Summer 2020

Characterization and Analysis of the Avian Epidermal Differentiation Complex and Implications in Feather Evolution

Anthony Chastine Davis

Recommended Citation

Davis, A. C.(2020). *Characterization and Analysis of the Avian Epidermal Differentiation Complex and Implications in Feather Evolution.* (Doctoral dissertation). Retrieved from https://scholarcommons.sc.edu/etd/6062

CHARACTERIZATION AND ANALYSIS OF THE AVIAN EPIDERMAL DIFFERENTIATION COMPLEX AND IMPLICATIONS IN FEATHER EVOLUTION

by

Anthony Chastine Davis

Bachelor of Science
University of South Carolina, 2016

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Biological Sciences

College of Arts and Sciences

University of South Carolina

2020

Accepted by:

Roger Sawyer, Major Professor

Shannon Davis, Committee Chair

Soumitra Ghoshroy, Committee Member

Wayne Carver, Committee Member

Robert Friedman, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

DEDICATION

To my parents, John and Laure, without whom none of this would have been possible. Thank you for everything and I love you both very much. Also, to my friends who stuck by my side the entire time. Thank you.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Roger H. Sawyer for all his time and assistance throughout this entire process. Dr. Sawyer gave me an opportunity to get involved in research as an undergraduate and was a continuous source of advice and inspiration. When I first joined Dr. Sawyer's lab, I had little idea of what I wanted to accomplish or the direction I wanted to take my studies. His thoughtful advice and understanding have allowed me to acquire a wide range of research and problem-solving skills that will help me in my future career. He always encouraged me to work hard and have confidence in myself. I would also like to thank my former lab mates Dr. Weier Bao and Dr. Matthew Greenwold. Both accepted me with open arms and never hesitated to help me. Without the continuous help of Dr. Greenwold, even after his departure from our lab, none of my work would have been possible and I am beyond thankful and lucky to have had such a great mentor and friend over these years. I would also like to thank my committee members Dr. Shannon Davis, Dr. Wayne Carver, Dr. Soumitra Ghoshroy, and Dr. Bob Friedman for all their time and support. Furthermore, I need to acknowledge the many faculty members at the University of South Carolina who taught and assisted me in my research over the years such as Dr. Jeff Dudycha, Dr. Mihaly Czako, Dr. Margaret Cirtain, and Dr. Robert Ogilvie, all were instrumental in my success.

Finally, thank you to the University of South Carolina for being my home for the last 8 years of undergraduate and graduate studies. I love this place with all my heart and will never forget my time here. The relationships I have made here and the greater USC/Columbia community are ones I will treasure for the rest of my life. Forever to thee.

ABSTRACT

Since the most recent major mass-extinction event ~65 million years ago, birds have expanded to now occupy a wide range of habitats and exhibit diverse lifestyles. A major reason for this evolutionary success is the mechanical resilience and diversity of their epidermal appendages such as feathers, scales, and beaks. The diversity of these appendages, specifically feathers has played a critical role in their evolutionary success. The feathers of birds vary substantially across different species, as well as at different life stages and anatomical locations on an individual bird. Several of the genetic elements involved in the development and structure of feathers are located at a specific genetic locus known as the Epidermal Differentiation Complex (EDC). To gain a better understanding of the genes and proteins involved in these processes as well as how genetic variation in these elements has accompanied the evolution of diverse lifestyles and phenotypes in birds, we have characterized the organization and architecture of the EDC locus across 48 diverse bird species. We have also investigated two specific gene families within the avian EDC, loricrins and a group of EDC genes rich in aromatic amino acids, which also contain a conserved sequence of Methionine-Threonine-Phenylalanine (MTF) residues at their start (EDAA/EDMTFs), to analyze their evolution in birds as well as their roles in epidermal development. Our results demonstrate that the avian EDC is conserved across birds and evolved from a common amniote ancestor. Furthermore, we show that these ancestral EDC genes have expanded in birds into large gene families but have not translocated to other parts of the genome. We also provide evidence that these gene families of the EDC have expanded via significant amounts of gene loss and duplication events many of which are lineage specific. Finally, given that the amino acid compositions of structural proteins play a significant role in function, we investigate the amino acid contents of identified avian EDC genes and demonstrate that they contain amino acid residues commonly associated with epidermal development and structure. Overall, our results support that the evolution of the avian EDC accompanied the evolution of bird species with diverse feather morphologies and phenotypes, which has played a key role in their evolutionary success.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS

β	represents Beta. Used to define Beta-keratins

LIST OF ABBREVIATIONS

EDAA	Epidermal Differentiation Protein rich in Aromatic Amino Acids
EDC	Epidermal Differentiation Complex
EDMTF Epidermal	Differentiation Protein with an MTF sequence
EDCRP	Epidermal Differentiation Protein Rich in Cysteine
HRP/fp	Histidine Rich Protein/Fast Protein
LOR	loricrin
PCA	Principle Component Analysis

CHAPTER 1

INTRODUCTION

1.1 - Introduction

The evolution of the epidermis as a physical barrier to limit water loss to the environment was crucial in the transition of amniotes to terrestrial lifestyles (Strasser et al. 2014). Adaptations to the amniotic epidermis have played a key role in the evolution of fully terrestrial organisms with diverse habitats and lifestyles (Chuong 2002). Mechanically resilient appendages such as hair, scales, and feathers perform a variety of important functions including but not limited to thermoregulation, camouflage, and sensory applications (Pierard et al. 2000; Strasser et al. 2014). Analyzing the evolutionary origin as well as developmental and genetic aspects of epidermal appendages is critical in understanding the history of life on earth and the adaptation of novel structures.

The amniotic epidermis is a stratified organ that consists of an outer stratum corneum layer made up of dead cells which is constantly shed and replaced (Gilbert 2014). The specific layers of the epidermis change depending on the type of organism as well as the stage of development, with some layers only being present during embryonic development and others becoming parts of epidermal appendages. Progressing from the Basal layer, the epidermis contains a spinous layer, a granular layer, and the outermost Cornified layer (Candi et al. 2005, Eckhart et al. 2013). The primary cell type of the epidermis is the keratinocyte. As keratinocytes differentiate, they move from the basal cell layer of the epidermis, through the spinous and granular layers before accumulating in the stratum corneum (Candi et al. 2005). Differentiation of keratinocytes involves a specialized mode of programmed cell death known as cornification which takes place as they migrate through the epidermis and results in the formation of dead corneocytes whose plasma membranes have been replaced by a cornified envelope (CE). The CE of these terminally differentiated keratinocytes is what confers many of the unique physical properties to the epidermis and helps facilitate its barrier function (Kalinin et al. 2002). These dead cells are continuously desquamated, or shed from the top layer, and replaced by new cells from the basal layer (Eckhart et al. 2013).

The specialized process of terminal differentiation of keratinocytes known as cornification confers many of the mechanically resilient properties to the epidermis by replacing the cells' plasma membrane with an insoluble Cornified envelope (Candi et al. 2005). The formation of the CE is a stepwise process which includes Initiation, Reinforcement, Lipid-envelope formation and finally desquamation. During the initiation stage, structural proteins and lipids are synthesized intracellularly by the epidermal cells of the spinous layer and packaged for transport toward the cell surface. Concurrently, important scaffolding components such as Envoplakin and Periplakin are anchored to the internal surface of the cell membrane (Candi et al. 2005). During reinforcement, the newly synthesized CE proteins and lipids are covalently attached within the granular layer and packaged into lamellar bodies for transport while important structural proteins such as loricrin and small-proline rich proteins (SPRs) are crosslinked to the cell membrane. In mammals, the primary method of crosslinking observed in CE assembly is transglutamination, however other types of protein-crosslinking such as disulfide bonding are also observed (Saathoff et al. 2004). Lipid-envelope formation entails the attachment and exposure of the CE proteins and lipids on the outside of the cell membrane and occurs concomitantly with reinforcement (Eckhart et al. 2013). The desquamation phase entails the shedding of the superficial corneocytes, which are then replaced with freshly cornified cells from the basal layers via repeating the entire process from initiation. The specific physical properties of a cell's CE depend on the identity of the protein components which comprise it, as well as the method and degree of protein crosslinking that takes place during the development (Candi et al. 2005).

Avian feathers are hierarchically organized and branched structures of variable complexity, made of up cornified keratinocytes (Alibardi 2016). The anatomy and structure of avian feathers provide them with unique and physical properties. These properties are key in providing the aerodynamic prerequisites required for flight (Kondo et al. 2018). Much like mammalian epidermal development, the development of the avian epidermis and appendages is a complex process which involves several elements. Specifically, Feather development is a complex spatiotemporal process which entails a close association of the epidermis with the underlying mesenchymal dermal cells (Alibardi 2017). The anatomy of a mature feather generally consists of a hollow calamus or shaft anchored to a feather germ, which serves as the source of undifferentiated cells for the growing feather and an anchor point between the epidermis and underlying mesenchymal tissue. Extending from the calamus is the central rachis forming the central feather vane. Branching off from the central vane are the barbs and fused to the barbs are branches of smaller barbules, which can also contain hooklets (Ostmann et al. 1963) (Figure 1.1 – Kazilek 2009). It is the branching organization of the barbs and barbules from the central rachis which give feathers their complex hierarchical organization. These distinct anatomical aspects of feathers are made up by specialized populations of cells such as barb/barbule cells as well as other supportive cells, capable of differential expression of specific genes, which later cornify and fuse in a process unique to feather formation.

In general, feathers derive from thickenings in the embryonic epithelium of feather placodes in which the differentiation of barb/barbule and the degeneration of other supportive cells, forms a branching organization (Figure 1.2) (Alibardi 2017). Throughout embryonic development, several epidermal layers are formed, including the periderm and archosaur-specific subperiderm (Alibardi et al. 2009). In avian scale development, the subperiderm is sloughed off prior to hatching, while in feathers, studies have provided evidence it is incorporated as part of the mature feather (Sawyer et al. 1974, Sawyer and Knapp 2003, Alibardi 2016). Initial feather formation begins with the formation of dermal condensations which in turn induce the thickening of the overlying epidermis (Sawyer and Knapp 2003). Next, the dermal condensations rise and form a cylindrical outgrowth composed of a mesodermal core surrounded by an epithelial sheath (Figure 1.2). As epithelial cells divide longitudinally, the feather grows longer and epidermal folds known as barb ridges form to accompany the additional volume of cells, which point inward towards the mesodermal core (Alibardi 2016). At the apex of each barb ridge, large columnar cells differentiate and fuse into elongated shafts know as Barbs which will eventually fuse with the rachis resulting in the branching pattern observed in feathers (Figure 1.2). With the help of additional supportive cells, the outermost cells of the barbs branch off into barbules. Eventually these supportive cells degenerate, the remaining barbs and barbule cells cornify and fuse and the epithelial sheath is lost resulting in the branched structure of a mature feather (Alibardi 2017). The fusion of the barb and barbule cells is a unique process only observed in feather development (Sawyer and Knapp 2003)

Feathers come in a large variety of colors, shapes and sizes across different bird species, and vary at different life stages as well as at anatomical locations on an individual bird. Plumaceous feathers, or down feathers, provide insulation to newly hatched chicks, while the structure of pennaceous or flight, feathers provide the aerodynamic structure that is key for flight (Kischer 1963, Norberg 1985). The development of plumaceous and pennaceous feathers both follow the same general process mentioned earlier, however the specific genetic and protein elements involved differ. Furthermore, differences in the distribution and molecular makeup of feathers can result in a wide range of physical properties. For example, the pennaceous feathers of penguins display anti-icing properties which prevent the accumulation of ice on their wet feathers (Li et al. 2014, Wang et al. 2016)**.** The structural diversity and unique physical properties observed across feathers are the result of differences in the many genetic elements involved in their development (Strasser et al. 2014).

In mammals, many of the genetic elements responsible for the development and structure of hair and nails are found at a genetic locus known as the Epidermal Differentiation Complex (EDC) (Kypriotou et al. 2012). These genes contain several sequence elements which are indicative of structural proteins, and they are expressed in various epidermal tissues throughout development as well as in mature epidermis and appendages (Strasser et al. 2014). There is also evidence that local, gene-specific factors are key in the regulation of EDC gene expression indicating that some EDC genes may be involved as enhancers or repressors of gene expression during epidermal differentiation (Elder and Zhao 2002). Mammalian EDC genes are rich in amino acid residues such as cysteine, tyrosine and serine, residues which are known to be involved in the cornification process by facilitating and participating in protein crosslinking (Candi et al. 20015). Disulfide bonding, transglutamination and enzymes such as serine proteases have been shown to be crucial in proper epidermal development and function (Hynes and Destree 1977, Robinson et al. 1997, Leyvraz et al. 2005). During the development of mammalian epidermal appendages, these genetic elements serve various purposes including signal transduction, matrix assembly as well as direct structural roles (Kypriotou et al. 2012). For example, the mammalian EDC gene Loricrin, a major component of the cornified envelope, has been shown to influence both tensile and flexural properties of the epidermis and appendages (Steinert et al. 1991). Moreover, Loricrin-deficient mice did not initially develop a healthy or normal epidermis. Their epidermal cells exhibited several problems with epidermal barrier-function and maintaining homeostasis with desquamated epithelial cells during early development (Ishida-Yamamoto et al. 1998). Interestingly, loricrin-deficient mice regained mostly normal epidermal function as they aged indicating the complexity of the elements involved in epidermal development and that there is possible genetic redundancy in the process. Other mammalian EDC genes such as Cornulin (CRNN) have also been shown to play important roles in the assembly of the CE and development of the epidermis and appendages (Contzler et al. 2005).

Until recently, many of the genes involved in the development and structure of feathers were unknown, however recently a genomic locus homologous to the mammalian EDC has been identified in birds and other reptiles (figure 1.3) (Strasser et al. 2014, Strasser et al. 2015, Holthaus et al. 2015, Alibardi et al. 2016, Holthaus et al. 2018, Holthaus et al. 2018, Lachner et al. 2019). The Avian EDC contains mammalian EDC homologs such as Lorcrin and CRNN, but also contains several avian-specific genes including a cluster of genes which encode β-keratins, the primary components of mature feathers (Strasser et al. 2014, Greenwold and Sawyer 2014). All avian EDC genes identified in the chicken were found to be expressed in at least 1 epidermal tissue (scale, claw, beak, feather, skin) on day 18 of embryonic development (Strasser et al. 2014). Additional studies on specific avian EDC genes such as Epidermal Differentiation Cysteine rich protein (EDCRP) and Epidermal Differentiation protein containing DPCC motifs (EDDM) have demonstrated that diversification and differential expression of EDC genes was instrumental in facilitating the evolution of complex structures such as feathers and scales (Strasser et al. 2015, Lachner et al. 2019).

This dissertation aims to characterize the avian EDC as well as investigate its' evolutionary history in order to gain a better understanding of how genetic diversity in the EDC has accompanied the adaptation and development of novel and diverse epidermal structures in birds. These aims are achieved by first identifying the EDC loci across 48 diverse bird species by performing iterative rounds of reciprocal blast searches. A clear understanding of the architecture of the avian EDC will give us insight into its evolution from early amniote and archosaur ancestors. In order to closer examine the evolutionary relationships of birds with one another as well as in the greater tree of life, we use phylogenetic methods to build detailed gene trees. We also investigate the possible function of several EDC genes by analyzing their respective amino acid and nucleotide compositions to provide evidence that EDC genes may function as important structural elements of epidermal appendages. Overall these results give insight into how evolution of genetic elements of epidermal differentiation has accompanied the adaptation of diverse and morphologically complex appendages such as feathers. Furthermore, we provide a basis for future studies regarding the specific function of avian EDC genes in epidermal development.

CHAPTER 2 of this study identifies and characterizes the avian EDC loci across 48 diverse bird species, explores their respective architectures and serves a general introduction into the results of this detailed investigation of the avian EDC. It details the conserved organization of the avian EDC as well as the difficulties presented in identifying several EDC genes. We also provide evidence that avian EDC genes could be classified as Dark DNA, genomic regions or genes which contain higher than average G/C contents and often go undetected or are reported as missing from genomic databases (Hron et al. 2015, Bornelöv et al. 2017). Specifically, we analyze the G/C content of avian EDC genes as well as the presence of G/C nucleotide stretches and demonstrate that many avian EDC genes meet the criteria laid out by previous studies for Dark DNA.

CHAPTER 3 investigates genes homologous to human Loricrin, a major component of the mammalian cornified envelope. We use phylogenetic analyses to explore the complex evolutionary history of avian loricrins as well as sequence and amino acid analyses which examine the repetitive yet extremely diverse loricrin sequences observed across birds. Finally, we provide evidence that avian loricrins are candidates to take on a specialized protein conformation known as a Glycine-loop which is key in contributing elasticity and tensile strength to the epidermis and its appendages.

CHAPTER 4 analyzes a conserved avian EDC gene family known as Epidermal Differentiation protein rich in aromatic amino acids containing MTF motifs

(EDAAs/EDMTFs), including the previously reported chicken Histidine-Rich Protein (HRP). Like analysis of avian loricrins, we use phylogenetic methods to examine the evolutionary history of EDAAs in reptiles and birds. We demonstrate that the EDAA gene family originated in a common archosaur ancestor and has expanded in birds, crocodilians and testudines respectively. We also use sequence and amino acid analyses to investigate the possible function of EDAAs in epidermal development and provide a basis for future studies.

CHAPTER 5 serves as a general conclusion to the dissertation and presents thoughts on future studies based on these results.

1.2 Figures

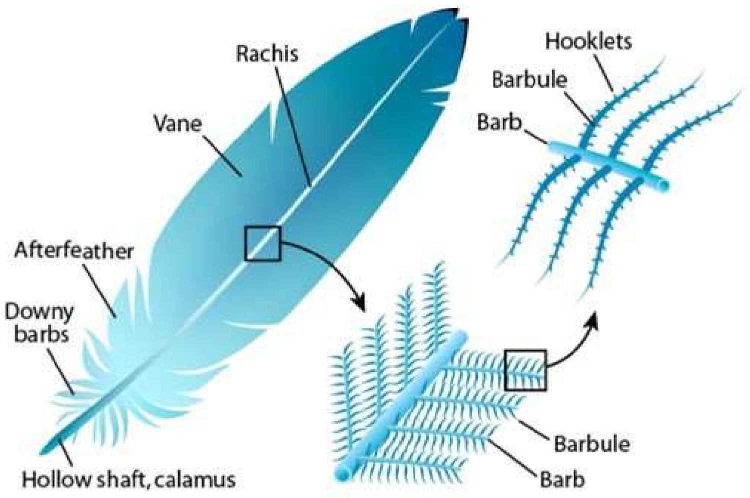


Figure 1.1 : Feather Anatomy and structure. Image taken from Kazilek (2009). (https://askabiologist.asu.edu/explore/feather-biology). Image details the anatomy of a feather. The feather generally consists of a hollow shaft or calamus anchored to the bird which extends outward into the rachis. From the rachis branch of the Barbs. Branching off from the barbs are the barbules and from them the hooklets. Together this elements form the complex hierarchical branching structure of a feather.

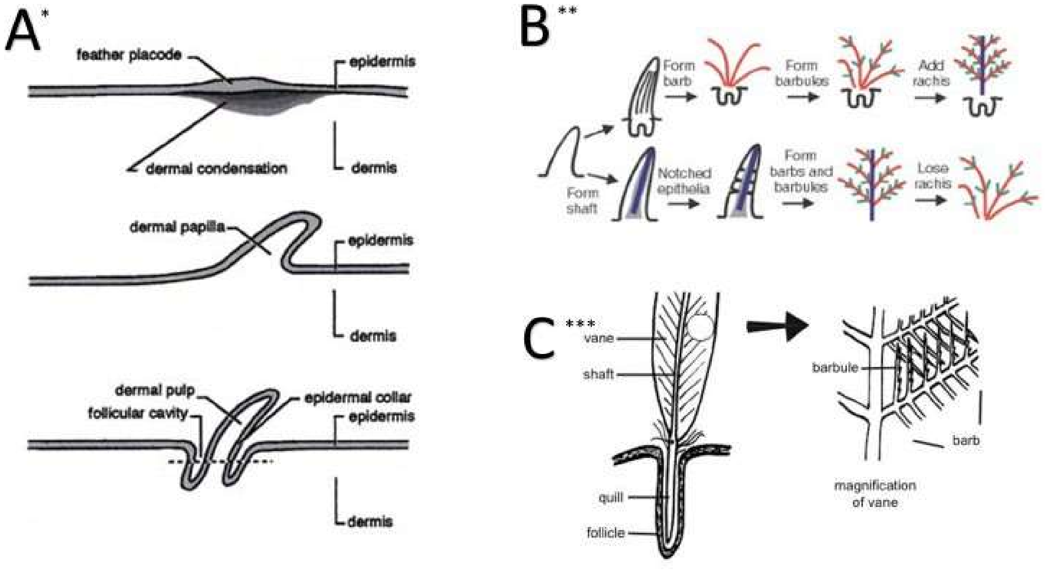


Figure 1.2 – Details the general process of feather formation. (A) shows the initial formation of the feather placode via thickening of the epidermis and underlying dermis. This is followed by the formation of a cylindrical outgrowth with an epidermal sheath and resulting follicular cavity. (B) Details the general process of degeneration of the epidermal sheath and the branching of the barbs and barbules. (C) The final steps are the appearance of mature feather branching structure, complete with barbs, barbules and hooklets. * taken from National Center for the Study of Cladistic Existentialism (http://ncsce.org/pages/feathers.html). ** from Yu et al. 2002. *** from Science Learning Hub – The University of Waikato. (2007). *Feathers and Flight*. Retrieved from https://www.sciencelearn.org.nz/resources/308-feathers-and-flight

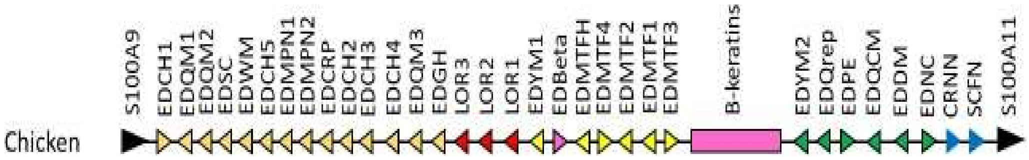


Figure 1.3 – Organization of Chicken EDC locus identified by Strasser et al. (2014). Figure depicts the organization and architecture of the EDC identified in the Chicken by Strasser et al. 2014. All genes were annotated by Strasser et al. 2014 . EDC genes focused on in this study have a black outline. The colors of the arrows indicate different groups of genes based upon amino acid contents and speculated ancestry. The box labeled "β-keratins" represents more than 50 genes.

