Kawagoshi et al. 2009), and warrants further investigations on whether this is a result of identity by descent or independent sex chromosome evolution from the same autosomal pair. This work opens the door for genome-level studies to test how generalizable are the patterns observed here, which are increasingly facilitated by the availability of reptilian genomic resources in lizards (Alföldi et al. 2011), snakes (Vicoso et al. 2013), alligators (St John et al. 2012) and turtles (Shaffer et al. 2013; Wang et al. 2013a; Badenhorst et al. 2015). Our work reveals the extensive nonsynonymous divergence at sex-linked loci between closely and distantly related taxa will facilitate further investigations
into the consequences of sex chromosome evolution, including understanding speciation events, the accumulation of sex-biased genes and the efficacy of natural selection on sex chromosomes.

**Acknowledgements:** We thank R. A. Literman for guidance on gene annotation, and D. Lavrov for guidance on the molecular evolution analysis. This work was funded in part by NSF grant MCB 1244355 to N.V.

**Author contributions:** N.V conceived of the project and helped edit the manuscript. S.R performed all bioinformatics analysis and wrote the paper. J.W.T. and A.G provided turtle DNA samples for analysis.
### 4.5 Tables and Figures

Table 4.1: X-, Z-linked and Control genes chosen from Human (HSA), Mouse (MMU), Chicken (GGA), the musk turtle (STR), the softshell turtles (ASP, PSI) and the Anole lizard (ACA).

<table>
<thead>
<tr>
<th>X- or Z-linked</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSA/MMU (X)</td>
<td>GGA (Z)</td>
</tr>
<tr>
<td>Slc25a5</td>
<td>Lingo2</td>
</tr>
<tr>
<td>Tspan7</td>
<td>Rp11</td>
</tr>
<tr>
<td>Rp2</td>
<td>Nxn12</td>
</tr>
<tr>
<td>Glra2</td>
<td>Smad7</td>
</tr>
</tbody>
</table>

* also Z-linked in GGA
Table 4.2: Fast-X computed from branch-model. SL = Sex-linked, A = Autosomal, C = Control genes, FC = Foreground control genes, OC = Other control genes, CI = confidence interval. Species acronyms are described in Table 4.1.

<table>
<thead>
<tr>
<th>Species</th>
<th>ASP</th>
<th>PSI</th>
<th>STR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Category</td>
<td>Z</td>
<td>C</td>
<td>Z</td>
</tr>
<tr>
<td>Original alignment length (bp)</td>
<td>11004</td>
<td>4820</td>
<td>11004</td>
</tr>
<tr>
<td>GBlocks alignment (bp)</td>
<td>10821</td>
<td>3975</td>
<td>10821</td>
</tr>
<tr>
<td>PAML alignment (bp)</td>
<td>8451</td>
<td>3612</td>
<td>8451</td>
</tr>
<tr>
<td>Mean $\omega$ (Sex-linked sequences) [SL]</td>
<td>0.0412</td>
<td>0.1333</td>
<td>0.1363</td>
</tr>
<tr>
<td>95% CI (left)</td>
<td>0.0407</td>
<td>0.1319</td>
<td>0.133</td>
</tr>
<tr>
<td>95% CI (right)</td>
<td>0.0417</td>
<td>0.1348</td>
<td>0.1396</td>
</tr>
<tr>
<td>Mean $\omega$ (Autosomal sequences) [A]</td>
<td>0.0412</td>
<td>0.1333</td>
<td>0.0362</td>
</tr>
<tr>
<td>95% CI (left)</td>
<td>0.0407</td>
<td>0.1319</td>
<td>0.0357</td>
</tr>
<tr>
<td>95% CI (right)</td>
<td>0.0417</td>
<td>0.1348</td>
<td>0.0367</td>
</tr>
<tr>
<td>Relative rate of evolution = [SL:A] or [FC:OC]</td>
<td>1</td>
<td>1</td>
<td>3.765</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Species</th>
<th>GGA</th>
<th>HSA/MMU</th>
<th>ACA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Category</td>
<td>Z</td>
<td>C</td>
<td>X</td>
</tr>
<tr>
<td>Original alignment length (bp)</td>
<td>4788</td>
<td>4820</td>
<td>4161</td>
</tr>
<tr>
<td>GBlocks alignment (bp)</td>
<td>4242</td>
<td>3975</td>
<td>3777</td>
</tr>
<tr>
<td>PAML alignment (bp)</td>
<td>3396</td>
<td>3612</td>
<td>3402</td>
</tr>
<tr>
<td>Mean $\omega$ (Sex-linked sequences) [SL]</td>
<td>0.0389</td>
<td>0.0868</td>
<td>0.4608</td>
</tr>
<tr>
<td>95% CI (left)</td>
<td>0.0371</td>
<td>0.0846</td>
<td>0.2363</td>
</tr>
<tr>
<td>95% CI (right)</td>
<td>0.0407</td>
<td>0.0890</td>
<td>0.6853</td>
</tr>
<tr>
<td>Mean $\omega$ (Autosomal sequences) [A]</td>
<td>0.0636</td>
<td>0.1418</td>
<td>0.0803</td>
</tr>
<tr>
<td>95% CI (left)</td>
<td>0.0625</td>
<td>0.1403</td>
<td>0.0790</td>
</tr>
<tr>
<td>95% CI (right)</td>
<td>0.0647</td>
<td>0.1433</td>
<td>0.0816</td>
</tr>
<tr>
<td>Relative rate of evolution = [SL:A] or [FC:OC]</td>
<td>0.6116</td>
<td>0.612</td>
<td>5.738</td>
</tr>
</tbody>
</table>
Table 4.3: Percentage similarity between Z-chromosomal coding sequences in PSI and ASP. Differences < 90% are highlighted in bold.