1.3 – References

1. Alibardi L, Valle LD, Nardi A, Toni M. (2009). Evolution of hard proteins in the sauropsid integument in relation to the cornification of skin derivatives in amniotes. *J. Anat.* 214. 560-586. doi:10.1111/j.1469-7580.2009.01045.

2. Alibardi L, Holthaus KB, Sukseree S, Hermann M, Tschaler E, Eckhart L. (2016). Immunolocalization of a Histidine-Rich Epidermal Differentiation Protein in the Chicken Supports the Hypothesis of an Evolutionary Developmental Link between the Embryonic Subperiderm and Feather Barbs and Barbules. *PLOS One* 11(12): e0167789. doi:10.1371/journal.pone.0167789.

3. Alibardi L. (2017). Review: cornification, morphogenesis and evolution of feathers. *Protoplasma*. 254(3):1259-1281. doi: https://doi.org/10.1007/s00709-016-1019-2

4. Candi E, Schmidt R, Melino G. (2005). The cornified envelope: a model of cell death in the skin. *Nat. Rev. Mol. Cell Bio*. 6(4): 328-340. doi:10.1038/nrm1619

5. Chuong CM et al. (2002). What is the 'true' function of skin? *Exp Dermatol.* 11:159-187. doi: https://doi.org/10.1034/j.1600-0625.2002.00112.x.

6. Contzler R, Favre B, Huber M, Hohl D. (2005). Cornulin, a New Member of the "Fused Gene" Family, is Expressed During Epidermal Differentiation. *J. Invs. Derm.* Vol. 124, Issue 5: 990-997. Doi:https://doi.org/10.1111/j.0022-202X.2005.23694.x.

7. Elder JT, Zhao X. (2002). Evidence for local control of gene expression in the epidermal differentiation complex. *Exp. Derm.* Vol. 11, Issue 5: 406-412. doi:https://doi.org/10.1034/j.1600-0625.2002.110503.

8. Eckhart L, Lippens S, Tschaler E, Declercq W. (2013). Cell death by cornification. *BBA-Mol. Cell Res.* 1833(12): 3471-3480. doi: https://doi.org/10.1016/j.bbamcr.2013.06.010.

9. Gilbert SF. 2014. Developmental Biology, 10th edition. Pgs. 79-83.

10. Hunter S. (2002). Feathers: Whats Flight got to do – got to do with it. http://ncsce.org/pages/feathers.html.

11. Holthaus KB, Mlitz V, Strasser B, Tschaler E, Alibardi L, Eckhart L. (2017). Identification and comparative analysis of the epidermal differentiation complex in snakes. *Sci. Rep.* 7, 45338 doi: 10.1038/srep45338.

12. Holthaus KB, Strasser B, Sipos W, Schmidt HA, Mlitz V, Sukseree S, Weissenbacher A, Tschaler E, Alibardi L, Eckhart L. (2015). Comparative genomics Identifies Epidermal Proteins Associated with the Evolution of the Turtle Shell. *Mol. Biol. Evol.* 33(3):726-737. doi: 10.1093/molbev/msv265.

13. Holthaus KB et al. (2018). Comparative analysis of epidermal differentiation genes of crocodilians suggests new models for the evolutionary origin of avian feather proteins. *Gen. Biol. Evol.* 10(2): 694-704. doi: 10.1093/gbe/evy035.

14. Hynes RO, Destree A. (1977). Extensive disulfide bonding at the mammalian cell surface. *Proc. Natl. Acad. Sci.* Vol. 74, No. 7: 2855-2859.

15. Ishida-Yamamoto A., Takahashi H., Iizuka H. (1998). Loricrin and human skin diseases: molecular bases of loricrin keratodermas. *Histol. Histopathol.* 13(3):819-826. doi: 10.14670/HH-13.819

16. Kazilek CJ. (2009). Feather Biology. ASU – Ask A Biologist. Retrieved April 16, 2020 from https://askabiologist.asu.edu/explore/feather-biology

17. Kondo M, Sekine T, Miyakoshi T, Kitajima K, Egawa S, Seki R, Abe G, Tamura K. (2018). Flight feather development: its early specialization during embryogenesis. *Zoological Lett.* 4:2. doi: 10.1186/s40851-017-0085-4

18. Kalinin AE, Kajava A, Steinert PM. (2002). Epithelial barrier function: assembly and structural features of the cornified cell envelope. *Bioessays*. 24:789-800.

19. Kischer CW. (1965). Fine structure of the developing down feather 1. 1963. *J. Ultstr. Res.* Vol. 8, (3-4): 305-321.

20. Kypriotou M, Huber M, Hohl D. (2012). The human epidermal differentiation complex: cornified envelope precursors, S100 proteins and the 'fused genes' family. *Exp. Dermatol.* 21(9):643-649. doi: https://doi.org/10.1111/j.1600-0625.2012.01472.x.

21. Lachner J, Ehrlich F, Mlitz V, Harmann M, ALibardi L, Tschaler E, Eckhart L. (2019). Immunolocalization of phylogenetic profiling of the feather protein with the highest cysteine content. *Protoplasma*. doi:https://doi.org/10.1007/s00709-019-01381-3.

22. Li, C., Zhang, Y., Li, J., Kong, L., Hu, H., Pan, H., Xu, L., Deng, Y., Li, Q., Jin, L., Yu, H., Chen, Y., Liu, B., Yang, L., Liu, S., Zhang, Y., Lang, Y., Xia, J., He, W., Shi, Q., … Zhang, G. (2014). Two Antarctic penguin genomes reveal insights into their evolutionary history and molecular changes related to the Antarctic environment. *GigaScience*, *3*(1), 27. https://doi.org/10.1186/2047-217X-3-27

23. Leyvraz C, Charles RP, Rubera I, Guitard M, Rotman S, Breiden B, Sandhoff K, Hummerl E. (2005). The epidermal barrier function is dependent on the serine protease CAP1/Prss8. *J. Cell Bio.* 170(3): 487-496. doi:https://www.jcb.org/cgi/doi/10.1083/jcb.200501038.

24. Norberg UM. (1985). Evolution of Vertebrate Flight: An Aerodynamic Model for the Transistion from Gliding to Active Flight. *American Naturalist*. Vol. 126, No. 3. doi:10.1086/284419

25. Ostmann OW, Ringer RK, Tetzlaff M. (1963). The anatomy of the feather follicle and its immediate surroundings. *Poultry Sci.*, 42957-969.

26. Pierard G, Goffin V, Hermanns-Le T, Pierard-Franchimont C. (2000). Corneocyte desquamation. *Int. J. Mol. Med.* 6(2):217-238. doi: https://doi.org/10.3892/ijmm.6.2.217.

27. Robinson NA, Lapic S, Welter JF, Eckert RL. (1997). S100A11, S100A10, Annexin I, Desmosomal Protiens, Small Proline-rich Protiens, Plasminogen Activator Inhibitor-2, and Involucrin are components of the cornified envelope of cultured human epidermal keratinocytes. *J. Biol. Chem.* Vol 272, No. 18: 12035-12046.
28. Sawyer RH, Abbott UK, Fry GN. (1974). Avian Scale Development III: Ultrastructure of the keratinizing cells of the outer and inner epidermal surfaces of the scale ridge. *J. Exp. Zool.* Vol. 190, No.1: 57-70.
29. Sawyer RH, Knapp LW. (2003). Avian skin development and the evolutionary origin of feathers. *J. Exp. Zool. MDE.* 298B:57-72. doi:10.1002/jez.b.00026.
30. Saathoff M, Blum B, Quast T, Kirfel G, Herzog V. (2004). Simultaneous cell death and desquamation of the embryonic diffusion barrier during epidermal development. *Exp. Cell Res.* 299:415-426. doi:10.1016/j.yexcr.2004.06.009.
31. Science Learning Hub – The University of Waikato. (2007). *Feathers and Flight*. Retrieved from https://www.sciencelearn.org.nz/resources/308-feathers-and-flight
32. Steinert P, Mack J, Korge B, Gan SQ, Haynes S, Steven A. (1991). Glycine Loops in Proteins: their occurrence in certain intermediate filament chains, loricrins and single-stranded RNA binding protiens. *Int. J. Biol. Macromol.* 13(3):130-139. doi: https://doi.org/10.1016/0141-8130(91)90037-U.
33. Strasser B et al. (2014). Evolutionary Origin and Diversification of epidermal barrier proteins in amniotes. *Mol Biol Evol*. 31(12): 3194-3205. doi: 10.1093/molbev/msu251.

34. Strasser B, Miltz V, Hermann M, Tschachler E, Eckhart L. (2015). Convergent evolution of cysteine-rich-proteins in feathers and hair. *BMC Evol Biol*. 15:82. doi: https://doi.org/10.1186/s12862-015-0360-y.

35. Veltri A, Lang C, Lein WH. (2017). Concise Review: Wnt Signaling Pathways in Skin Development and Epidermal Stem Cells. *Stem Cells.* 36:22-35. http://dx.doi.org/10.1002/stem.2723.

36. Yu M, Wu P, Widelitz RB, Chuong CM. (2002). The morphogenesis of feathers. *Nature*. 420(6913):308-312. doi:10.1038/nature01196.

37. Wang S, Yang Z, Gong G, Wang J, Wu J, Yang S, Jiang L. (2016). Icephobicity of penguins Spheniscus humboldti and and artificial replica of penguin feather with air-infused hierarchical rough structures. *J. Phys. Chem.* 120. 15923-15929. doi:10.1021/acs.jpcc.5b12298.

CHAPTER 2

IDENTIFICATION AND CHARACTERIZATION OF THE AVIAN EPIDERMAL DIFFERENTIATION COMPLEX LOCUS

2.1 Abstract

The adaptation of novel epidermal appendages such as feathers and hair has played a key role in the evolution of amniotes into different terrestrial lifestyles and environments. They serve several purposes such as thermoregulation, collecting food as well as mating across different amniotes. These appendages form as the result of a tightly regulated spatiotemporal process which involves several elements (Alibardi et al. 2015). In mammals, several of the genetic elements involved in the development as well as the mechanical resilience of hair are found at a specific genetic locus known as the Epidermal Differentiation Complex (EDC). Recently, a locus homologous to the mammalian EDC has been identified in several archosaurian species including the chicken, anole lizard, snakes, turtles, and crocodilians. In order to better characterize the avian EDC, we screened the genomes 48 diverse bird species for EDC genes. We demonstrate that the EDC is conserved in birds, despite being difficult to identify due to several factors including its' complex evolutionary history. Furthermore, our results support the hypothesis for evolution of the avian EDC from a single or small number of ancestral genes. We also provide support for the theory that avian genomes contain "Dark" DNA, areas with high G/C contents which can impair abilities to sequence and characterize them.

2.2 Introduction

The adaptation of unique integumentary structures has played a major role in the evolution of reptiles, birds and mammals (Holthaus et al. 2015). The ability of the embryonic epidermis to form discrete cell lineages capable of producing major structural proteins was key in the evolution of scales, feathers and hair respectively (Sawyer and Knapp 2003). These diverse appendages are all made up of cells which have undergone a specialized mode of programmed cell death called cornification, which confers many of their mechanically resilient properties (Eckhart et al. 2013). In these cells, the plasma membrane has been replaced by an outer Cornified envelope made up of several covalently linked protein elements (Candi et al. 2005).

Previous studies have found that the evolution of several of the structural proteins in the outermost epidermal layers of mammals, which are involved in the development of these epidermal appendages was driven by the diversification of several genes located at a specific genomic locus known as the Epidermal Differentiation Complex (EDC) (Kypriotou et al. 2012). The EDC is thought to have originated in an early amniote ancestor and was critical for the adaptation of a mechanically resilient physical barrier which protected and limited water loss to the environment (Strasser et al. 2014). The expansion of EDC gene families in birds, reptiles and mammals respectively has resulted in the diversity observed across their epidermal appendages such as hair and feathers. These appendages are adapted to fulfill specific functions such as thermoregulation, camouflage and protection against the environment.

Birds specifically exhibit a large amount of physical variation in their epidermal appendages such as feathers, which correlate with the wide range of environments they inhabit and their diverse lifestyles. Feather morphogenesis is a complex process that is heavily reliant on a strict spatiotemporal regulation of gene expression as well as assembly of the protein components into mature structures (Alibardi et al. 2016). The feather formation process entails several interactions between the epidermis and the underlying dermis which ultimately results in the complex hierarchical branching structure observed in feathers. Distinct cell populations of the embryonic epidermal layers differentiate into the barb and barbule cells which fill with several structural proteins including β-keratins, the primary component of mature barbs and barbules, and then fuse together in a unique process resulting in the mature feather (Shames et al. 1993, Sawyer et al. 2000, Alibardi 2003, Greenwold et al. 2014).

The EDC has been extensively characterized in humans and contains genes which are involved in early Cornified Envelope (CE) formation and the mature structure of the epidermis and its' appendages via processes such as disulfide bonding and translgutamination (Hynes and Destree 1977, Steinert et al. 1991, Robinson et al. 1997). Previous studies have identified homologous EDC loci in birds, crocodilians, testudines, lizards and snakes which contain genes of similar exon-intron organization to those of the mammalian EDC, contain amino acid residues associated with epidermal differentiation processes and are expressed in several epidermal tissues throughout development as well as in mature epidermal appendages (Strasser et al. 2014, Strasser et al. 2015, Holthaus et al. 2015, Alibardi et al. 2016, Holthaus et al. 2018, Holthaus et al. 2018, Lachner et al. 2019). The EDC of the chicken was found to contain several unique EDC genes, as well as a cluster of β-keratins containing members of all β-keratin subfamilies of claw, feather, scale and keratinocyte (Greenwold et al. 2014). Studies have provided evidence that translocation of these β-keratins to additional loci outside of the EDC, as well as the respective expansion and diversification of their subfamilies has played a major role in evolution of the feather and the adaptation of birds into multiple ecological niches (Greenwold et al. 2014).

Further studies focusing on specific avian EDC genes such as epidermal differentiation cysteine rich protein (EDCRP) and epidermal differentiation protein with an MTF motif and rich in Histidine (EDMTFH) have found that some EDC genes are conserved across diverse bird species, however there was significant variation observed. For example, the sequences and total number of the repetitive units comprising EDCRP varied across closely related species (Alibardi et al. 2015). Moreover, the respective amino acid compositions of respective EDC genes, such as EDMTFH, also varied significantly (Alibardi et al. 2016). While these studies did identify that specific EDC, genes are conserved across a small subset of avian species, they did not characterize the overall conservation of the EDC across birds.

In order to further analyze the role diversification of EDC genes has played in the evolution of avian epidermal appendages, as well as analyze the evolutionary history and expansion of specific EDC gene families, the EDC loci must be characterized across a larger sample of diverse bird species. Recently, as part of a larger coordinated effort to sequence the genomes of several diverse bird species, the genomes of 48 phylogenetical diverse bird species were sequenced and made available (Jarvis et al. 2014). Here we identify and characterize the EDC loci of these 48 bird species in order to better understand the role it has played in the evolution of the feather and the adaptation of birds into multiple ecological niches. We also investigate repetitive nature and genetic variation across the sequences of identified avian EDC genes. Our results support the hypothesis that the avian EDC evolved from a single or small number of ancestral EDC genes. Furthermore, we provide evidence that avian EDC genes contains biased GC nucleotide contents which can lead to problems with genomic library preparation and blast algorithms resulting in the failure to identify some EDC genes which are indeed present.

2.3 Methods

2.3.1 Identification of Avian EDC genes

All genomes were downloaded from NCBI genomic databases in Fasta formats (supplemental table 2.1). Avian EDC genes were identified using NCBI Blast+ (Altshcul et al. 1990, Gish and States 1993). Amino acid sequences of avian EDC genes identified by Strasser et al. (2014) as well as those of humans (*Homo sapiens*) and the green anole lizard (*Anolis carolinensis*) were added to the query file and *tblastn* searches were performed. Since initial searches using a cutoff e-value and blast score did not return significant results, we reduced the specificity of our searches by adjusting the e-value to 0.1. Furthermore, in order to improve results, any identified amino acid sequences of avian EDC genes were added to the original query file and reciprocal blast searches were performed.

Candidate sequences identified by blast searches were extracted as nucleotide fasta files and translated into amino acids using ExPasy Translate online analysis tool (Gasteiger et al. 2003). Identified sequences were aligned with chicken EDC genes as well as additional identified avian EDC genes using ClustalW (Thompson et al. 1997). Several sequences were only partially identified or contained unknown nucleotides (NNNs) which could not be translated. Genes which were only partially identified were not considered missing if at least one highly conserved sequence element was identified, and the genes' orientation and genomic position corresponded with its location in other avian genomes. Partially identified and genes which contained NNNs exceeding 15% of the coding sequence or disrupting a start or stop codon were excluded form all future analysis. A list of all complete and partially identified avian genes and their qualities investigated in this work can be found in supplemental table 2.3.

The genomic organizations of avian EDCs were assembled using the chicken as a model (Strasser et al. 2014). Scaffolds were placed in linear organization by arranging the blast extraction positions of identified genes. In some cases, the N-terminal region of a gene was contained on the end of one genomic scaffold and the C-terminal region fond on the next. If at least 3 genes were present on a genomic scaffold, their linear organization could be discerned. The chromosomal orientation of EDC genes as well as the genetic rearrangements in the Turkey and Zebra Finch were also identified using this method. The species in figure 2.1 were selected because they represent phylogenetically diverse groups of birds as well as several different lifestyles. Their EDC regions were also contained on less that 3 genomic scaffolds increasing the accuracy of characterizing the linear organization of the EDC as well as determining if missing genes were not identified due to technical issues or truly missing.

In supplemental table 2.2, a gene was considered not found if no elements of the gene were identified at all by blast, or by manual screening of its’ suspected locus. Manual screening for genes was done in combination with assembly of genomic organization by extracting large genomic regions between identified genes and manually screening the translations for EDC gene sequence. This method allowed us to identify several complete and partial avian EDC genes which were not found by blast algorithms and was able to confirm the conserved linear organization of the avian EDC.

Gene sizes in supplemental table 2.3 are represented as the number of amino acids in the second coding exon of EDC genes and were calculated using ExPasy ProtParam online analysis tool (Gasteiger et al. 2005). All genes containing length values in supplemental table 2.2 were identified and contained a start and stop codon and no more than 15% unknown residues (NNNs). Genes that did contain NNNs in this table are denoted with (XX) and genes which contained a resolved frameshift with (fs) following their length values respectively. 6 genes in supplemental table 2.2 contained more than the allowed value of NNNs and were excluded from analysis and are denoted with X, however partial evidence for these genes was identified. Genes that were not found are denoted with n/a. The species of Zebra Finch (*Taeniopygia guttata*), Ground tit (*Pseudopodoces humilis*), Budgerigar (*Melopsittacus undulates*), Bald Eagle (*Haliaeetus leucocephalus*), Adélie Penguin (*Pygoscelis adeliae*), Chicken (*Gallus gallus*) and Turkey (*Meleagris gallopavo*) were selected to represent a phylogenetically and physically diverse subset of birds. These species also contained fewer missing and incomplete EDC genes relative to others.

2.3.2 GC Content Calculation

GC content of avian EDC genes was done using a standard online GC content calculator (Endmemo.com/bio/gc.php). GC content calculations are based on the second coding exon only of EDC genes. Scatter plot of EDC GC contents was done using the calculated GC content as a percent on the y-axis and the average length of all GC stretches present in the gene on the x-axis. A GC stretch was defined as an undisrupted sequence of at least three consecutive G or C nucleotide residues. The average length of GC stretches was calculated by dividing the number of identified GC stretches by the total number of GC residues present within those stretches. Only complete avian EDC genes containing no unknown nucleotides were used this analysis. 36 genes were selected from phylogenetically diverse species including the Zebra Finch (Passeriformes), Chicken (Galliformes) and the Emperor Penguin (Sphenisciformes) and are listed in supplemental table 2.4.

2.4 Results

2.4.1 The EDC is Conserved in Birds

Strasser et al. 2014 identified 30 genes within the chicken EDC with the exon-intron organization identical to that of human EDC genes called Simple EDC genes (SEDCs). In order to determine if the organization and genes of the chicken EDC identified by Strasser et al. 2014 was conserved across all birds, we analyzed the genomes of 48 diverse bird species listed in supplemental table 2.1 using Blast+ using the amino acid sequences of chicken, anole lizard and human EDC genes as queries (supplemental table 2.1). Initial Blast searches did not return results for several avian EDC genes across different species. Identified EDC genes were added to the original query file and reciprocal blast searches were performed.

Our results demonstrate that the EDC locus is present and conserved in all birds investigated. We confirmed the conservation of all avian SEDC genes identified by Strasser et al. (2014) except for EDGH, for which no complete ORS could be located outside of the chicken. Interestingly, the terminal sequence of EDGH was identified in most species, however no complete ORF with both start and stop codons were identified. For this reason, EDGH was excluded from our analyses. Supplemental table 2.2 details the conservation of the remaining 29 EDC genes identified by Strasser et al. (2014) across the 48 bird species analyzed (supplemental table 2.2). The gene EDCH1 was only identified in the chicken and the turkey indicating it may be Galliforme-specific. We identified additional EDC genes in 4 species, the chicken, hummingbird, common cuckoo, and the zebra finch. In the Zebra Finch we identified 2 additional EDCH genes and in Anna's Hummingbird, a single additional copy of the EDCH gene (supplemental table 2.2). In the Common Cuckoo we identified 2 additional copies of the Epidermal differentiation proteins rich in Aromatic amino acids containing MTF motifs (EDAA/EDMTF). We also identified an additional EDMTF in the chicken annotated EDMTF5. All additionally identified genes were duplicates. EDMTF2 was only identified in 5 of the 48 species investigated, including the chicken, and the previously characterized EDMTFH gene was not identified in any Passeriformes except for the Golden-Collared Manakin (*Manacus vitellinus*). Many of the genes identified contained sequence artifacts such as frameshifts or unknown nucleotides which impaired our ability to identify and further analyze them. The organization and completion status of each all EDC genes across the 48 species investigated and all values' meanings are listed supplemental table 2.3 in the table key.

We found that the linear organization of the EDC in the chicken is also conserved in birds, with S100A genes located on the borders, and the central region contains genes which consist of a 5'-terminal noncoding exon and a second exon which comprises the entire coding region (Figure 2.1). We identified 2 possible genetic rearrangements in the zebra finch and turkey respectively, however the chromosomal orientation of all genes was conserved (Figure 2.1). We outline the potential rearrangement in the turkey EDC in

Figure 2.2. All other avian EDC genes and gene families were conserved in syntenic locations across all species, however as mentioned previously, there is lineage-specific variation present in their exact copy-numbers (supplemental table 2.2). Unlike β-keratins which have expanded to other parts of the avian genome, all EDC genes identified were localized within the borders of the EDC.

Previous studies found that the coding sequence of the conserved EDCRP gene is made up of alternating repetitive units that vary among different species, however contain highly conserved cysteine residues (Alibardi et al 2015). Our results demonstrate that several avian EDC genes are also made up of repetitive units, which can vary extensively across species resulting in a large amount of genetic variation across conserved EDC genes. Specifically, there are large differences in gene size corresponding to the numbers and sizes of the respective repetitive units which make of genes such as avian Loricrins, EDCRP, EDQrep, EDQM, and EDDM (supplemental table 2.4).

Additionally, the amino acid contents of avian EDC genes are rich in amino acids associated with epidermal development and the structure of mechanically resilient appendages such as cysteine, serine and tyrosine. These amino acid residues are often associated with structural functions, specifically in extracellular matrix assembly on a cellular level and in the overall keratinization of epidermal appendages. Moreover, these residues are often more highly conserved than other residues indicating they are important in function.

2.4.2 Support for the presence of "Dark DNA"

Hron et al. (2015) and Bornelöv et al. (2017) provided evidence that a subset of genes previously reported as missing in birds but present in most other vertebrate lineages were indeed present in the chicken genome. They attribute the underrepresentation of these genes in genomic databases to their strongly biased G/C patterns, with all the genes exceeding 60% G/C content due to studies which have found that excessive G/C contents present technical problems with genomic library preparation and in database searching algorithms (Hron et al. 2015, Bornelöv et al. 2017). In our current study, there were several species where some EDC genes were not identified (supplemental table 2.2 – Genes not found). In species where these genes were identified, their sequences often contain biased G/C contents of above 60%, as well as the presence of several G/C nucleotide runs (supplemental table 2.4, Figure 2.3). Specifically, EDQM1, EDQM2 or both were not identified in 25 different species however these genes contain a highly repetitive sequence elements as well as high G/C contents in species where they were identified (supplemental table 2.5). As mentioned previously, except for EDMTFH which was not identified in Passerine birds, we did not find any evidence among closely related species for loss of any EDC genes across an entire clade. For example, EDQM genes were not identified in the Northern Fulmar or Great Cormorant, however they were identified in both the Crested ibis as well as the Adélie Penguin which are closely related. It is possible that the high G/C contents of several avian genes, as well as their repetitive nature directly correlates with the failure to identify several avian EDC genes across closely related species, as well as the large number of incomplete and partial sequences identified.

These results demonstrate that the EDC region as well as its genomic organization is conserved across birds. They also show that much like EDCRP, several EDC genes are made up of highly repetitive units. Differences in the sequences of these units has led to a large amount of genetic variation across otherwise conserved genes, even among closely related species. Moreover, we found there are several conserved gene families in the avian EDC. These families include EDQM1/2, EDCH1/2/3/4, LOR1/2/3, and EDAA/EDMTF1/2/3/4/H. The EDQM and EDCH groups consist of duplicate genes which have not diverged in sequence. The Loricrins and EDAA/EDMTF groups consist of both duplicate copies as well as more divergent copies which may have taken on different functions (supplemental table 2.2) Loricrins and the EDAA/EDMTF gene families will be investigated and explored in chapters 3 and 4 respectively. Finally, we provide evidence that avian EDC genes contain biased G/C contents characteristic of "Dark" DNA, which has been shown to cause problems during genomic library preparation resulting in genes falsely being reported as missing.

2.5 Discussion

The results of this study demonstrate that the EDC is conserved across 48 diverse species of birds and characterizes its organization as a conserved cluster of genes. These results also confirm that the genomic organization and architecture of the chicken EDC presented by Strasser et al. (2014) is also conserved across birds. We did not identify any avian EDC genes outside of the borders of the EDC locus; however, we did identify 2 possible genomic rearrangements in the Turkey and Zebra Finch respectively, however all genes were still localized within the borders of the EDC and their chromosomal orientation was also conserved, indicating that it may be important in gene transcription.

While no avian EDC genes were identified outside of the EDC locus, several EDC genes have evolved into conserved multigene families via tandem duplications such as EDCH, Loricrins, and EDAA/EDMTF genes. This supports the hypothesis presented by Strasser et al. (2014) that the EDC has evolved from a single or small number of ancestral EDC genes primarily through tandem gene duplications followed by diversification and ultimate neofunctionalization. Previous studies have found that members of all 4 β-keratin subfamilies are conserved in a cluster within the avian EDC, but the exact number of genes and proportions of the specific subfamilies varies significantly across different species (Greenwold et al. 2014). While feather β-keratins specifically have translocated and expanded in other parts of the genome and this is thought to have played a key role in the diversification of feathers, the presence of all 4 subfamilies of β-keratins within the EDC indicates that the transposition of duplicated genes on the same locus is adequate to induce a relatively high level of sequence divergence and possible neofunctionalization (Greenwold et al. 2014).

It has also been shown that like β-keratins, individual members of the EDC gene groups such as EDAA/EDMTFs and loricrins are capable of differential expression in developing epidermal tissues such as feathers and scales (Strasser et al. 2014, Alibardi et al. 2016). Our results this indicate that avian EDC genes are excellent candidates for gene duplication and subfunctionalization, which has been shown to ultimately lead to neofunctionalization (Rastogi and Liberales 2005). It is possible that the expansion of

EDCH and EDAA/EDMTF gene families and the capability for differential expression throughout development has also played a key role in the evolution of feathers.

While several genes of the EDC are evolving via whole-gene duplication events, other avian EDC genes have undergone several intragenic duplications of specific sequence elements resulting in genes composed of repetitive units which substantially in size and number across different species. Strasser et al. (2015) analyzed the conserved Epidermal Differentiation Protein Rich in Cysteine (EDCRP) and found that it was composed of a highly variable number of repetitive units, which resulted in significant variation the size of the gene across different species. We found that several other avian EDC genes, Loricrins, EDQMs, EDQrep and EDDM are also composed of highly variable repetitive units. These genes are rich in amino acids such as cysteine, glycine and serine, all residues which are important in development and structure of epidermal appendages. If these genes do play structural roles in avian feathers, differences in their sizes due to variation in the number of repetitive units making them up, or specific amino acid composition of those units, could ultimately result in differences in physical properties such as elasticity and flexibility of feathers. This could have major implications on major aspects of avian lifestyle such as flying, hunting and habitat selection.

The presence of these highly variable repetitive units within EDC genes also likely plays a direct role in the difficulty identifying them. It is known that highly repetitive DNA sequences present several technical challenges for sequence alignment and assembly programs (Treangen and Salzberg 2011). Strasser et al. (2015) identified an artificial frameshift in the Zebra Finch EDCRP sequence, which was resolved by direct sequencing. We encountered several frameshifts and unknown nucleotides in the ORFs of EDC genes which impaired our ability to further analyze them. It is likely that many of these frameshifts and unknown sequences would be resolved due to direct sequencing, however given the high level of inter and intragenic duplications within the avian EDC, it is also possible that several duplicate genes do contain frameshift mutations.

This study also provides support that avian genomes contain areas of “Dark” DNA which results in several genes being underrepresented in genomic databases. Dark DNAs are sequences which contain highly biased G/C nucleotide contents as well as many G/C stretches. These biased G/C contents present problems with genomic library preparation which results in conserved genes being reported as missing (Hron et al. 2015). Already, several important genes which were originally reported as missing in birds, have now been identified and found to contain highly biased G/C contents (Bornelov et al. 2017). Our results show that of many of the avian EDC genes identified across diverse species such as the chicken, penguin, zebra finch and ground tit, contain biased G/C contents, characteristic of genes originally reported as missing by previous studies. Similarly, studies on the conservation of avian β-keratins have suggested that the very large discrepancy observed between numbers of identified genes could in part be related to problems associated with genomic library preparation as well as sequence alignment and identification algorithms (Greenwold et al. 2015). Except for EDGH, for which we did not identify a conserved start codon outside of the chicken, we did not identify any evidence for a conserved frameshift in avian EDC genes resulting in loss of the ORF. Given the overall conservation of the EDC across phylogenetically diverse species, it is more likely that many of these genes are indeed present and are missing from genomic databases due to technical issues, than they have been lost in individual lineages.

In this chapter we identified and characterized the avian EDC as a chromosomal gene cluster across a phylogenetically diverse group of birds. We found that while the EDC is conserved across birds, there is significant interspecific and intragenic variation observed. Our results support the hypothesis presented by Strasser et al. (2014) for evolution of the EDC from a single or small number of ancestral genes. The presence of several multigene families within the avian EDC indicates that much like β-keratins, the duplication and divergence of avian EDC genes has played a role in the adaptation of avian species to diverse lifestyles and habitats. We suggest the failure to identify several avian EDC genes, as well as the presence of unknown nucleotides and artificial frameshifts in many species is in part due to their repetitive nature and high G/C contents. It is likely that the direct sequencing would resolve several of these issues.

2.6 Figures

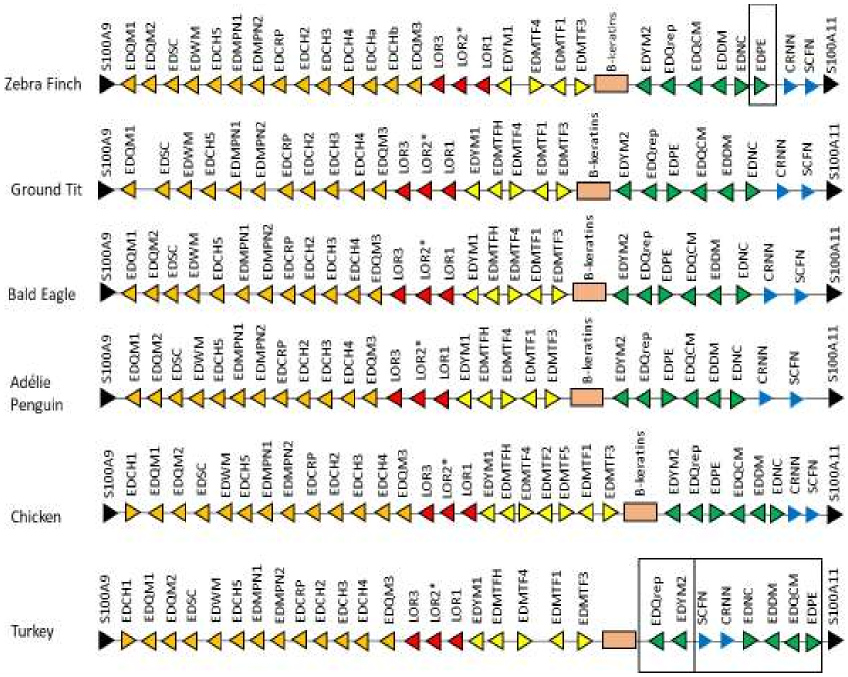


Figure 2.1 : Organization of avian EDC across diverse species. This figure demonstrates the conservation of the avian EDC across 6 phylogenetically diverse species. The color of the genes correspond to those in Strasser et al. (2014). The boxes around Zebra Finch EDPE and in the turkey represent possible inversion events which have resulted in the variation in organization, however none of the genes orientation was changed. We identified an extra copy of EDMTF in the chicken which we annotated EDMTF5.

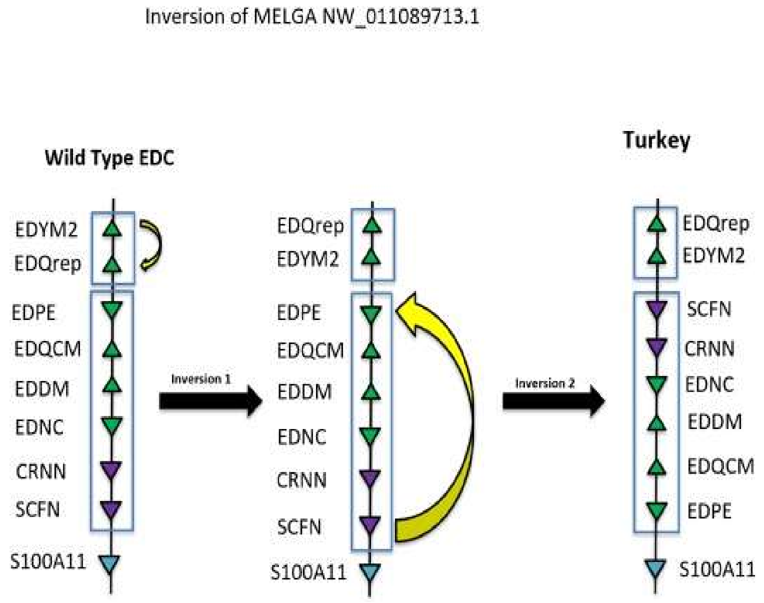


Figure 2.2: Possible Inversion Event in the EDC of the Turkey. The turkey and Zebra Finch were the only avian species where the organization of some EDC genes varied from that of the chicken. We present the possible inversion event in the turkey which may have led to its observed architecture. None of the genes' orientations were affected by this potential inversion event.

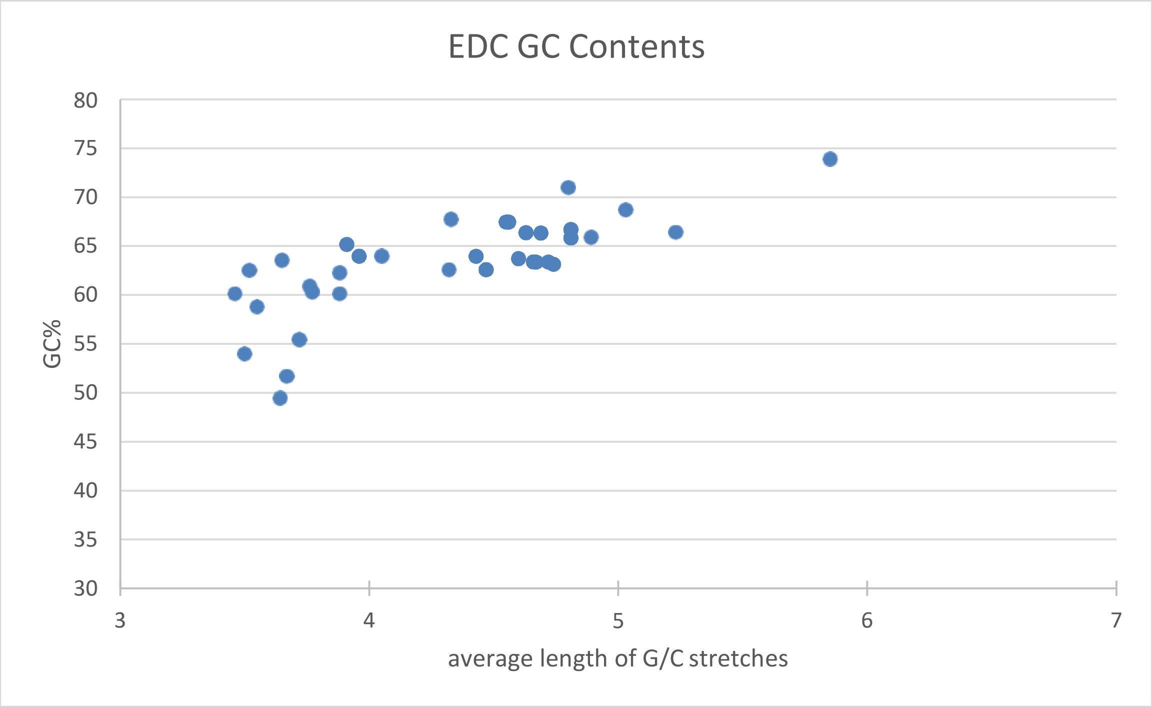


Figure 2.3 : Scatter plot of avian EDC G/C contents. Y-axis represents overall percentage of G/C residues making up the coding sequence. X-axis indicates the average length of the G/C stretches present. G/C stretches were defined as runs of 3 or more uninterrupted G/C nucleotides. The scatter plot demonstrates that the majority of Avian EDC genes identified contained highly biased G/C contents. These biased G/C contents meet the criteria presented by Hron et al. (2015) of not being identified due to technical issues with genomic library preparation and search algorithms.

2.7 References

1. Alibardi L, Valle LD, Nardi A, Toni M. (2009). Evolution of hard proteins in the sauropsid integument in relation to the cornification of skin derivatives in amniotes. *J. Anat.* 214. 560-586. doi:10.1111/j.1469-7580.2009.01045.

2. Alibardi L, Holthaus KB, Sukseree S, Hermann M, Tschaler E, Eckhart L. (2016). Immunolocalization of a Histidine-Rich Epidermal Differentiation Protein in the Chicken Supports the Hypothesis of an Evolutionary Developmental Link between the Embryonic Subperiderm and Feather Barbs and Barbules. *PLOS One* 11(12): e0167789. doi:10.1371/journal.pone.0167789.

3. Alibardi L. (2017). Review: cornification, morphogenesis and evolution of feathers. *Protoplasma*. 254(3):1259-1281. doi: https://doi.org/10.1007/s00709-016-1019-2

4. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

5. Bornelöv S, Seroussi E, Yosefi S, Pendavis K, Curgess SC, Grabherr M, Friedman_einat M, Andersson L. (2017). Correspondence on Levell et al.: Identification of chicken genes previously assumed to be evolutionarily lost. *Genome Biology.* 18: 1 12. Doi:10.1186s13059-017-1231-1.

6. Eckhart L, Lippens S, Tschaler E, Declercq W. (2013). Cell death by cornification. *BBA-Mol. Cell Res.* 1833(12): 3471-3480. doi: https://doi.org/10.1016/j.bbamcr.2013.06.010.