<table>
<thead>
<tr>
<th>PSI Scaffold</th>
<th>Gene</th>
<th>% Similarity (Query coverage * % Identity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JH212668.1</td>
<td>Mtmr3</td>
<td>95.5</td>
</tr>
<tr>
<td>JH211589.1</td>
<td>Ascc2</td>
<td>94.1</td>
</tr>
<tr>
<td>JH211589.1</td>
<td>Zmat5</td>
<td>22.79</td>
</tr>
<tr>
<td>JH209995.1</td>
<td>Cabp7</td>
<td>87.6</td>
</tr>
<tr>
<td>JH209995.1</td>
<td>Nf2</td>
<td>71.2</td>
</tr>
<tr>
<td>JH209995.1</td>
<td>Nipsnap1</td>
<td>73.55</td>
</tr>
<tr>
<td>JH209995.1</td>
<td>Ccdc157</td>
<td>49.99</td>
</tr>
<tr>
<td>JH209995.1</td>
<td>Sf3a1</td>
<td>79.15</td>
</tr>
<tr>
<td>JH204957.1</td>
<td>Top3b</td>
<td>98.7</td>
</tr>
<tr>
<td>JH204957.1</td>
<td>Sdf2l1</td>
<td>95.74</td>
</tr>
<tr>
<td>JH204957.1</td>
<td>Ydjc</td>
<td>93.6</td>
</tr>
<tr>
<td>JH204957.1</td>
<td>Ubel2l3</td>
<td>93.08</td>
</tr>
</tbody>
</table>
Figure 4.1: Phylogenetic relationships among selected vertebrates used in this study and their sex-determining mechanism.
Figure 4.2: (Left panel) Molecular evolution [dN/dS (ω) ratio] of sex-linked (green) and autosomal sequences (red) measured in selected vertebrates as determined by the branch model. Analyses used the sex-linked sequence in each focal species (x-axis) as foreground and the autosomal orthologs from every other species (Other) as background. A blue ω value denotes that no difference exists in the evolution of sex-linked relative to autosomal orthologs as determined by the branch model, i.e. cases where the green and red dots overlap. (Right panel) Molecular evolution measured by an identical analysis of control genes from each focal species (green), background species (red), and cases where no difference was detected (blue). [Legend for species: ACA: *Anolis carolinensis*, the green anole; ASP: *Apalone spinifera*, the spiny softshell turtle, HSA: *Homo sapiens*, human; MMU: *Mus musculus*, mouse; GGA: *Gallus gallus*, chicken; STR: *Staurotypus triporcatus*, the musk turtle, PSI: *Pelodiscus sinensis*, the Chinese softshell turtle]
Figure 4.3: Rates of molecular evolution [dN/dS (w) ratio] in the genes of interest and control genes measured in focal vertebrate TSD species (CPI or AMI; green) against GSD species (red). A blue w value denotes that no difference exists in the evolution of TSD sequences relative to their GSD orthologs as determined by the branch model (i.e. cases where the green and red dots overlap). CPI = Chrysemys picta; AMI = Alligator mississippiensis. Error bars represent 95% confidence interval ranges from bootstrapped alignments among all sex-linked and all control genes in the foreground and the background species.
Figure 4.4: Rates of molecular evolution [dN/dS (ω) ratio] in the genes of interest and control genes measured in focal clades (green) against background clades (red). A blue ω value denotes that no difference exists in the evolution of focal clade sequences relative to their orthologs in the background clade as determined by the branch model, i.e. cases where the green and red dots overlap.
Figure 4.5: Overlaid structural alignment of (a) NF2 proteins in *Pelodiscus sinensis* and *Apalone spinifera* (b) Tspan7 proteins in *Chrysemys picta* and *Glyptemys insculpta*. Structures were computed using the i-Tasser web server. The rmsd between structures is indicated by a spectrum of colors, with higher rmsd indicated in red, lower rmsd in blue. Unalignable regions are indicated in grey.
Figure C1: Coding DNA and amino acid sequence alignment of Nf2 in Pelodiscus sinensis (PSI) and the consensus sequence of other studied amniotes. DNA binding and FERM domains in Nf2 are identified. Positively selected residues undergoing a change in charge, hydrophobicity and polarity are highlighted.
Figure C2: Coding DNA and amino acid sequence alignment of Tspan7 in Chrysemys picta (CPI) and the consensus sequence of other studied amniotes. Transmembrane family and DNA binding domains in Tspan7 are identified. Positively selected residues undergoing a change in charge, hydrophobicity and polarity are highlighted.
CHAPTER 5: CONCLUSION