7. Gilbert SF. (2014). Developmental Biology, 10th edition. Pgs. 79-83.

8. Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D., Bairoch A. (2003) *ExPASy: the proteomics server for in-depth protein knowledge and analysis* Nucleic Acids Res. 31:3784-3788

9. Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. (2002). *Protein Identification and Analysis Tools on the ExPASy Server;*

(In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press pp. 571-607

10. Gish, W. & States, D.J. (1993) "Identification of protein coding regions by database similarity search." Nature Genet. 3:266-272.

11. Greenwold MJ, Bao W, Jarvis ED, Hu H, Li C, Gilbert MTP, Zhang G, Sawyer RH. (2014). Dynamic Evolution of the alpha and beta keratins has accompanied integument diversification and the adaptation of birds into novel lifestyles. *BMC Evo. Biol.* 14:249. doi: 10.1186/s12862-014-0249-1.

12. Holthaus KB, Strasser B, Sipos W, Schmidt HA, Mlitz V, Sukseree S, Weissenbacher A, Tschaler E, Alibardi L, Eckhart L. (2015). Comparative genomics Identifies Epidermal Proteins Associated with the Evolution of the Turtle Shell. *Mol. Biol. Evol.* 33(3):726-737. doi: 10.1093/molbev/msv265.

13. Holthaus KB et al. (2018). Comparative analysis of epidermal differentiation genes of crocodilians suggests new models for the evolutionary origin of avian feather proteins. *Gen. Biol. Evol.* 10(2): 694-704. doi: 10.1093/gbe/evy035.

14. Hron t, Pajer P, Paces J, Bartunek P, Elleder D. (2015). Hidden genes in birds. *Genome Biology*. 16:164. Doi:10.1186/s13059-015-0724-z.

15. Hynes RO, Destree A. (1977). Extensive disulfide bonding at the mammalian cell surface. *Proc. Natl. Acad. Sci.* Vol. 74, No. 7: 2855-2859.

16. Jarvis ED et al. (2014). Whole genome analyses resolve the early branches in the tree of life of modern birds. *Science.* 346(6215):1320-1331. doi: 10.1126/science.1253451.

17. Kypriotou M, Huber M, Hohl D. (2012). The human epidermal differentiation complex: cornified envelope precursors, S100 proteins and the 'fused genes' family. *Exp. Dermatol.* 21(9):643-649. doi: https://doi.org/10.1111/j.1600-0625.2012.01472.x.

18. Lachner J, Ehrlich F, Mlitz V, Harmann M, ALibardi L, Tschaler E, Eckhart L. (2019). Immunolocalization of phylogenetic profiling of the feather protein with the highest cysteine content. *Protoplasma*. doi:https://doi.org/10.1007/s00709-019-01381-3.

19. Li, C., Zhang, Y., Li, J., Kong, L., Hu, H., Pan, H., Xu, L., Deng, Y., Li, Q., Jin, L., Yu, H., Chen, Y., Liu, B., Yang, L., Liu, S., Zhang, Y., Lang, Y., Xia, J., He, W., Shi, Q., … Zhang, G. (2014). Two Antarctic penguin genomes reveal insights into their evolutionary history and molecular changes related to the Antarctic environment. *GigaScience*, *3*(1), 27. https://doi.org/10.1186/2047-217X-3-27

20. Robinson NA, Lapic S, Welter JF, Eckert RL. (1997). S100A11, S100A10, Annexin I, Desmosomal Protiens, Small Proline-rich Protiens, Plasminogen Activator Inhibitor-2, and Involucrin are components of the cornified envelope of cultured human epidermal keratinocytes. *J. Biol. Chem.* Vol 272, No. 18: 12035-12046.

21. Sawyer RH, Glenn T, French JO, Mays B, Shames RB, Barnes GL, Rhoses W, Ishikawa Y. (2000). The expression of Beta (β) keratins in the epidermal appendages of reptiles and birds. *American Zoologist* Vol. 40 (4), 530-539. https://doi.org/10.1093/icb/40.4.530.

22. Sawyer RH, Knapp LW. (2003). Avian skin development and the evolutionary origin of feathers. *J. Exp. Zool. MDE.* 298B:57-72. doi:10.1002/jez.b.00026.

23. Shames RB, Knapp LW, Barnes GL, Sawyer RH. (1993). Region-Specific Patterns of Beta Keratin Expression During Avian Skin Development. *Developmental Dynamics* 196:283-290.

24. Steinert P, Mack J, Korge B, Gan SQ, Haynes S, Steven A. (1991). Glycine Loops in Proteins: their occurrence in certain intermediate filament chains, loricrins and single-stranded RNA binding protiens. *Int. J. Biol. Macromol.* 13(3):130-139. doi: https://doi.org/10.1016/0141-8130(91)90037-U.

25. Strasser B et al. (2014). Evolutionary Origin and Diversification of epidermal barrier proteins in amniotes. *Mol Biol Evol*. 31(12): 3194-3205. doi: 10.1093/molbev/msu251.

26. Strasser B, Miltz V, Hermann M, Tschachler E, Eckhart L. (2015). Convergent evolution of cysteine-rich-proteins in feathers and hair. *BMC Evol Biol*. 15:82. doi: https://doi.org/10.1186/s12862-015-0360-y.

27. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25(24): 4876-4882.

28. Treangen TJ. Salzberg S. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*; 13(1): 36-46. doi:10.1038/nrg3117.

CHAPTER 3

COMPLEX GENE LOSS AND DUPLICATION EVENTS HAVE FACILITATED THE EVOLUTION OF MULTIPLE LORICRIN GENES IN DIVERSE BIRD SPECIES[1]

[1] Davis A, Greenwold MJ, Sawyer RH. 2018. Complex gene loss and duplication events have facilitated the evolution of multiple loricrin genes in diverse bird species. Oxford Genome Biology and Evolution, Vol 11 (3); 984-1001

3.1 Abstract

The evolution of a mechanically-resilient epidermis was a key adaptation in the transition of amniotes to a fully terrestrial lifestyle. Skin appendages usually form via a specialized type of programmed cell death (PCD) known as cornification which is characterized by the formation of an insoluble cornified envelope (CE). Many of the substrates of cornification are encoded for by linked genes located at a conserved genetic locus known as the epidermal differentiation complex (EDC). Loricrin is the main protein component of the mammalian CE and is encoded for by a gene located within the EDC. Recently, genes resembling mammalian loricrin, along with several other proteins most likely-involved in CE formation, have been identified within the EDC of birds and other reptiles. To better understand the evolution and function of loricrin in birds, we screened the genomes of 50 avian species and 3 crocodilians to characterize their EDC regions. We found that loricrin is present within the EDC of all species investigated, and that 3 loricrin genes were present in birds. Phylogenetic and molecular evolution analyses found evidence that gene deletions and duplications as well as concerted evolution has shaped the evolution of avian loricrins. Our results suggest a complex evolutionary history of avian loricrins which has accompanied the evolution of bird species with diverse morphologies and lifestyles.

3.2 Introduction

The major event that facilitated the adaptation of amniotes to a fully terrestrial lifestyle was the evolution of a mechanically resilient epidermis which provided a protective barrier and limited water loss to the environment (Chuong et al. 2002). The development of the amniotic epidermis is largely characterized by the cornification of keratinocytes, which represent the main cellular component of the epidermis. (Candi et al. 2005). Cornification of keratinocytes is a multi-step process that ultimately results in the formation of the terminal layer of the epidermis known as the stratum corneum (SC). The SC confers mechanically resilient properties to the epidermis and its appendages such as hair, nails, and feathers which have aided amniotes in diversifying and inhabiting nearly every ecological niche on the planet, as well as adapt to changing environments, resource availabilities, and climate conditions (Pierard et al. 2000, Strasser et al. 2014).

The stratum corneum is composed of terminally differentiated keratinocytes, or corneocytes, in which the plasma membrane has been replaced by an insoluble protein structure known as the Cornified Envelope (CE). The CE provides mechanically resilient properties such as flexibility and elasticity to the epidermis and its appendages (Candi et al. 2005, Eckhart et al. 2013). The process of CE formation requires strict spatiotemporal regulation of the expression of several different genes and protein substrates (Alibardi et al. 2016). Many of the genes which encode protein substrates involved in CE assembly and structure in mammals are clustered on the human chromosomal region 1q21, which has been termed the Epidermal Differentiation Complex (EDC) (Kypriotou et al. 2012). The EDC of mammals contains genes such as filaggrin, involucrin and loricrin which are expressed during CE assembly and are critical for proper function of the epidermis and its appendages (Hohl et al. 1990, Robinson et al. 1997, Chuong et al. 2002, Candi et al. 2005). Recently, a genetic locus homologous to the mammalian EDC has been characterized in the chicken and anole lizard, which contains genes of similar exon-intron organization, amino acid composition and expression profiles (Strasser et al. 2014, Strasser et al. 2015). Since then, the EDC locus has been identified in crocodilians, snakes and turtles that also contain genes characteristic of being involved in the CE assembly. This indicates the EDC locus was present before the divergence of birds and reptiles from mammals. There is no current evidence of an EDC in the genomes of ray-finned fishes (*Takifugu rubripes*), amphibians (*Xenopus tropicalis, x. laevis*), or the coelacanth (*Latimeria chalumnae*) supporting the hypothesis that the evolution of the EDC coincided with the adaptation of amniotes to a fully terrestrial lifestyle. (Holthaus et al. 2015, Holthaus et al. 2017).

The main component of the mammalian CE is loricrin, and previous studies have suggested it constitutes 70-85% of the total CE protein content (Hohl et al. 1990, Candi et al. 2005, Eckhart et al. 2013). A more recent study found that while loricrin is a major protein of the CE, they calculated that loricrin has a 11.8-21.5 relative abundance in wild-type mice (Rice et al. 2016). Loricrin is a highly crosslinked structural protein which is extremely rich in glycine as well as polar residues. Studies have found that mutations in loricrin are associated with human skin diseases such as Vohwinkel's syndrome (VS) and progressive symmetric erythrokeratoderma (PSEK) (Ishida-Yamamoto et al. 1998, Candi et al. 2005). In mammals, loricrin is preferentially crosslinked by transglutaminases (TGases) and provides both elasticity as well as mechanical resistance to the CE (Steinert et al. 1991). Mammals possess a single loricrin gene which contains 2 exons with the entire coding sequence contained in the second exon. The coding sequence (CDS) is composed of conserved N and C terminal domains rich in lysine and glutamine separated by 3 central Gly-Ser-Cys rich repeat domains of variable lengths which are interspersed by short Glutamine-rich regions. This central domain is thought to confer some of the mechanically resilient properties to the CE by taking on a specialized conformation known as the Glycine-loop (Gly-Loop) (Hohl et al. 1990, Steinert et al. 1991). Gly-loops form when at least 2 quasi-peptide repeats of the form $x(y)_n$ are arranged in tandem, where x is an aromatic or aliphatic residue, y is usually a polar residue (glycine or serine) and n is the number of polar residues and is highly variable. Sequencing and proteolysis of normal human corneocytes has demonstrated that loricrin is primarily crosslinked to other loricrins via isodipeptide bonds, but loricrin was also found to be crosslinked with Small Proline-rich proteins (filaggrin, and Keratin Intermediate Filaments (KIF). These crosslinked proteins form a matrix referred to as the KIF-matrix-protein complex. Crosslinking of loricrin with the KIF-matrix-protein complex may provide a means of coordinating cellular structure (Robinson et al. 1997, Wang et al. 2000).

Loricrin has been localized to the EDC in the chicken, two turtles, two snakes and the anole lizard, however the number of loricrin genes varied across different groups of organisms. Three loricrin genes were identified in the chicken, two in squamates and only a single loricrin was identified within the EDCs of crocodilians and testudines (Strasser et al. 2014, Holthaus et al. 2015, Holthaus et al. 2017, Holthaus et al. 2018). Furthermore, they found that the three chicken loricrin genes are differentially expressed in the beak, scale, comb, claw, feather and skin of both embryonic and adult individuals (Strasser et al. 2014). Recently, the genomes of several diverse avian species have been sequenced and published allowing researchers to further analyze the conservation and function of avian EDC genes (Strasser et al. 2015, Alibardi et al. 2016). Given the importance of loricrin in the structural properties of the mammalian epidermis as well as loricrins' expression patterns found in the epidermal appendages of the chicken, studies focusing on loricrins in birds and reptiles may provide insight into the formation epidermal appendages (Alibardi 2017). Previous studies have also demonstrated that the avian-specific β-keratins confer very specialized structural properties useful in a variety of applications ranging from applications as low-cost building materials to medicinal uses (Barati et al. 2016). Despite these advances, little is known about how these proteins interact with other epidermal proteins including loricrin (Barati et al. 2017). To gain a better understanding of the evolution of loricrin genes in birds and reptiles, as well as the roles they play in the development of feathers and scales, we used comparative genomics to screen for loricrin genes in 50 phylogenetically diverse species of birds (Cai et al. 2013, Fankl et al. 2013, Jarvis et al. 2014, Zhang et al. 2014).

3.3 Methods

3.3.1 Identification and characterization of the Epidermal Differentiation Complex in Birds and Reptiles

All genomes were downloaded from the NCBI FTP site in fasta format (supplemental table 1). All genomes had been previously assembled as unplaced genomic scaffolds with the exception of the chicken (*Gallus gallus*) and the zebra finch (*Taeniopygia guttate*) which were assembled at the chromosomal level (Jarvis et al. 2012, Zhang et al. 2014, Yang et al. 2015) Blast databases of each genome were created using blast-2.7.1+ *makeblastdb*. Using the *tblastn* command each nucleotide database was screened for EDC genes using the amino acid sequences of EDC genes from Strasser et al. (2014) as queries (Altschul et al. 1990, Altschul et al. 1997, Camacho et al. 2009, Pierard et al. 2000, Holthaus et al. 2015, Holthaus et al. 2017).Potential EDC genes identified by *tblastn* searches were extracted using the *blastdbcmd* command as nucleotide sequences in fasta format. These sequences were then translated using the ExPASy translate online analysis tool, and aligned using ClustalW online analysis tools (Thompson et al. 1997, Jeanmougin et al. 1998, Gasteiger et al. 2003).

The genomic organization of avian EDC loci was predicted by aligning identified EDC genes with their respective positions in the chicken. The linearity of DNA sequences was then used to align various genomic scaffolds to recreate each avian EDC region. Several EDC genes, including loricrins, were often not identified by *tblastn* algorithms, however manual screening of genome sequences often found evidence of loricrin genes.

3.3.2 Phylogenetic Analysis of Loricrins

The loricrin sequences of 15 avian species, 9 mammalian species, 2 crocodilian, 2 testudine, and 3 squamates, which are listed in supplemental table 2, were used to construct Bayesian and maximum-likelihood (ML) phylogenetic trees. These avian species were selected because they each possess 3 loricrin genes with both start and stop codons, had no premature stop codons or frameshift mutations, and less than 70% of their central domain was composed of unknown nucleotides (NNNs). Amino acid alignments of loricrin sequences were done using ClustalW2 (Thompson et al. 1997) local alignment tools and edited using Bioedit software (Hall 1999). Using MEGA7 (Kumar et al. 2016) the substitution matrix PROTGAMMAJTTF (JTT+G) was determined to be the best fit substitution model based on Bayesian Information Criterion (BIC), Akaike Information

Criterion, corrected (AIC_c) and the substitution rate (BIC^{JTT+G}=1299.488, AIC_c^{JTT+G}=839.458). Bayesian analysis was done using Mrbayes-v3.2 tool (Huelsenback and Ronquist 2001, Ronquist and Huelsenbeck 2003). We ran 10,000,000 generations and checked for convergence using the Potential Scale Reduction Factor method (PSRF) (TL:PSRF=1.0 ; alpha: PSRF=1.0) (Gelman and Rubin 1992). ML analysis was performed on the same alignment file using RAxML-v8.2.10 by first using MRE-based bootstrapping until convergence was reached, followed by inferring the best tree produced from generating 1000 thorough ML trees, then mapping the MRE bootstrap values onto the best ML tree (Stamatakis 2014). Generated Bayesian and ML trees were viewed and edited using FigTree-v1.4.3 (Rambaut 2012). Protein sequence alignment (supplementary figure 2) was generated using T-Coffee online analysis tool (Notredame et al. 2000).

3.3.3 Gene Conversion Tests

Gene conversion analysis was done using GENECONV (Sawyer 1989). Loricrin sequences of only 6 phylogenetically diverse avian species (*Struthio camelus, Manacus vitellinus, Chaetura pelagica, Gallus gallus, Haliaeetus leucocephalus,* and *Pseuopodoces humilis*) were used due to GENECONV analysis requiring that no NNNs be present in the sequences.

3.3.4 Prediction and Analysis of Gly-loop domains of avian loricrins

The Gly-loop domains of 6 avian species (*Gallus gallus, Haliaeetus leucocephalus, Chaetura pelagica, Manacus vitellinus, Pseudopodoces humilis, and*

Melopsittacus undulatus) as well as the orca (*Orcinus orca*) were predicted using the $x(y)_n$ motif described by Hohl et al 1990. The number and size of Gly-loops of human and mouse loricrins were calculated using the schematic representations proposed by Steinert et al. 1991. Avian species were selected because they were phylogenetically diverse and possessed complete loricrin sequences. Furthermore, with the exception of the budgerigar (*Melopsittacus undulatus*) (supplemental table 3.1) they had no NNNs in their central domains. The total number of Gly-loops was predicted by counting the number of gly-ser-rich stretches of sequence present in the central domain ($(y)_n$) that were also bordered by either an aromatic or an aliphatic residue (x). Loop sizes were predicted by counting only the number of residues located between aromatic/aliphatic residues which were thought to form gly-loops. The schematic representations of the Gly-loops of chicken LOR3 and LOR1 (figure 4 A and B) are based on the schematic representations of human and mouse loricrins proposed by Steinert et al. 1991 and are not intended to predict specific secondary structure.

3.3.5 Amino Acid Composition and Statistical Analysis of Loricrin

Amino Acid Analysis was performed using avian loricrin sequences classified as complete and which were composed of <15% NNNs's, as well as mammalian, crocodilian, and squamate loricrin sequences. Translated amino acid loricrin sequences were analyzed for amino acid composition using ExPASy ProtParam tool (cite ExPASy). In order to account for the large amount of variation in size observed across loricrin genes, all amino acid analyses were done using the percentage of each amino acid present in the sequence as opposed to the total number of residues. The resulting percentage of each amino acid residue for each loricrin sequence analyzed can be found in supplemental table 4 A and B. These data were used to generate the principal component analysis (PCA) in R (Figure 5) by means of the BiocLite- pcaMethods package by BioConductor. The PCA was done using thing Singular Value Decomposition (SVD) method.

Further amino acid analyses were performed by comparing the percentage of each of the 20 amino acid residues observed across the respective loricrins of each species examined in order to identify significant differences in the amino acid contents of respective amino acid residues. Significance was determined using Analysis of Variance (ANOVA) and Welch's t-test analysis which was performed using Microsoft Excel: Data Analysis ToolPak.

3.4 Results

3.4.1 Loricrin conservation within the EDC across birds and reptiles

In order to establish whether the loricrin genes identified in the chicken and anole lizard by Strasser et al. [3] are conserved across birds and reptiles, we screened the genomes of 2 crocodilian species (*Alligator mississippiensis* and *Crocodylus porosus*) and 50 phylogenetically diverse avian species (supplemental table 1) using the amino acid sequences of the chicken, king cobra, burmese python, chinese soft-shelled turtle, western painted box turtle and the anole lizard EDC genes as BLAST queries (Altschul et al. 1990, Altschul et al. 1997, Camacho et al. 2009, Pierard et al. 2000, Holthaus et al. 2015, Holthaus et al. 2017). Bird genomes searched in this analysis came from the recently sequenced genomes of 48 diverse bird species (Jarvis et al. 2012, Zhang et al., 2014). We also searched the genomes of the ground tit (*Pseudopodoces humilis)* and atlantic canary (*Serinus canaria*) (Cai et al. 2013, Fankl et al. 2013). All genomes were obtained from NCBI and were previously assembled at the scaffold level with the exception of the chicken, zebra finch and turkey which were assembled to the chromosome level. Identified loricrin (LOR) genes were added to the query file and iterative rounds of BLAST searches were performed on the avian genomes.

The results of these BLAST searches confirmed evidence of at least a single copy of loricrin in the two crocodilian species and the 50 bird species (supplemental table 1). When multiple loricrin genes were identified in the bird genomes, we found them to be tandemly arranged in the same orientation and conserved within the EDC between the EDGH and EDYM1 genes (figure 1). We found evidence of only a single loricrin gene in the crocodilian genomes, which is in agreeance with a recent study characterizing the crocodilian EDC (Holthaus et al., 2018). Previous studies found a single loricrin gene in turtles whereas two loricrin genes are present in squamates (figure 1) (Strasser et al. 2014, Holthaus et al. 2014, Holthaus et al. 2017). In birds, evidence of three loricrin genes was identified in 39 of the 50 species examined, however, in many species this region of the EDC (in which loricrins are located) was either incomplete (assembled across multiple scaffolds) or composed almost entirely of unknown nucleotides (NNN's) (supplemental figure 1 A and B). This resulted in only the ground tit (*Pseudopodoces humilis*), bald eagle (*Haliaeetus leucocephalus*) and chicken (*Gallus gallus*) having three uninterrupted, complete loricrin sequences (supplemental table 3.1).

In order to analyze the number of loricrin genes conserved across birds, we narrowed our results by selecting species in which the loricrin containing region of the

EDC (Figure 1B) was assembled on a single scaffold. Twenty-five phylogenetically diverse avian species (supplemental table 3.1) were found to have this portion of the EDC; however, 22 of these species still possessed loricrin sequences containing NNN's. We found evidence suggesting the presence of three loricrin genes in all but one (pigeon) of these 25 species (supplemental table 3.1). The pigeon (*Columbia livia*) was found to have only two loricrins with no evidence of a third loricrin. We did not find evidence suggesting the presence of more than three or less than two loricrin genes in any of these 25 bird species.

3.4.2 Phylogenetic analyses suggest a complex and dynamic evolutionary history of loricrins in birds

Like the mammalian loricrin, avian loricrins are composed of highly conserved N- and C- terminal domains separated by a highly variable glycine rich repeat domain (figure 2 and supplemental figure 2) (Hohl et al. 1990). Likely due to the highly repetitive nature of loricrins, many loricrin genes did not assemble well in the avian genomes and are composed of unknown nucleotides (NNNs) (Milinkovitch et al. 2010, Hron et al. 2015). Therefore, we used specific parameters to screen loricrin genes for inclusion in phylogenetic analyses. Loricrin sequences were considered complete provided that: (1) the N- and C- termini were both present without any NNNs, (2) within the central domain, at least 3 tandemly arranged repeat units are present without NNNs, and (3) no more than 15% of the central domain contained NNNs. Loricrin sequences in compliance with (1) and (2), but contained > 15% but less than 70% NNNs were considered partial sequences. This resulted in 15 avian species having 3 complete or partial loricrin genes which were used in phylogenetic analyses (supplemental table 2, supplemental figure 3, figure 3). In addition to these 15 avian species and their three loricrin genes, we included a single loricrin gene from nine mammals, two loricrin copies from two snakes and a lizard (Holthaus et al., 2017), one loricrin from two turtle species (Holthaus et al., 2015) and a single loricrin we identified from the two crocodilian species (supplemental table 2) in the phylogenetic analyses.

Bayesian (supplemental figure 3) and maximum likelihood (figure 3) analyses were performed using MrBayes v. 3.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003; Ronquist et al., 2011) and RAxML v. 8.0.0 (Stamatakis, 2014). The topology of these phylogenies largely agreed with a few exceptions related to node support values. The nine loricrins of mammals were used to root the phylogeny with the reptile and avian loricrins forming a well-supported monophyletic clade. Due to a low bootstrap value in the maximum likelihood phylogeny (figure 3), the reptile and avian loricrins are composed of four monophyletic clades comprising a crocodilian clade, squamate clade, testudines and avian loricrin 1 clade and an avian loricrin 2 and 3 clade. In contrast, a high posterior probability support value indicates that the crocodilian clade is the outgroup to all other reptile and avian loricrins (figure 3). These results conflict with the currently accepted topology of reptiles and birds which indicates that crocodilians and birds form the monophyletic clade of archosaurs and that squamates (excluding tuatara) are the outgroup to other reptiles (turtles, crocodilians) and birds (Miller et al. 2012, Crawford et al. 2012).

The squamate loricrin clade consists of two subclades composed of a squamate loricrin 1 gene and a squamate loricrin 2 gene indicating that a duplication occurred early in squamate evolution. Interestingly, the avian loricrin 1 and testudine loricrin genes form a monophyletic clade possibly indicating convergent evolution. In contrast, this may indicate that the avian loricrin 1 gene is highly conserved and represents the ancestral loricrin of turtles and archosaurs. The final clade (LOR2/LOR3 Clade; Figure 4) of avian loricrins consists of multiple loricrin copies with a dynamic duplication history.

The LOR2/LOR3 clade (figure 3) containing avian loricrins was made up of 2 major sister groups. One of these sister groups (LOR2; figure 3), contained passerine loricrin sequences as well as a single loricrin gene belonging to the Hoatzin (OPHHO; *Ophisthocomus hoazin*). While most of the passerine species had only one loricrin gene in this clade (LOR2), the budgerigar (MELUN; *Melopsittacus undulatus*) had two copies which were annotated as LOR2 and LOR2B. The other sister group (LOR3/LOR3B) contained representatives from all species including passerines, however the latter only contained a single loricrin gene while the former all contained 2 copies which displayed a lineage-specific duplication history. These Loricrin sequences were designated as LOR3 and LOR3B (figure 3). LOR2B and LOR3B, or 'B-type' loricrins, were nearly identical to their paralogous LOR2 and LOR3 gene, respectively.

The loricrin genes of the Hoatzin (OPHHO; *Ophisthocomus hoazin*) and Anna's Hummingbird (CALAN; *Calypte anna*) displayed unique evolutionary histories relative to other avian species' LOR2 and LOR3 sequences. The hoatzin was the only non-passerine bird which possessed a loricrin gene in the passerine LOR2 sister group. The hoatzin's other loricrin gene was closely related to the LOR3/LOR3B gene of the adélie penguin, bald eagle, crested ibis, peregrine falcon and killdeer. In the case of Anna's Hummingbird, one loricrin gene formed a sister group with both chimney swift loricrin genes (LOR 3 and 3B) and the other formed a sister group with LOR3 of passerine birds. Our phylogenetic results within the LOR1 clade and the LOR2 and LOR3 clade were largely in agreeance with recent comprehensive avian phylogenies proposed by Prum et al. (2015) which places the enigmatic Hoatzin as a sister group to other landbirds (Ericson et al. 2002, Jarvis et al. 2014, Prum et al. 2015) (figure 3).

The results of these phylogenetic analyses suggest 2 possible scenarios for the evolution of avian loricrins. The first scenario is detailed in figure 3 and involves multiple lineage specific duplications and deletions where, **(1)** Duplication of the ancestral loricrin gene (Anc_LOR) resulted in 2 copies of loricrin (LOR1 and LOR2) before the emergence of the crown birds (Prum et al., 2015). **(2)** Duplication of LOR2 resulted in LOR2 and LOR3 genes. **(3)** Following the divergence of Passeriformes, deletion of LOR2 in all other major orders of birds resulted in a single copy of loricrin (LOR3) in most orders of birds whereas LOR2 was retained in Passeriformes. **(4)** In non-passerine lineages, LOR3 duplicated and produced LOR3B found in Palaeognathae, Galloanserae and Neoaves (excluding Passeriformes) species. **(5)** In the case of Psittacisformes, a suborder of passerine birds, the retained LOR2 duplicated and produced Psittaciforme-specific LOR2B in budgerigar. No evidence was found of a 4th loricrin gene in Psittaciformes suggesting that the LOR3 present in other birds may have been lost in this lineage (figures 3 and 4).

The second possible scenario is that concerted evolution of LOR2B and LOR3B with LOR2 and LOR3 respectively, has resulted in the phylogenetic distribution of loricrin paralogs (figure 4). This second scenario may have occurred though gene conversion, a mechanism of concerted evolution. Gene conversion events occur through unequal recombination where a stretch of DNA is replaced by a homologous region such as those found in duplicate genes that results in the homogenization of both genes (Daiquing 1999). We used GENECONV (Sawyer 1989) to assess the likelihood that gene conversion led to concerted evolution of LOR3/LOR3B. Due to complications associated with incomplete sequences and NNN's, we were left with 7 diverse avian species (chicken, ostrich, ground tit, chimney swift, golden-collared manakin, bald eagle, Atlantic canary) which contained complete loricrin genes and no NNN's. The results of the GENECONV analysis found strong evidence of a gene conversion event between LOR3 and LOR3B for one species, the chimney swift (*Chaetura pelagica)* (BC KA p= 0.00213), which possessed a 91 nucleotide long global fragment that contained 43 polymorphic sites. No other significant gene conversion events were detected between LOR2/LOR2B and LOR3/LOR3B in the other species (supplemental table 3). These results support scenario one, which is detailed in figure 5.

3.4.3 Avian loricrin genes form Gly-loops of variable size and number

The central domain of mammalian loricrin is thought to take on a specialized structural conformation termed the Glycine-loop (Gly-loop) which results from tandemly arranged quasi-repetitive, glycine-rich peptide sequences. The Gly-loop conformation is a key structural motif which provides both barrier-like function as well as elasticity to the epidermis and its appendages (Hohl et al. 1990, Candi et al. 2005). The properties conferred by the Gly-loop motif depend heavily on the relative composition of amino acids which make up the peptide repeats of the central domain as well as the presence of specific residues in the N- and C- termini (Steinert et al. 1991). In order for a protein sequence to take on the Gly-loop conformation, it must have the general form $x(y)_n$, where “x” is an aromatic or aliphatic residue, “y” is usually glycine or a polar residue such as serine and “n” is the number of polar residues. Another key characteristic of mammalian Gly-loops is a large amount of variation in the size and the number of repeating peptide units present in the central domain (Steinert et al. 1991). Here, we calculated the size and total number of Gly-loops for six diverse avian species as well as three mammals (supplemental table 3.2). This allowed us to analyze the interspecific and intraspecific amino acid variation of avian Gly-loops as well as the variation in the number and size of avian Gly-loops.

We found that the repetitive units which comprise the central domain of avian loricrins conform to the general form ($x(y)_n$) required for the formation of Gly-loops and that there is a significant amount of variation in the amino acid composition and organization of avian loricrins. In general, there are distinct amino acid differences between Gly-loops formed by the avian loricrin genes. LOR1 glycine rich loops are interspersed by glutamine and proline residues and are indexed primarily on aliphatic isoluecines, or the “x” of the $x(y)_n$ conformation (figure 5A). The glycine rich loops of LOR2 and LOR3 are interspersed by conserved lysine and cysteine residues. However, LOR2 loops are indexed primarily on aliphatic methionines, while LOR3 loops are indexed on either tyrosine or isoluecine residues (figure 4B). B-type loricrins (LOR2B and LOR3B) conform to the same general Gly-loop amino acid characteristics as their duplicates LOR2 and LOR3.

Although all avian loricrins conform to the general form $x(y)_n$, we observed considerable variation in the number and the size of Gly-loops (supplemental table 3.2). As previous studies (Hohl et al. 1990, Steinert et al. 1991) have shown, we found that mammalian loricrins vary extensively in both size and number of Gly-loops. The lorcrin gene of the Orca (*Orcinus orca*) was predicted to contained 6 Gly-loop domains, while human (*Homo sapiens*) and mouse (*Mus musculus*) loricrin contained 21 and 22 Gly-loop domains respectively. Furthermore, the Gly-loops of mouse loricrin were generally longer (average loop size = 18.18) than those of humans (average loop size = 10.62) and the Orca (average loop size = 16.67) (supplemental table 3.2). Overall, the variation observed across mammalian loricrins is thought to result in slight differences in the mechanical properties exhibited by the CE (Steinert et al. 1991).

Similar to mammalian loricrins, we observed significant variation in the size and number of Gly-loop domains of avian loricrins. Out of six avian species analyzed, the longest glycine loop contained 30 residues between “x” residues ($x(y)_n$) (MANVI LOR2) and the shortest contained two (MANVI + PHUMI LOR2). The highest amount of interspecific variation in the number of Gly-loops was in LOR3/LOR3B, where the total number of predicted Gly-loops ranged from 8 in LOR3B of the bald eagle to 48 in LOR3B of the chicken (supplemental table 3.2). While there was considerable interspecific variation in the total number of Gly-loops making up LOR3/LOR3B, the size parameters of those loops were more conserved (LOR3: average loop size=10.52, SD=1.48, n=8) relative to the size parameters of the loops of other avian loricrins (LOR1: average loop size=11.98, SD=2.61, n=6 ; LOR2: average loop size=19.52, SD=2.99, n=2) (supplemental table 3.2). Despite the high amount of interspecific variation observed across the size and number of Gly-loops in LOR3/LOR3B, there was relatively little variation observed within species. For instance, LOR3 and LOR3B of the chicken are predicted to contain 43 and 48 total loops respectively, whereas LOR3 and LOR3B of the bald eagle are predicted to contain 10 and 8 loops respectively (supplemental table 3.2). Overall, these results demonstrate that avian loricrins, much like mammalian loricrins, exhibit a large amount of variation in the size and number of the Gly-loops even between closely related species. Due to uncertainty with the number of actual NNNs present in incomplete avian loricrin genes, our analysis was restricted to a small sample size (n = 6). Therefore, more complete avian loricrin sequences are needed to make inferences relating the size and number of Gly-loops to functional properties of avian loricrins.

3.4.4 Amino acid compositional differences between avian loricrin genes suggests functional diversity

Similar to mammalian loricrins, the amino acid composition of avian loricrins are extremely biased with over 50% of the gene is composed of glycine and serine (supplemental table 4 A and B). Other prevalent amino acids are cysteine, tyrosine, lysine and glutamine, which are all associated with protein cross-linking (Hohl et al., 1990; Steinert et al., 1991; Candi et al., 2005; Eckhart et al., 2013).

In order to further assess the potential functional properties of avian loricrins, we analyzed the amino acid composition of all loricrin sequences identified as having less than 10% NNNs. Using the ExPASy ProtParam tool (Gastieger et al. 2005) we calculated the percent composition of the 20 amino acids for 48 avian, 8 reptilian, and 9 mammalian loricrin genes (supplemental table 4 A and B). Using these data, we generated a Principle Component Analysis (PCA) using the Bioconductor pcaMethods package in R (Stacklies et al. 2007, R Studio Team 2015). The PCA plot (figure 6) was able to explain 46.79% (PC1 = 0.2764% , PC2 = 0.1915%) of the total variance between the amino acid composition of loricrin sequences. The PCA also found that principle component 1 (PC1) differentiated avian LOR1 into a distinct cluster relative to all other loricrin genes. The amino acid composition of the remaining loricrin sequences failed to sort into unique clusters; however, LOR2 and LOR3 of birds did group together but could not be differentiated from one another. The loricrins of crocodilians, snakes and some mammals increased the vertical spread (PC2). Overall, these results demonstrate that avian LOR1 has a conserved and unique amino acid composition, while avian LOR2 and LOR3/LOR3B loricrins could not be differentiated from reptilian and mammalian loricrin genes (figures 6 and 7). Together with our phylogenetic results (figures 3 and 4), these results suggest that avian LOR1 diverged early in the evolution of birds and has remained conserved within birds.

To characterize which amino acid residues were primarily contributing to the PCA analysis results, we performed an analysis of variance (ANOVA) to analyze the differences of the mean amino acid content between avian loricrin genes. We observed statistically significant differences in 7 amino acid residues between LOR1 and LOR2 and 11 amino acid residues between LOR1 and LOR3 (supplemental table 5). The most significant amino acid differences between LOR1 from LOR2 and LOR3 were observed in serine (LOR1: $\bar{x}$=12.64%, n=17 ; LOR2: $\bar{x}$=21.86%, n=4, $F_{17,5}$=119.59, p<0.001 ; LOR3: $\bar{x}$=27.79%, n = 15, $F_{17,14}$=372.9, p<0.001), cysteine (LOR1: $\bar{x}$=3.88%, n=17 ; LOR2: $\bar{x}$=8.82%, n=4, $F_{17,5}$=156.83, p<0.001 ; LOR3: $\bar{x}$=6.12%, n=15, $F_{17,14}$=43.15, p<0.001) and proline (LOR1: $\bar{x}$=3.38%, n=17 ; LOR2: $\bar{x}$=0.66%, n=4, $F_{17,5}$=222.51, p < 0.001 ; LOR3: $\bar{x}$=1.07%, n=15, $F_{17,14}$=166.63, p<0.001).

We also conducted an ANOVA analysis of the amino acid composition between LOR2 and LOR3, which did not form unique groups in our PCA (figure 6). For these analyses, we included the LOR2 gene of the White-rumped munia (*Lonchura striata*) in the interest of identifying more subtle differences in amino acid composition (Yang et al. 2015). We identified significant differences for 5 amino acid residues (serine, glycine, histidine, tyrosine, and cysteine) between LOR2 and LOR3. Moreover, 4 (serine, glycine, tyrosine, and cysteine) of these amino acid residues are known to be important throughout the cornification process as well as in maintaining the structure of the epidermis and its's appendages (Candi et al. 2005) (figure 6 and supplemental table 5). We found that LOR2 contained significantly higher amounts of glycine (LOR2: $\bar{x}$ =47.76%, n=5 ; LOR3: $\bar{x}$ =38.3%, n=15 ; $F_{5,15}$=22.28, p<0.001) and cysteine (LOR2: $\bar{x}$ =8.82% n=5 ; LOR3: $\bar{x}$ =6.1%, n=15 ; $F_{5,15}$=27.68, p<0.001), and significantly less serine (LOR2: $\bar{x}$=21.86%, n=5 ; LOR3: $\bar{x}$=27.69%, n=15 ; $F_{5,15}$=20.31, p<0.001) relative to LOR3. Additionally, LOR2 was found to have very low amounts of tyrosine ($\bar{x}$ =0.175%, n=4, σ=0.12), with LOR2 of the Atlantic Canary (*Serinus canaria*) containing no tyrosine residues. This contrasts with avian LOR3 ($\bar{x}$ =4.12%, n=15, σ=1.99) as well as mammalian loricrins ($\bar{x}$ =3% n=9, σ=1.36) which have tyrosine residues conserved throughout their central domains.

Our phylogenetic analyses (figure 3) found that with the exception of the Hoatzin, only passerine birds possess LOR2. However, most avian species, except Budgerigar, possess a LOR3 gene. In order to determine if the passerine LOR2 (P-LOR2) and passerine LOR3 (P-LOR3) differ in amino acid composition (Unlike LOR3/LOR3B of non-passerine birds); we repeated our ANOVA analyses between LOR2 and LOR3 using either only passerine genes or only non-passerine genes (NP). Only tyrosine content (P-LOR3: $\bar{x}$ =5.68%, n=6 ; LOR3-NP: $\bar{x}$ =3.08, n=9 ; $F_{6,9}$=10.14, p<0.01) was found to significantly differ between P-LOR3 and LOR3 of non-passerine birds indicating they have nearly identical amino acid compositions. In contrast, significant differences were observed in cysteine (P-LOR2: $\bar{x}$ =8.82%, n=5 ; P-LOR3: $\bar{x}$ =5.87%, n=6 ; $F_{5,6}$=27.3, p<0.001), glycine (P-LOR2: $\bar{x}$ =47.76%, n=5 ; P-LOR3: $\bar{x}$ =40.57%, n=6 ; $F_{5,6}$=12.76, p<0.01), serine (P-LOR2: $\bar{x}$ =21.86%, n=5 ; P-LOR3: $\bar{x}$ =26.57%, n=6 ; $F_{5,6}$=18.42, p<0.001), tyrosine (P-LOR2: $\bar{x}$ =0.22%, n=5 ; P-LOR3: $\bar{x}$ =2.08%, n=6 ; $F_{5,6}$=50.24, p<0.001) and valine (P-LOR2: $\bar{x}$ =0.54%, n=5 ; P-LOR3: $\bar{x}$ =2.08%, n=6 ; $F_{5,6}$=25.2, p<0.001) between P-LOR2 and P-LOR3 (supplemental table 5). These results support the hypothesis that LOR2 is distinct from other avian loricrins and was most likely lost in most lineages of birds following the divergence of Passeriformes from other crown birds (figure 4).

3.5 Discussion

The results of this study demonstrate that loricrin genes are conserved within the EDC of birds and reptiles indicating that loricrin is an essential component of not only the mammalian cornified envelope (CE) (Hohl et al. 1990, Candi et al. 2005), but most likely of all amniotes. All loricrins identified were tandemly arrayed and found in the same orientation within the EDC between the genes EDQL (formerly EDQM3) and EDYM1. Although all species investigated had complete genome assemblies available on NCBI, the quality of the assemblies varied significantly (Zhang et al. 2014 and Jarvis et al. 2014; supplemental table 2). However, we were not able to find a relationship between the quality of loricrins and genome quality (results not shown). A frequent problem observed is the interruption of loricrin genes due to scaffold breaks (supplemental figure 1). Additionally, we encountered an abundance of unknown nucleotides interrupting the CDS of loricrin sequences resulting in artificial frameshifts. These problems are consistent with the results of previous studies (Milinkovitch et al. 2010, Hron et al. 2015, Peona et al. 2018) which have found that genome assemblers have difficulty resolving highly repetitive and GC rich regions of the genome, which can result in large numbers of gaps (i.e. fragmented sequences). Loricrins are highly repetitive which likely contributes to these problems. Similar problems have been encountered and resolved in other avian EDC genes through direct sequencing. Strasser et al. (2015) encountered a frameshift in the central domain of another avian EDC gene, the cysteine-rich EDCRP gene of the zebra finch. It has been demonstrated that EDCRP is expressed in the embryonic subperiderm of chickens as well as in the barbule cells of developing feathers (Strasser et al. 2015), which suggests it plays a role in the morphogenesis and structure of feathers and scales. Upon direct sequencing of zebra finch EDCRP, the frameshift was resolved and a single continuous open reading frame was identified (Strasser et al. 2015). Therefore, it’s likely the frameshifts and premature stop codons observed in several loricrins are artificial and would be resolved upon direct sequencing.