Sex determination in vertebrates is widely studied because it occurs in a startling variety of modes, spanning a spectrum of genotypic- (or GSD) and environmental (although mainly temperature-dependent, or TSD) sex determination. The resolution of sexual fate through both genetic and environmental cues has raised numerous questions over the years on topics such as the adaptive nature of TSD and the evolutionary transitions between GSD and TSD systems. Further questions have been raised on the consequences of the emergence of sex chromosomes from ancestral autosomes, such as the existence of the differences in mutation rate and effectiveness of natural selection on sex chromosomes relative to autosomes, and whether genes with sex-specific fitness effects tend to differentially accumulate on sex chromosomes. These are recalcitrant mysteries about sex determination and sex chromosome evolution that have evaded scientific explanation for years. Turtles are a reptile lineage possessing both TSD and GSD and hence are an ideal system to fill some of these gaps. The molecular basis of sex determination in these vertebrates remains a jigsaw whose pieces have only been partially identified, unlike mammals whose urogenital networks have been better characterized. This dissertation contributes to solve this puzzle by illuminating the genetic and epigenetic basis of TSD, and revealing the effect of sex-linked versus autosomal chromosomal context on the evolution of DNA coding sequences.

In Chapter 2, I show that gonadal transcriptomes of turtles with TSD and GSD regulatory networks consist of previously known plus unknown elements that are mostly common to both systems but deployed differentially by temperature in TSD and GSD turtle species. This transcriptomal study paves the way for future functional tests of multiple
candidate genes identified here that could potentially mediate the temperature cue and determine sex in TSD turtles. In Chapter 3, I investigate the potential link between epigenetic mechanisms as a molecular underpinning of TSD, by first computing the extent of DNA methylation in a TSD turtle in silico, and validating these results experimentally using MeDIPseq in TSD turtle hatchlings. I propose that the normalized CpG content is a reasonable estimator of methylation status in TSD turtles and show that DNA methylation occurs differentially by temperature in male and female TSD turtle hatchlings. Further, results suggest that multiple candidates showing higher transcription at temperatures that produce a single sex in Chapter 2 could have been methylated at temperatures that produce the opposite sex; suggesting that DNA methylation could be the mediator of the temperature cue that ultimately decides sexual fate. As with Chapter 2, I identify multiple genes that are differentially methylated by temperature that can guide future functional assays.