The number of loricrin genes identified varied across different groups of organisms. Previous studies have identified 2 loricrin genes in squamates, and only a single loricrin gene in turtles while there are 3 loricrin genes which have been identified in the chicken (Strasser et al. 2014, Holthaus et al. 2014, Holthaus et al. 2017). Birds were the only group of species in our study which possessed 3 loricrin genes, whereas we were only able to identify a single copy of loricrin in crocodilian species. The results of our analysis of the crocodilian EDC are consistent with the recently published findings of Holthaus et al. 2018. All avian loricrin genes identified were located at the same position within the EDC as the 3 loricrin genes identified in the chicken by Strasser et al. (2014) (figure 1). We identified evidence of 3 loricrins in all bird species where the entire region of the EDC in which loricrins are located was assembled on a single scaffold with the exception of the pigeon, *Columbia livia*, where only 2 copies were identified (supplemental table 3.1). We did not identify any avian species that contained more than 3 copies or less than 2 copies of loricrin indicating that 3 copies of loricrin were most likely present in the most recent common ancestor (MRCA) of crown birds. These results, together with those of previous studies (citations), demonstrate a complex and dynamic duplication history of loricrins in birds and reptiles.

Our phylogenetic analyses identified four major clades of loricrins across birds and reptiles (figure 3 and supplemental figure 3). In contrast to accepted comprehensive species phylogenies, crocodilian loricrins formed the outgroup to all other birds and reptiles, and the testudine loricrins grouped with avian LOR1 (St John et al. 2012, N.G. Crawford et al. 2012, Miller et al. 2012, Holthaus et al. 2018). These results demonstrate the evolutionary uncertainty described in previous studies (St John et al. 2012, N.G. Crawford et al. 2012, Holthaus et al. 2015) associated with defining the basal clade of all sauropsids. It is known that the epidermal appendages of birds and reptiles are highly specialized adaptations which exhibit significant molecular and genetic diversity even across phylogenetically similar species (Gremillet et al. 2005 and Wang et al. 2016). It is possible that the results of our phylogenetic analyses reflect evolutionary adaptations associated with specialization of epidermal appendages such as crocodilian scales or the carapace of testudines, and are not indicative of the true phylogenetic history of birds and reptiles. These results suggest crocodilian loricrins have undergone little evolutionary divergence relative to those of birds and other reptiles. Additionally, these results suggest the possible convergent evolution of testudine loricrins with avian LOR1. Testudines, like birds, possess evolutionarily unique appendages in their shell and scutes, however unlike avian LOR1, testudines loricrins are ubiquitously expressed throughout the epidermis and its appendages (Strasser et al. 2014 and Holthaus et al. 2015). The presence of NNNs in the loricrin sequences of both testudine species (Green sea turtle = 35.4% NNNs, Painted turtle = 54.5% NNNs) may have impacted our phylogenetic results. Finally, PCA analysis demonstrated that the amino acid composition of avian LOR1 is distinct from that of testudine loricrins (figure 6).

The LOR2/2B group of the second clade of avian loricrins contained only passerine loricrin sequences and LOR3 of the Hoatzin. Conversely, LOR2 of the Hoatzin grouped with other loricrins in the LOR3/LOR3B group. The nomenclature for Hoatzin loricrins, as all other species, was based on the genomic orientation of loricrins relative to other EDC genes (Figure 1). These data suggest a genomic inversion of LOR2 and LOR3 of the Hoatzin (Figure 1). We also found another, larger inversion in a different region of the turkey's EDC indicating that inversions may be a major contributor to the evolution of EDC in birds (Holthaus et al., 2018; figure 1-A).

These phylogenetic results support two likely scenarios for the evolution of avian loricrins (figure 4). The first scenario entails the loss of an ancestral LOR2 from most orders of birds and its retention in Passeriformes, followed by recent lineage-specific duplications of LOR3 in most orders of birds. Alternatively, scenario 2 entails concerted evolution which has homogenized the LOR3/LOR3B and LOR2/LOR2B genes. Concerted evolution takes place when genes undergo gene conversion resulting in the homogenization of their DNA sequences (Daiquing 1999, Sawyer 1989). We found evidence of a statistically significant gene conversion event between LOR3 and LOR3B of the chimney swift (*Chaetura pelagica)* (CHAPE BC KA p=0.00213)(supplemental table 3). The likely concerted evolution in LOR3/LOR3B of the chimney swift, in combination with the absence of evidence supporting additional gene conversion events in other avian species suggest that a combination of concerted evolution, gene deletions and gene duplications have shaped the evolution of avian loricrins .

In the first scenario of the evolution of avian loricrins (figure 4), the recent gene duplications of LOR3 in most species analyzed resulted in the nearly identical LOR3/LOR3B and LOR2/LOR2B genes. Gene duplications have long been accepted as a major mechanism promoting evolutionary change (Holland et al. 1994). The most commonly observed mechanism of gene duplication, which occurs at high frequencies in birds, is unequal crossing over which generates tandem duplicates that are nearly identical in sequence and are genetically linked (Zhang 2003). Previous studies (Ahlroth et al. 2001) have found that in the chicken, unequal crossing over has resulted in wide variation in the copy number of the avidin gene between individuals. The tandem linkage of avian loricrins is characteristic of gene duplications by unequal crossing over. Interestingly, previous studies (Dawson et al. 2007, Völker et al. 2010, Backström et al. 2010) have provided evidence that recombination-based processes play a major role in avian evolution. This may correlate with the general absence of apparent loricrin "duplicates" (LOR3B/LOR2B) from all passerine birds except for the budgerigar (figure

3). In the case of the budgerigar, LOR3 may have been lost, and instead LOR2 was duplicated into LOR2B. These results highlight the dynamic evolutionary nature of avian loricrins, even at the species level. Future studies which include additional loricrins will further elucidate if the similarities observed between avian LOR3/LOR3B and LOR2/LOR2B are primarily the result of recent gene duplications, concerted evolution or the result of both mechanisms.

In mammals, loricrin functions as the major reinforcement protein of the CE, but also provides high levels of flexibility to the epidermis and its' appendages. These key properties are thought to be achieved through a specialized conformation known as a Gly-Loop which results from the tandemly arranged quasi-repetitive peptide units which make up the central domain of loricrins. These highly flexible loops consist of long stretches of primarily glycine and serine residues, but they do tolerate substitutions of other residues. These stretches of glycine and serine residues with occasional substitutions of polar residues are indexed upon aromatic and aliphatic residues which may associate to form a three-dimensional rosette-like array (Hohl et al. 1990, Steinert et al. 1991). Mammalian loricrins vary extensively in their size, exact organization and amino acid content, however they maintain the general form $x(y)_n$ required for the formation of Gly-loops. This variation in mammalian Gly-loops is thought to play a major role in the mechanical properties conferred to the CE, such as flexibility, and tensile strength (Ishida-Yamamoto et al. 1998). There are also known to be allelic variants of loricrin with slightly different amino acid compositions within individual populations which influence the properties of the epidermis and its' appendages (Hohl et al. 1990, Steinert et al. 1991, Eckhart et al. 2013). Gly-loops provide their barrier function via weak hydrophobic interactions between the glycine and serine residues of adjacent Gly-loops, as well as other components of the CE such as keratins and filaggrin (figure 5 A and B). These interactions are thought to be easily interrupted upon application of stress which induces the formation of a separate but similar set of interactions. Once the stress is released, these new interactions are released to form yet another set of interactions similar but not identical to the original unstressed state. This is termed the "Velcro hypothesis" and accounts for the known flexibility and elastic recovery of the mammalian CE (Steinert et al. 1991).

Our results demonstrate that while there is significant variation across avian loricrins, they still adhere to the general form $x(y)_{n.}$ The observation that the sizes and sequences of avian Gly-loops are highly variable, but that the common structural motif of $x(y)_n$ is conserved implies that the structural motif is more important for proper loricrin function than the exact sequence itself. It is possible that this variation also contributes to the large amount of diversity observed in the feathers and scales of different species of birds; however more data is needed to identify any correlations between Gly-loop sequences and specific epidermal properties.

The Glycine-loop domains of avian loricrins differ from those of mammals primarily in the identity of the aromatic/aliphatic amino acids upon which the loops are indexed. In mammals, these residues are primarily tyrosines, but there are occasional isoleucines, alanines, phenylalanines, and methionines. For example, the Gly-loops of mouse loricrin are indexed almost exclusively on tyrosine residues, whereas in human loricrin the loops are indexed on a combination of phenylalanine, tyrosine, isoleucine and valine residues. The general consensus repetitive unit of LOR1 is

HQ(G/S)QGPICI(**G**$_x$)SG which maintains the general form of $x(y)_n$. The Isoleucine (I) residues serve as long-chain aliphatic residues which are known to associate with one another to form a hydrophobic core, while the variable stretches of glycine and serine residues form the ‘loops’ of the Gly-loop (figure 4A). The sequence HQ(G/S)Q is conserved preceding the glycine-rich loop sequences. These glutamine residues are possibly involved in transglutamination via transglutaminases. In avian LOR1, the primary residue upon which loops are indexed are aliphatic isoleucines while in LOR2/3 the identity of these residues is more variable but primarily are tyrosines, isoleucines and methionines. Furthermore, in avian loricrins, long-chain aliphatic residues are often found as dimers or trimers, whereas Gly-loops associated with aromatic amino acids are generally indexed upon only a single residue. This may result from the strength of the respective interactions. It is known that an extended row of aromatic residues is likely to stack in an ordered manner so that the phenyl rings align at a preferential distance and these interactions contribute 1-2.5 kcal/mol per aromatic pair toward the overall stability of the protein (Burley et al. 1985, Singh et al. 1985). In contrast, aliphatic residues do not by themselves associate to form highly ordered arrays, but it is well known that they do associate to form a hydrophobic core. It is possible that the presence of multiple adjacent aliphatic residues aids in the association of aliphatic residues packing together to form a hydrophobic core (Rose et al. 1980, Zhu et al. 1993).

Mammals possess a single loricrin gene which is preferentially crosslinked by different TGases throughout the process of cornification, whereas we found there are generally 3 loricrin genes in birds. It has been demonstrated that variation in the composition of amino acid residues which make up structural proteins, often correlates with different functionality (Candi et al. 2005). We analyzed the variation in amino acid content of the different avian loricrin genes and found that the amino acid contents of each respective amino acid in LOR1 were significantly different from those of other avian loricrins (supplemental tables 4 and 5). Along with expression data from Strasser et al. (2014) which demonstrates LOR1 is differentially expressed in the chicken relative to LOR3/3B, these results indicate that the Gly-loops formed by the central domain of avian LOR1 likely have a unique functional role which is distinct from those of other loricrins. There were also significant differences in the amino acid compositions of LOR2 vs. LOR3/LOR3B, specifically in cysteine, glycine, serine and tyrosine contents all of which are known to be involved in the process of keratinocyte cornification (Rice et al. 2013, Eckhart et al. 2013). While we did observe that LOR3/LOR3B exhibited increased variation relative to LOR2, we contribute this to the fact that LOR2 is only found in Passeriformes while LOR3/LOR3B are represented by a much more diverse group of avian orders. There was no significant variation observed in the amino acid contents of type-B loricrins from their respective duplicates. This may be expected given that in the chicken, LOR3 and LOR3B have identical expression profiles in epidermal tissues (Strasser et al. 2014). Due to significant differences between the amino acid contents of LOR2 and LOR3/LOR3B, we predict that in passerine species LOR2 most likely exhibits a different expression profile than that of LOR3, and possibly a distinct function.

The feathers and scales of different bird species are novel adaptations which possess highly specialized properties which correspond to the diverse environments and lifestyles associated with birds. For example, the feathers of the great cormorant (*Nipponia nippon)* exhibit a unique morphological-functional adaptation to diving which balances the constraints of buoyancy and thermoregulation (Gremillet et al. 2005). The feathers of the Humboldt penguin (*Spheniscus humboldti*) exhibit unique hydrophobicity and anti-adhesion characteristics which endow them with excellent anti-icing properties and allow them to survive in arctic environments (Wang et al. 2016). Along with the variation previously described between different loricrin orthologs, we also observed interspecific variation in the amino acid contents of respective loricrin genes. This variation was most prevalent in LOR3, which was found in all species examined and is ubiquitously expressed in epidermal tissues (Strasser et al., 2014). This interspecific variation resembles that observed across mammalian loricrins, which is known to influence the mechanical properties endowed to the resulting CE (Hohl et al. 1990, Steinert et al. 1991). We propose that this variation in amino acid composition may correspond to specific evolutionary adaptations of feathers and other avian epidermal appendages . The least amount of interspecific variation in amino acid content was observed with LOR1 which interestingly is not expressed in feathers.

In mammals, loricrins are crosslinked primarily by the process of transglutamination via TGases. TGases catalyze the formation of N-(γ-glutamyl)-lysine isodipeptide bonds through the preferential, step-wise covalent cross-linking of glutamine and lysine residues located in both the N- and C- termini as well as interspersed throughout the central domain (Eckhart et al. 2013). We found that avian loricrins also possess several glutamine and lysine residues located at conserved positions. In LOR1, there are conserved glutamine residues in the H**Q**(G/S)**Q** portion of each repeat (figure 5 A). In LOR2 and LOR3, like mammalian loricrins, there are conserved glutamine residues in both the N- and C-termini that are located adjacent to lysines in the sequence

QQK. There are also conserved lysine residues located near the aromatic/aliphatic residues upon which the glycine loops or indexed, furthermore these lysines are occasionally located adjacent to glutamine residues (figure 5 B). Further analysis of the composition of avian epidermal appendages is required to determine if loricrins are indeed the primary substrates for crosslinking by TGases in birds, however due to the large amount of molecular similarities with mammalian loricrin and the presence of conserved glutamine and lysine residues, a similar function may be inferred.

Along with transglutamination, it is also known that disulfide bonding between adjacent cysteine residues plays a major role in facilitating the development of the epidermis and epidermal appendages in both mammals and birds (Hynes et al. 1977, Kalinin et al. 2002). Strasser et al. (2015) characterized a cysteine-rich SEDC protein (EDCRP) in the chicken which is expressed in the subperiderm of feathers and scales. EDCRP consists of over 50% cysteine and most likely participates in disulfide bonding throughout epidermal development. Moreover, the cysteine content of several SEDC genes identified in the chicken exceeded 20% (Strasser et al. 2014). We identified adjacent cysteine residues located at conserved sites near the apex of many of the larger loricrin loops in avian LOR2 and LOR3. These residues potentially participate in disulfide bonding with other SEDC proteins such as EDCRP, other loricrins as well as various β-keratins. Disulfide bonding may also help facilitate the anchoring of loricrin and its associated proteins to the CE via interactions with SEDC genes similar to mammalian involucrin (Vanhoutteghem et al. 2008, Strasser et al. 2014). The presence of conserved adjacent cysteine residues throughout LOR2 and LOR3 suggest loricrins participate in not only transglutamination but may also take part in a combination of covalent cross-linking interactions that result in the unique mechanical properties observed in feathers and other avian epidermal appendages.

Overall, the results of this study demonstrate a complex and dynamic evolutionary history of loricrins in archosaurs which likely involved gene duplications and deletions as well as concerted evolution and chromosomal inversions. The availability of more complete avian genomes is necessary to gain further insight into the evolution of avian loricrins. Given the conservation of the Gly-loop structure and expression profile of the loricrins in the chicken (Strasser et al., 2014) it is likely that avian loricrins constitute a major portion of the CE. Future studies which focus on a detailed expression profile of loricrins in other birds such as the passerines may provide further insight into the evolution of avian loricrin genes as well as the role they play in conferring the unique mechanical properties observed across the feathers of birds.

3.6 Figures

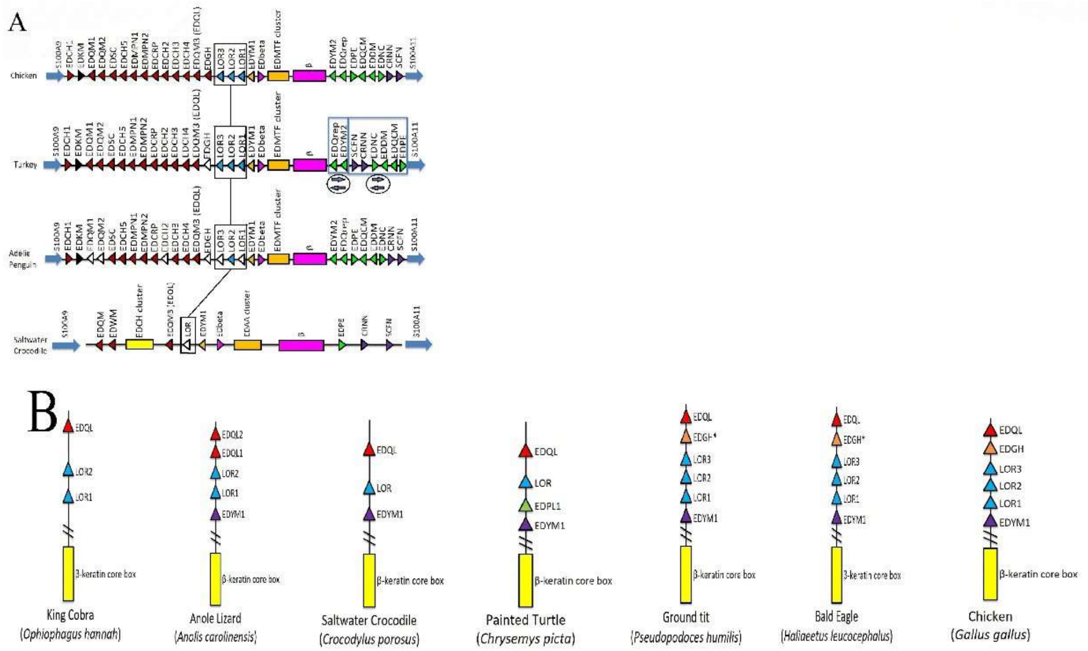
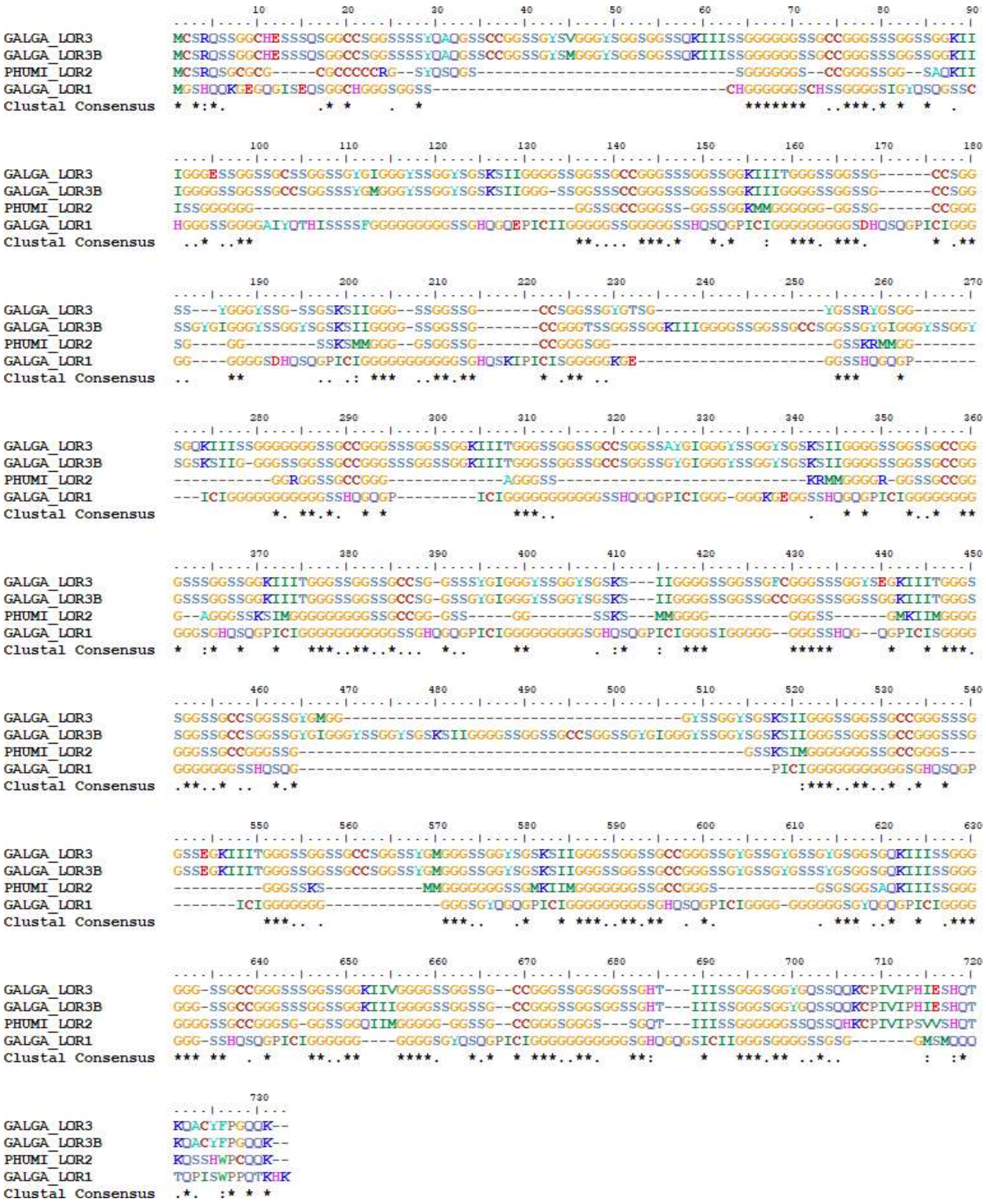


Figure 3.1 : Genomic organization of Loricrin within the EDC of archosaurs. (A) schematic overview of the conservation of the entire EDC of the Chicken (*Gallus gallus*), Turkey (*Meleagris gallopavo*), Adélie penguin (*Pygoscelis adeliae*), and Saltwater crocodile (*Crocodylus porosus*). Chicken EDC organization identical to that proposed by Strasser et al. (2014), with the exception of the identification of EDMTF5. Filled in arrows with black outlines represent complete SEDC genes, arrows with white fills indicate incomplete genes. Gene number annotations LOR1, LOR2, LOR3 come from annotations of chicken loricrins by Strasser et al. (2014). Colors correspond to classifications by Strasser et al. (2014). **(B)** Schematic representation of the region of the EDC which contains loricrins between the conserved genes EDQL (formerly EDQM3) and the β-keratin core box. The genes EDQL and EDYM1 are conserved across all species examined. EDGH sequences were identified in all avian species, however, the start codon from the chicken identified by Strasser et al. (2014) was not present in other bird species(*). The loricrin copy number varied across different groups of organisms, but in general squamates possessed 2, crocodilian and testudine species contained 1, and birds 3 loricrin genes. Arrow colors represent related genes. Parallel lines between EDYM1 and β-box indicate presence of variable number of lineage-specific EDC genes. King Cobra EDC genes identified by Holthaus et al. (2016), anole lizard EDC genes identified by Strasser et al. (2014), and Painted turtle EDC genes identified by Holthaus et al. (2015).



```
                            10        20        30        40        50        60        70        80        90
                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
GALGA_LOR3         MCSRQSSGGCHESSSQSGGCCSGGSSSSYQAQGSSCCGGSSGYSVGGGYSGGSGGSSQKIIISSGGGGGGSSGCCGGGSSSGGSSGGKII
GALGA_LOR3B        MCSRQSSGGCHESSSQSGGCCSGGSSSSYQAQGSSCCGGSSGYSMGGGYSGGSGGSSQKIIISSGGGGGGSSGCCGGGSSSGGSSGGKII
PHUMI_LOR2         MCSRQSGCGCG-----CGCCCCCRG--SYQSQGS----------------------------SGGGGGGS--CCGGGSSGG--SAQKII
GALGA_LOR1         MGSHQQKGEGQGISEQSGGCHGGGSGGSS---------------------------------CHGGGGGGSCHSSGGGGSIGYQSQGSSC
Clustal Consensus  * *:*.          .* *    .  *                                    *******  ..***.* *  *  .

                           100       110       120       130       140       150       160       170       180
                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
GALGA_LOR3         IGGGESSGGSSGCSSGGSSGYGIGGGYSSGGYSGSKSIIGGGGSSGGSSGCCGGGSSSGGSSGGKIIITGGGSSGGSSG------CCSGG
GALGA_LOR3B        IGGGGSSGGSSGCCSGGSSSYGMGGGYSSGGYSGSKSIIGGG-SSGGSSSCCGGGSSSGGSSGGKIIIGGGGSSGGSSG------CCSGG
PHUMI_LOR2         ISSGGGGGG------------------------------------GGSSGCCGGGSS-GGSSGGKMMGGGGGG-GGSSG------CCGGG
GALGA_LOR1         HGGGSSGGGGAIYQTHISSSSFGGGGGGGGGSSGHQGQEPICIIGGGGGSSGGGGGSSHQSQGPICIGGGGGGGGGSDHQSQGPICIGGG
Clustal Consensus   ..* ..**                                  **.... ***.*   *.*   :  ***. ***.        * .**

                           190       200       210       220       230       240       250       260       270
                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
GALGA_LOR3         SS---YGGGYSSG-SSGSKSIIGGG--SSGGSSG-------CCSGGSSGYGTSG------------------YGSSRYGSGG-------
GALGA_LOR3B        SSGYGIGGGYSSGGYSGSKSIIGGGG-SSGGSSG-------CCGGGTSSGGSSGGKIIIGGGGSSGGSSGCCSGGSSGYGIGGGYSSGGY
PHUMI_LOR2         SG----GG--------SSKSMMGGG--GSGGSSG-------CCGGGSGG------------------------GSSKRMMGG-------
GALGA_LOR1         GG----GGGGSDHQSQGPICIGGGGGGGGGGSGHQSKIPICISGGGGGKGE--------------------GGSSHQGQGP-------
Clustal Consensus  ..     **          .. .: ***   ..**.**       * .** ..                       ***    *

                           280       290       300       310       320       330       340       350       360
                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
GALGA_LOR3         SGQKIIISSGGGGGGGSSGCCGGGSSSGGSSGGKIIITGGGSSGGSSGCCSGGSSAYGIGGGYSSGGYSGSKSIIGGGGSSGGSSGCCGG
GALGA_LOR3B        SGSKSIIG-GGGSSGGSSGCCGGGSSSGGSSGGKIIITGGGSSGGSSGCCSGGSSGYGIGGGYSSGGYSGSKSIIGGGGSSGGSSGCCGG
PHUMI_LOR2         -----------GGRGGSSGCCGGG-------------AGGGSS-----------------------------KRMMGGGGR-GGSSGCCGG
GALGA_LOR1         ---ICIGGGGGGGGGGGSSHQGQGP---------ICIGGGGGGGGGGGSSHQGQGPICIGGG-GGGKGEGGSSHQGQGPICIGGGGGGGG
Clustal Consensus             *. **.*.  * *              ***..                            .   * *    *..*  **

                           370       380       390       400       410       420       430       440       450
                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
GALGA_LOR3         GSSSGGSSGGKIIITGGGSSGGSSGCCSG-GSSSYGIGGGYSSGGYSGSKS---IIGGGGSSGGSSGFCGGGSSSGGYSEGKIIITGGGS
GALGA_LOR3B        GSSSGGSSGGKIIITGGGSSGGSSGCCSG-GSSGYGIGGGYSSGGYSGSKS---IIGGGGSSGGSSGCCGGGSSSGGSSGGKIIITGGGS
PHUMI_LOR2         G--AGGGSSKSIMGGGGGGGGGGSSGCCGG-GSS-----GG-------SSKS---MMGGGG---------GGGSS------GMKIIMGGGG
GALGA_LOR1         GGGSGHQSQGPICIGGGGGGGGGGGSSGHQGQGPICIGGGGGGGGGSGHQSQGPICIGGGSIGGGGG--GGGSSHQG--QGPICISGGGG
Clustal Consensus  *  :*  *   *   ***..**..*...  *..      **         . :*   :  ***         *****      *   * ***.

                           460       470       480       490       500       510       520       530       540
                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
GALGA_LOR3         SGGSSGCCSGGSSGYGMGG-----------------------------------GYSSGGYSGSKSIIGGGSSGGSSGCCGGGSSSG
GALGA_LOR3B        SGGSSGCCSGGSSGYGIGGGYSSGGYSGSKSIIGGGGSSGGSSGCCSGGSSGYGIGGGYSSGGYSGSKSIIGGGSSGGSSGCCGGGSSSG
PHUMI_LOR2         GGGSSGCCGGGSSG--------------------------------------------------GSSKSIMGGGGGGGSSGCCGGGS---
GALGA_LOR1         GGGGGGGSSHQSQG-----------------------------------------------------PICIGGGGGGGGGGGSGHQSQGP
Clustal Consensus  .**..* ..  *.*                                                      :***..**..* .*  *

                           550       560       570       580       590       600       610       620       630
                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
GALGA_LOR3         GSSEGKIIITGGGSSGGSSGCCSGGSSYGMGGGSSGGYSGSKSIIGGGSSGGSSGCCGGGSSGYGSSGYGSSGYGSGGSGQKIIISSGGG
GALGA_LOR3B        GSSEGKIIITGGGSSGGSSGCCSGGSSYGMGGGSSGGYSGSKSIIGGGSSGGSSGCCGGGSSGYGSSGYGSSSYGSGGSGQKIIISSGGG
PHUMI_LOR2         ----------GGGSSKS-----------MMGGGGGGGSSGMKIIMGGGGGGGSSGCCGGGS-----------GSGSGGSAQKIIISSGGG
GALGA_LOR1         -------ICIGGGGGGG-------------GGGSGYQGQGPICIGGGGGGGGGSGHQSQGPICIGGGG-GGGGGGSGYQGQGPICIGGGG
Clustal Consensus            ***.. .                  ***..   .*   * ***..**.**  . *.           . *** ..*  *  .***

                           640       650       660       670       680       690       700       710       720
                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
GALGA_LOR3         GGG-SSGCCGGGSSSGGSSGGKIIVGGGGSSGGSSG--CCGGGSSGGSGGSSGHT---IIISSGGGSGGYGQSSQQKCPIVIPHIESHQT
GALGA_LOR3B        GGG-SSGCCGGGSSSGGSSGGKIIIGGGGSSGGSSG--CCGGGSSGGSGGSSGHT---IIISSGGGSGGYGQSSQQKCPIVIPHIESHQT
PHUMI_LOR2         GGGGSSGCCGGGSG-GGSSGGQIIMGGGGG-GGSSG--CCGGGSGGS---SGQT---IIISSGGGGGGSSQSSQHKCPIVIPSVVSHQT
GALGA_LOR1         GGG-SSHQSQGPICIGGGGGG----GGGGSGYQSQGPICIGGGGGGGGGGGSGHQGQGSICIIGGGSGGGGSSGSG-------GMSMQQQ
Clustal Consensus  *** **   . *     **..**     ****.    *.*  * ***..**.    **:      *    ***.** ..*..          :  :*

                           730
                   ....|....|...
GALGA_LOR3         KQACYFPGQQK--
GALGA_LOR3B        KQACYFPGQQK--
PHUMI_LOR2         KQSSHWPCQQK--
GALGA_LOR1         TQPISWPPQTKHK
Clustal Consensus  .*.   :* * *
```

Figure 3.2 : Sequences of identified loricrin genes of chicken (GALGA; LOR1, LOR3, LOR3B) and ground tit (PUMI; LOR2)

The identified sequences of loricrins identified in chicken (GALGA; LOR1, LOR3, LOR3B) and ground tit (PHUMI; LOR2). LOR1 contains unique N- and C-terminal sequences and a unique repeat unit compared with other loricrins. LOR3 and LOR3B are identical in sequences and differ only in individual amino acid substitutions and number of repeat units. LOR2 is found only in passerine birds and is highly similar to LOR3/LOR3B with the exception of the identify of aromatic/aliphatic residues in the repeat and a small Cysteine-rich stretch of amino acids at the N-terminus.

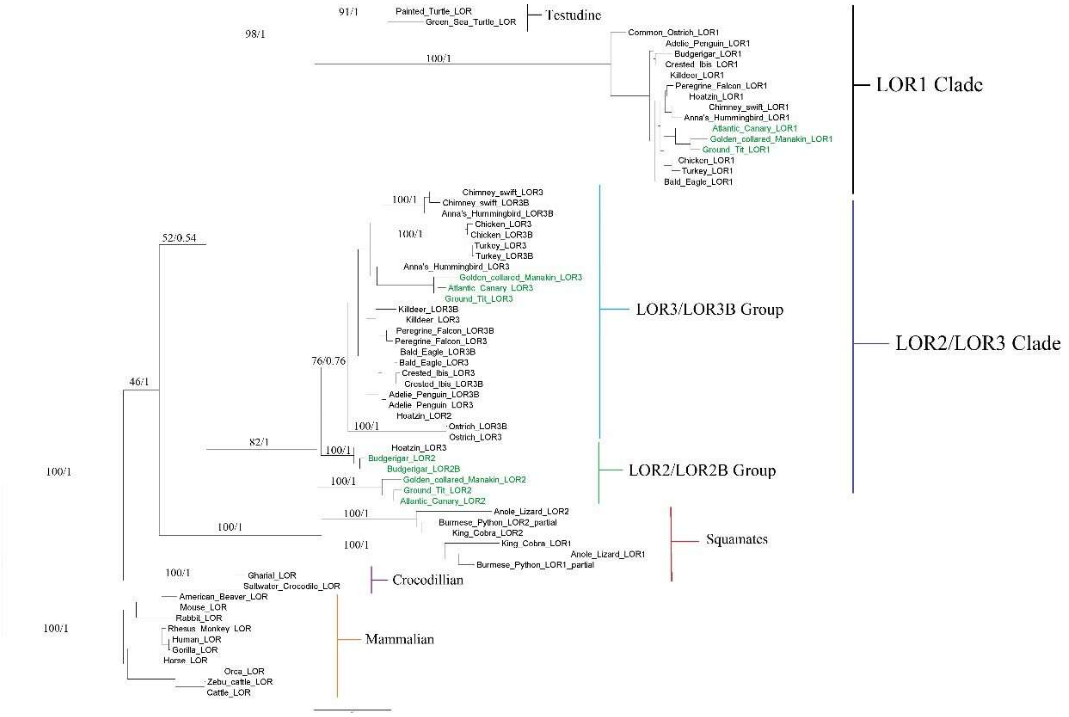


Figure 3.3 : Maximum Liklihood (ML) analysis of loricrin sequences. Phylogenetic tree generated by ML analyses. This tree is largely in agreeance with our baysian analysis tree (supplementary figure 3). Non-avian loricrins formed 3 distinct clades consisting of Mammals, Squamates, and crocodilian loricrin sequences respectfully. In contrast to currently accepted comprehensive phylogenetic data, our phylogeny places crocodilians as the basal group to all birds and reptiles. Avian loricrins were organized into 2 major clades. The First, LOR1 clade included all terminal avian loricrins that bordered EDYM1 annotated as LOR1 as well as testudine loricrins as a sister group. The second avian clade was LOR2/LOR3 clade which consisted of 2 major sister groups of LOR3 and LOR2 respectfully. Only passerine birds and the Hoatzin possessed LOR2 loricrins. All species possessed a LOR3 loricrin, and all species with the exception of Passeriformes, the hoatzin and Anna’s Hummingbird possessed a LOR3B gene organized in a lineage specific manner.

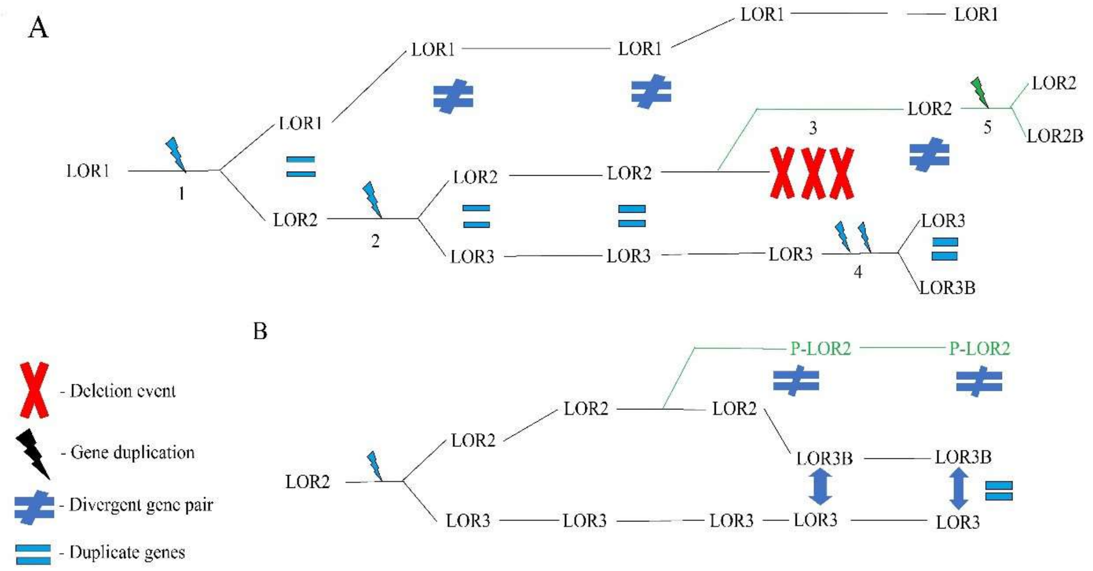


Figure 3.4 : schematic of possible scenario detailing evolutionary history of avian loricrins. **A.) (1)** Duplication of the ancestral loricrin gene (LOR1) resulted in 2 copies of loricrin (LOR1 and LOR2) before the emergence of the crown birds (Prum et al., 2015). **(2)** Duplication of LOR2 resulted in LOR2 and LOR3 genes. **(3)** Following the divergence of Passeriformes, deletion of LOR2 in other major orders of birds resulted in a single copy of loricrin (LOR3) in most orders of birds whereas LOR2 was retained in Passeriformes. **(4)** In non-passerine lineages, LOR3 duplicated and produced lineage specific LOR3B found in Palaeognathae, Galloanserae and Neoaves (excluding Passeriformes) species. **(5)** In the case of Psittacisformes, a suborder of passerine birds, the retained LOR2 duplicated and produced Psittaciforme-specific LOR2B in budgerigar. **B.)** Depicts a second scenario where the similarities between LOR3 and LOR3B are the result of concerted evolution of an ancestral duplication as opposed to similarity resulting from a recent duplication. Identical to scenario 1 until following duplication of LOR2. Ancestral LOR2 and LOR3 duplicate genes undergo gene conversion events resulting in concerted evolution. Evolution of LOR2 in Passeriformes (P-LOR2) resulted in its' divergence. Continued concerted evolution in non-passerine birds has maintained nearly identical loricrin paralogs. This scenario

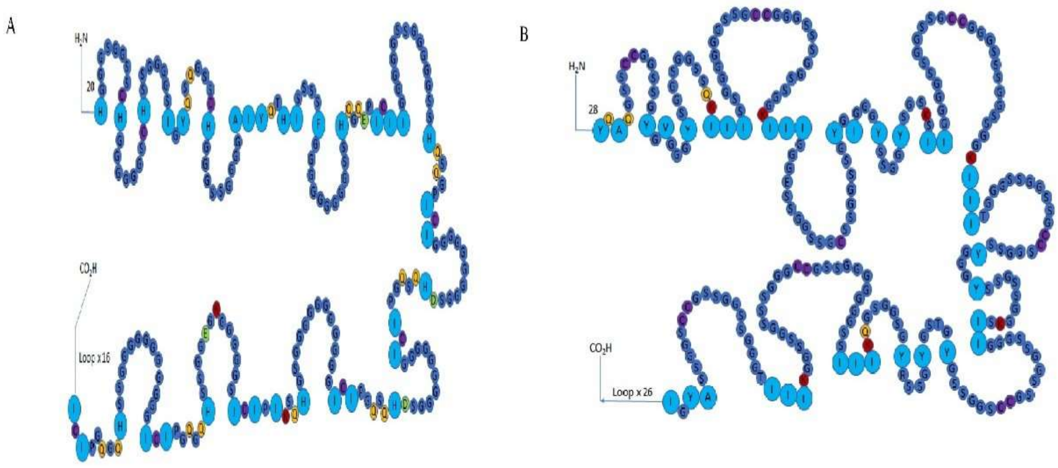


Figure 3.5 A + B : Schematic representation of the central domain of LOR Gly-loops in the chicken**. (A)** Chicken LOR1 likely contains extended arrays of glycine loops which, similarly to mammalian loricrins, are interspersed by glutamine rich domains of different structures. This representation does not infer any particular three-dimensional arrangement of the loops, but since mammalian loricrins are known to contain N-(γ-glutamyl) - lysine isodipeptide bonds, it is likely loricrins adopt a compact rosette-like structure. **(B)** There are 43 total predicted loops in GALGA LOR3. The loops range in size from 3 to 26 residues indexed on aromatic/aliphatic residues. Conserved adjacent cysteine residues are located at the apex of several of the larger loops and possible participate in disulfide bonding. This schematic does not infer any further three-dimensional structure of the loops. Glutamine and Lysine residues are also located at conserved positions throughout the sequence. Loops are generally indexed upon dimers/trimers of aliphatic residues or lone aromatic residues.