Importantly, since chapters 2 and 3 are genome-wide characterizations of temporal gene expression and DNA methylation, they represent the initial steps in a longer-term effort to build complete molecular networks across a broad range of taxa to better understand the genetic basis and the scale of epigenetic involvement in sex determination in TSD and GSD species. Further, phylogenetic comparisons of such molecular networks across various branches of the tree of life would help decipher the evolution of the genetic architecture of these systems and their underlying evolutionary drivers. Thus, my work contributes the first characterizations of temperature-dependent- regulation of temporal transcriptional patterns during embryonic development and of DNA methylation patterns in males and females in TSD and GSD turtle as a critical first step in this direction.
In Chapter 4, I unravel species-specific patterns of faster/slower molecular evolution in sex-linked coding sequences in some amniotes relative to their autosomal orthologs in others. I also reveal faster evolution of TSD-specific sequences relative to GSD sequences. I provide hypotheses that could have driven these patterns of evolution of sequences in closely and distantly related taxa. These divergence analyses are the first of their kind in reptilian species, and open the door to explore the sex chromosomal basis of reptilian speciation, as sex-linked gene divergence has been disproportionately linked to sterility in hybrid crosses in mammals. These analyses on the differential rate of molecular evolution based on sex linkage are the first step in undertaking larger reptilian investigations on deciphering the relative mutation rates and efficacy of selection on sex chromosomes and autosomes, and the build-up of genes inducing sex-specific fitness onto sex chromosomes. These bigger challenges demand larger annotated datasets spanning greater real estate of reptilian genomes, resources that are becoming increasingly available in this next-generation sequencing dominated era.

This body of work, largely performed with turtle datasets, has implications for human health because humans and turtles share a common ancestor, such that understanding sex determination in turtles illuminates on the evolution of the human genome and reproductive disorders that are controlled by common genes. Understanding the epigenetic underpinnings of TSD has implications for conservation since many TSD taxa are challenged by environmental change and epigenetic mechanisms are thought to mediate responses to environmental input. This work entails careful experimental design, robust next-generation sequencing technologies and statistically sound bioinformatics analyses and thus will emerge as a model to study the evo-devo of non-model organisms in
a genomic framework. Importantly, this study will illuminate the genetic and epigenetic basis of a complex system (sexual phenotype), its environmental susceptibility (TSD v GSD), and evolution (divergence among species). In essence, my thesis explores the molecular underpinnings of sex determination and unlocks avenues to investigate the genetic and epigenetic regulation of gonadal dimorphism as driven by temperature cues, and the consequences of sex chromosome evolution. I hope this body of work raises previously unanswered questions and invigorates further research, ultimately inching closer towards deciphering the molecular forces behind the puzzle that is sex determination.


http://www.genetics.org/content/196/4/1131.full.


http://dx.doi.org/10.1111/2041-210X.12228.


Mahajan MA, Samuels HH. 2008. Nuclear receptor coactivator/coregulator NCoA6(NRC) is a pleiotropic coregulator involved in transcription, cell survival, growth and development. *Nuclear receptor signaling* 6: e002.


Pearson K. 1900. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **50**: 157–175.


Wilson EB. 1905. *Studies on chromosomes*.


ACKNOWLEDGEMENTS

When I first heard about temperature determining sex from a kid’s program on The Discovery Channel back in the mid ‘90s, I remember being very taken with idea. Fast forward two decades – I’ve since had the opportunity to be able to work on the molecular basis of the very same system. Writing a dissertation on the topic has been a humbling experience that has made me gain an appreciation for some of the great things that nature is capable of.

I wish to thank my advisor, Dr. Nicole Valenzuela for her major hand at guiding me through this journey. This would not have been possible without her continued interest in my progress and constant constructive feedback.

I wish to thank all my POS committee members: Drs. Amy Toth, Peng Liu, Matt Hufford and Eric Henderson for their input on my work at various phases of my dissertation.

I wish to thank all members of the Valenzuela, Serb and Adams labs that have provided feedback on my manuscripts. I particularly thank Dr. Dean Adams for making statistics intuitive and fun, and more importantly, being a big inspiration on my running pursuits.

I wish to thank the EEOB department for their financial assistance by way of teaching assistantships over this period. I particular thank Dr. Jim Colbert and Linda Westgate for giving me an opportunity to teach biodiversity to undergraduate students, a four-year experience that has left me all the richer, and as I would like to believe, wiser.
I also wish to acknowledge Trish Stauble who has always fussed over and been there for all BCB students. I thank my BCB peers for all their help: Dr. Ali Berens and the soon-to-be Dr. Divita Mathur. I also particularly thank Dr. Rasna Walia, for having me in her thoughts. I wish them the very best in all their pursuits.

My dad, mom and sisters in India have been behind me all of this time. Their unconditional love and support through my lean and mean times propped me up and assured me that the light at the end of the tunnel was not that of an oncoming train. I love them and miss them so much.

And finally – I’ve shared all my highs and lows over this period with my wife Hema, who has been the bedrock of my life and happiness in the States. But for her words of encouragement and prolonged six-year endurance of a 2000 mile-long relationship, I wouldn’t be in a happy place today.