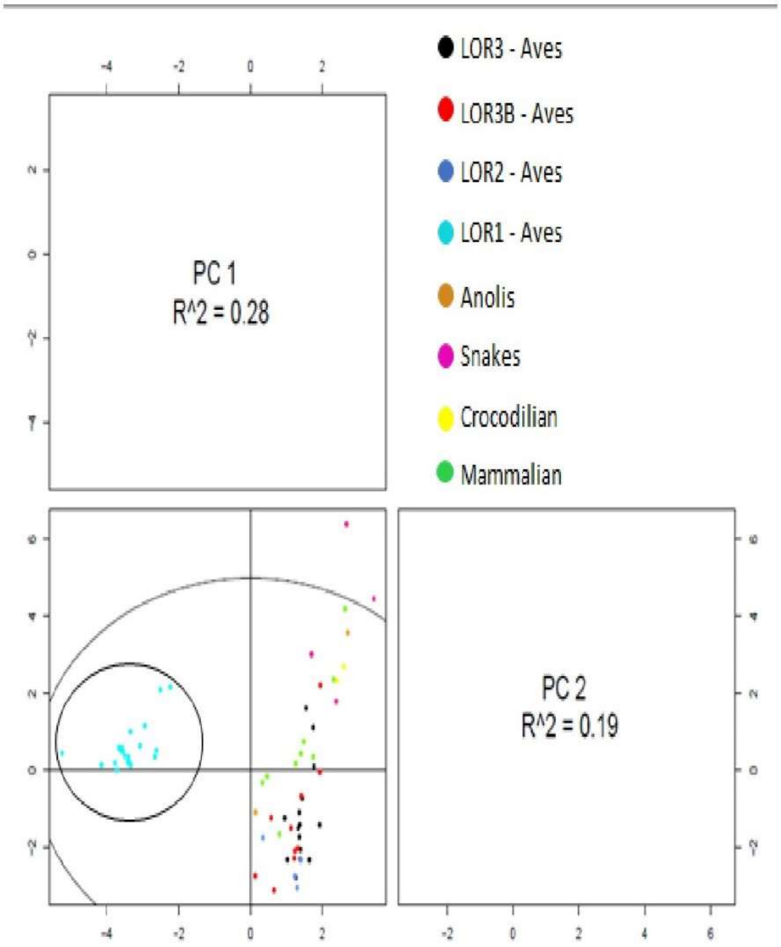


Figure 3.6 : Principle Component Analysis (PCA) of loricrin sequences. PCA generated in R, BiocLite-pcaMethods package using the singular value decomposition (SVD) method. Respective loricrin sequences are indicated by color as represented in the key. The black circle surrounds the avian LOR1 cluster. The amino acid contents of avian LOR1 were unique relative to all other loricrin sequences. All other loricrin genes including LOR2, LOR3 and LO3B of Aves failed to sort into distinct groups, highlighting the large amount of diversity observed across loricrin sequences. PCA plot was able to explain 46.79% (PC1 = 0.2764% , PC2 = 0.1915%) of the total variance between the amino acid contents of loricrin sequences.

Figure 7 – Loricrin Amino Acid Variation

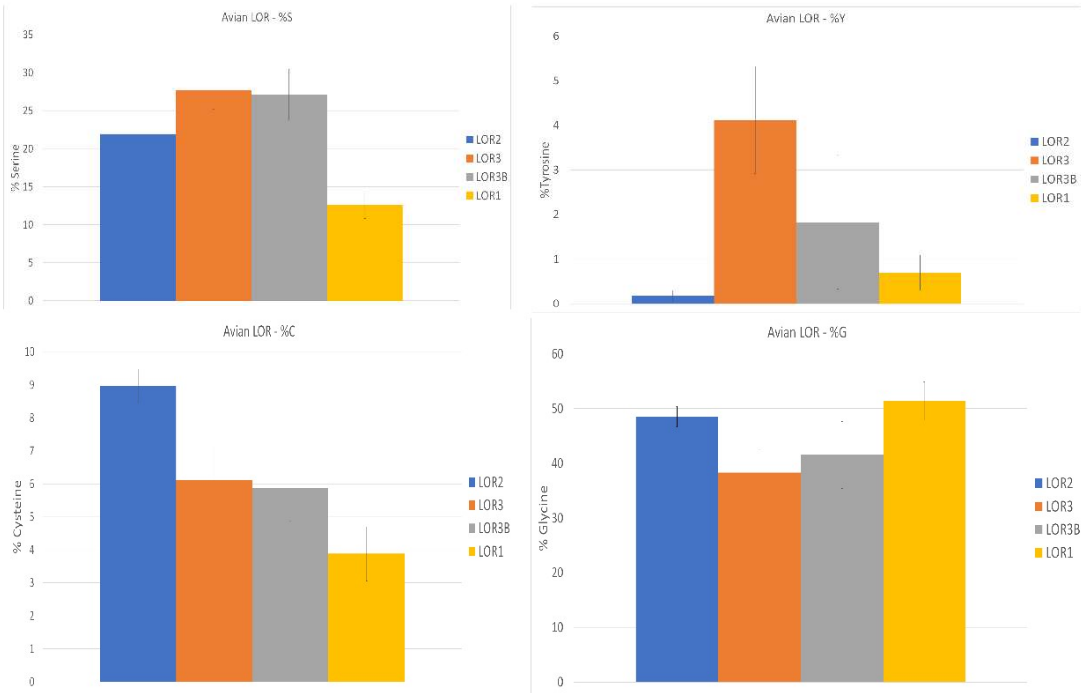


Figure 3.7 : Significant variation in amino acid residues associated with epidermal development and structure across avian loricrins. In clockwise order starting at the top left: the average percentage of serine (S), tyrosine (Y), Cysteine (C), and Glycine (G) across avian loricrins. For all 4 residues pictured, there were significant differences (**=$p<0.001$) observed in LOR2 and LOR1 from other loricrins. Data from AA analysis in supplementary table 4 A and B. Respective LOR orthologues distinguished by color and from left to right: LOR2, LOR3, LOR3B, LOR1.

3.7 References

1. Alibardi L, Holthaus KB, Sukseree S, Hermann M, Tschaler E, Eckhart L. (2016). Immunolocalization of a Histidine-Rich Epidermal Differentiation Protein in the Chicken Supports the Hypothesis of an Evolutionary Developmental Link between the Embryonic Subperiderm and Feather Barbs and Barbules. *PLOS One* 11(12): e0167789. doi:10.1371/journal.pone.0167789.

2. Alibardi L. (2017). Review: cornification, morphogenesis and evolution of feathers. *Protoplasma*. 254(3):1259-1281. doi: https://doi.org/10.1007/s00709-016-1019-2

3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J MOL BIOL*. 215(3):403-410. doi: https://doi.org/10.1016/S0022-2836(05)80360-2.

4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389-3402.

5. Backström N et al. (2010). The recombination landscape of the zebra finch Taeniopygia guttata genome. *Genome Res.* doi: 10.1101/gr.101410.109.

6. Barati D, Kader S, Shariati SRP, Moeinzadeh S, Sawyer RH, Jabbari E. (2017). Synthesis and Characterization of Phot-Cross-Linkable Keratin Hydrogels for Stem Cell Encapsulation. *Biomacromolecules*. 18(2):398-412. doi: 10.1021/acs.biomac.6b01493.

7. Burley SK, Petsko GA. (1985). Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science*. Vol. 229, 4708; 23-28. doi: 10.1126/science.3892686.

8. Cai Q, Qian X, Lang Y, et al. (2013). Genome sequence of ground tit Pseudopodoces humilis and its adaptation to high altitude. *Genome Biol*. 14(3):R29. doi:10.1186/gb-2013-14-3-r29.

9. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. (2009). Basic local alignment search tool. *BMC Bioinformatics.* 10:42. doi: 10.1186/1471-2105-10-421.

10. Candi E, Schmidt R, Melino G. (2005). The cornified envelope: a model of cell death in the skin. *Nat. Rev. Mol. Cell Bio*. 6(4): 328-340. doi:10.1038/nrm1619

11. Chuong CM et al. (2002). What is the 'true' function of skin? *Exp Dermatol.* 11:159-187. doi: https://doi.org/10.1034/j.1600-0625.2002.00112.x.

12. Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biol.* Lett. 8, 783-786. Doi: 10.1098/rsbl.2012.0331.

13. Daiquing, Liao. (1999). Concerted Evolution: Molecular Mechanism and Biological Implications. *Am. J. Hum. Genet.* 64(1): 24-30. doi: 10.1086/302221.

14. Dawson DA et al. (2007). Gene order and recombination rate in homologous chromosome regions of the chicken and a passerine bird. *Mol. Biol. Evol.* 24(7): 1537-1552. doi: 10.1093/molbev/msm07.

15. Eckhart L, Lippens S, Tschaler E, Declercq W. (2013). Cell death by cornification. *BBA-Mol. Cell Res.* 1833(12): 3471-3480. doi: https://doi.org/10.1016/j.bbamcr.2013.06.010.

16. Ericson PGP, Christidis L, Cooper A, Irestedt M, Jackson J, Johansson US, Norman JA. (2002). A Gondwanan origin of passerine birds supported by DNA sequences of the endemic New Zealand wrens. *P. Roy. Soc. B-Biol. Sci.* 269(1488): 234-241. doi: 10.1098/rspb.2001.1877.

17. Fankl C, Kuhl H, Weber M, Ralser M, Timmermann B, Gahr M. (2014). NCBI Serinus canaria Annotation Release 100. MPI Molgen, NCBI Eukaryotic Genome Annotation Pipeline. *NCBI Genome.* GCF_000534875.

18. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 31(13): 3784-3788.

19. Gelman A, Rubin DB. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science.* 7(4): 457-511.

20. Greenwold MJ, Bao W, Jarvis ED, Hu H, Li C, Gilbert MTP, Zhang G, SawyerRH. (2014). Dynamic evolution of the alpha (α) and beta (β) keratins has accompanied integument diversification and the adaptation of birds into novel lifestyles. *BMC Evolutionary Biology*. 14:249. doi: https://doi.org/10.1186/s12862-014-0249-1.

21. Greenwold MJ, Sawyer RH. (2013). Molecular Evolution and Expression of Archosaurian β-Keratins: Diversification and Expansion of Archosaurian β-Keratins and the Origin of Feather β-keratins. *J. Exp. Zool*. *Part* B. 320(6):393-405. doi: 10.1002/jez.b.22514.

22. Gremillet D, Chauvin C, Wilson RP, Meho YL, Wanless S. (2005). Unusual feather structure allows partial plumage wettability in diving great cormorants Phalacrocorax carbo. *J. Avian Biol.* 36(1): 57-63. doi: https://doi.org/10.1111/j.0908-8857.2005.03331.x.

23. Hall TA. (1999). BioEdit: a user-friendly biological sequence lignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp.* 41: 95-98.

24. Hohl D, Mehrel T, Lichti U, Turner ML, Roop DR, Steinert P. (1991). Characterization of Human Loricrin. *J. Biol. Chem.* 266(10) :6626-6636

25. Holland PWH, Garcia-Fernàndez J, Williams NA, Sidow A. (1994). Gene duplications and the origins of vertebrate development. *Development*. 125-133.

26. Holthaus KB, Mlitz V, Strasser B, Tschaler E, Alibardi L, Eckhart L. (2017). Identification and comparative analysis of the epidermal differentiation complex in snakes. *Sci. Rep.* 7, 45338 doi: 10.1038/srep45338.

27. Holthaus KB, Strasser B, Sipos W, Schmidt HA, Mlitz V, Sukseree S, Weissenbacher A, Tschaler E, Alibardi L, Eckhart L. (2015). Comparative genomics Identifies Epidermal Proteins Associated with the Evolution of the Turtle Shell. *Mol. Biol. Evol.* 33(3):726-737. doi: 10.1093/molbev/msv265.

28. Holthaus KB et al. (2018). Comparative analysis of epidermal differentiation genes of crocodilians suggests new models for the evolutionary origin of avian feather proteins. *Gen. Biol. Evol.* 10(2): 694-704. doi: 10.1093/gbe/evy035.

29. Hron T, Pajer P, Paces J, Bartunek P, Elleder D. (2015). Hidden genes in birds. *Genome Biol.* 16:164. doi: https://doi.org/10.1186/s13059-015-0724-z.

30. Huelsonbeck JP, Ronquist F. (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics*. 17: 754-755.

31. Hynes RO, Destree A. (1977). Extensive disulfide bonding at the mammalian cell surface. *P. Natl. A. Sci. USA*. 74(7); 2855-2859.

32. Ishida-Yamamoto A., Takahashi H., Iizuka H. (1998). Loricrin and human skin diseases: molecular bases of loricrin keratodermas. *Histol. Histopathol.* 13(3):819-826. doi: 10.14670/HH-13.819

33. Jarvis ED et al. (2014). Whole genome analyses resolve the early branches in the tree of life of modern birds. *Science.* 346(6215):1320-1331. doi: 10.1126/science.1253451.

34. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. (1998). Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* 23(10): 403-405.

35. John JA et al. (2012). Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biol.* 13(1): 415. doi:10.1186/gb-2012-13-1-415.

36. Kalinin AE, Kajava AV, Steinert PM. (2002). Epithelial barrier function: assembly and structural features of the cornified cell envelope. *BioEssays.* 24(9); 789-800. doi: https://doi.org/10.1002/bies.10144.

37. Kumar S, Stecher G, Tamura K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33: 1870-1874. doi: 10.1093/molbev/msw054.

38. Kypriotou M, Huber M, Hohl D. (2012). The human epidermal differentiation complex: cornified envelope precursors, S100 proteins and the'fused genes' family. *Exp. Dermatol.* 21(9):643-649. doi: https://doi.org/10.1111/j.1600-0625.2012.01472.x.

39. Milinkovitch MC, Helaers R, Depiereux E, Tzika AC, Gabaldon T. (2010). 2x genomes – depth does matter. *Genome Biol.* 11:R16. doi: https://doi.org/10.1186/gb-2010-11-2-r16.

40. Miller HC, Biggs PJ, Voelckel C, Nelson NJ. (2012). De novo sequence assembly and characterization of a partial transcriptome for an evolutionarily distinct reptile, the tuatara (Sphenodon punctatus). *BMC Genomics*. 13:439. doi: 10.1186/1471-2164-13-439.

41. Notredame C, Higgins DG, Heringa J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302:205-217. doi: 10.1006/jmbi.2000.4042.

42. Peona V, Weissensteiner MH, Suh A. (2018). How complete are 'complete' genome assemblies? – An avian perspective. *Mol. Ecol. Resour.* doi: 10.1111/1755-0998.12933.

43. Pierard G, Goffin V, Hermanns-Le T, Pierard-Franchimont C. (2000). Corneocyte desquamation. *Int. J. Mol. Med.* 6(2):217-238. doi:https://doi.org/10.3892/ijmm.6.2.217.

44. Prum OR, Berv JS, Dorburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature.* 534, S7-8. doi: 10.1038/nature15697.

45. Rambaut A. (2012). FigTree Phylogenetic viewing software. http://tree.bio.ed.ac.uk/software/figtree/.

46. Rice RH, Winters BR, Durbin-Johnson BP, Rocke DM. (2013). Chicken Corneocyte Cross-Linked Proteome. *J. Proteome Res.* 12(2): 772-776. doi: 10.1021/pr301036k.

47. Robinson NA, Lapic S, Welter JF, Eckert RL. (1997). S100A11, S100A10, Annexin I, Desmosomal Proteins, Small Proline-rich Proteins, Plasminogen Activator Inhibitor-2, and Involucrin are Components of the Cornified Envelope of Cultured Human Epidermal Keratinocytes. *J. Biol. Chem.* 272:12035-12046. doi: 10.1074/jbc.272.18.12035.

48. Ronquist F, Huelsenbeck JP. (2003). MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 19: 1572-1574.

49. Rose GD, Roy S. (1980). Hydrophobic basis of packing in globular proteins. *Proc. Natl. Acad. Sci.* 77(8): 4643-4647.

50. RStudio Team. (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

51. Sawyer S. (1989). Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* 6(5): 526–538. doi: https://doi.org/10.1093/oxfordjournals.molbev.a040567.

52. Singh J, Thornton JM. (1985). The interaction between phenylalanine rings in proteins. *FEBS Lett.* 191(1): 1-6. doi: https://doi.org/10.1016/0014-5793(85)80982-0.

53. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. (2007). pcaMethods – a Bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23(9):1164–1167. doi: 10.1093/bioinformatics/btm069.

54. Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 30(9):1312-1313. doi: 10.1093/bioinformatics/btu033.

55. Steinert P, Mack J, Korge B, Gan SQ, Haynes S, Steven A. (1991). Glycine Loops in Proteins: their occurrence in certain intermediate filament chains, loricrins and single-stranded RNA binding protiens. *Int. J. Biol. Macromol.* 13(3):130-139. doi: https://doi.org/10.1016/0141-8130(91)90037-U.

56. Strasser B et al. (2014). Evolutionary Origin and Diversification of epidermal barrier proteins in amniotes. *Mol Biol Evol*. 31(12): 3194-3205. doi: 10.1093/molbev/msu251.

57. Strasser B, Miltz V, Hermann M, Tschachler E, Eckhart L. (2015). Convergent evolution of cysteine-rich-proteins in feathers and hair. *BMC Evol Biol*. 15:82. doi: https://doi.org/10.1186/s12862-015-0360-y.

58. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25(24): 4876-4882.

59. Vanhoutteghem A, Djian P, Green H. (2008). Ancient origin of the gene encoding involucrin, a precursor of the cross-linked envelope of epidermis and related epithelia. *P. Natl. Acad. Sci-Biol.* 105(40): 15481-15486. doi: https://doi.org/10.1073/pnas.0807643105.

60. Völker M et al. (2010). Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome Res.* doi: 10.1101/gr.103663.109.

61. Wang H, Parry D, Jones LN, Idler WW, Marekov LN, Steinert P. (2000). In Vitro Assembly and Structure of Trichocyte Keratin Intermediate Filaments: A Novel Role for Stabilization by Disulfide Bonding. *J. Cell Biol.* 151(7): 1459-1468.

62. Yang F, Zhao G, Zhou L, Li B. (2015). Complete mitochondrial genome of White-rumped Munia Lonchura striata swinhoei (Passeriformes: Estrildidae). *Mitochondrial DNA.* 27(4): 3028-3029. doi: 10.3109/19401736.2015.1063052.

63. Zhang G et al. (2014) Comparative Genomics Reveals Insights into Avian Genome Evolution and Adaptation. *Science*. 346(6215): 1311-1320. doi: 10.1126/science.1251385.

64. Zhang J. (2003). Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18(6): 292-298. doi: 10.1016/S0169-5347(03)00033-8.

65. Zhu BY, Zhou NE, Kay CM, Hodges RS. (1993). Packing and hydrophobicity effects on protein folding and stability: Effects of β-branched amino acids, valine and isoleucine, on the formation and stability of two-stranded α-helical coiled coils/leucine zippers. *Protein Sci.* 2(3): 383-394. doi: 10.1002/pro.5560020310.

CHAPTER 4

EVOLUTION OF AN EPIDERMAL DIFFERENTIATION COMPLEX GENE FAMILY FROM A COMMON ARCHOSAUR ANCESTOR

4.1 Abstract

The transition of amniotes to a fully terrestrial lifestyle involved the adaptation of major molecular innovations to the epidermis, often in the form of epidermal appendages such as hair, scales and feathers. Feathers are diverse epidermal structures of birds, and their evolution has played a key role in the expansion of avian species to wide range of lifestyles and habitats. Like other epidermal appendages, feather development is a complex process which involves many different genetic and protein elements. In mammals, many of the genetic elements involved in epidermal development are located at a specific genetic locus known as the Epidermal Differentiation Complex (EDC). Studies have identified a homologous EDC locus in birds which contains several genes expressed throughout epidermal and feather development. A family of avian EDC genes rich in aromatic amino acids which also contain MTF motifs (EDAAs/EDMTFs) that includes the previously reported Histidine-rich or fast-protein (HRP/fp), an important marker in feather development, has expanded significantly in birds. Here we characterize the EDAA gene family in birds and investigate the evolutionary history and possible functions of EDAA genes using phylogenetic and sequence analyses. We provide evidence that the EDAA gene family originated in an early archosaur ancestor, and has since expanded in birds, crocodiles and turtles respectively. Furthermore, this study shows that the respective amino acid compositions of avian EDAAs are characteristic of structural functions associated with EDC genes and feather development. Finally, these results support the hypothesis that the genes of the EDC have evolved as the result of tandem duplications and diversification resulting in neofunctionalization and expansion of respective gene families, including the EDAA/EDMTF gene families.

4.2 Introduction

The adaptation of novel and complex appendages such as hair, scales and feathers, were critical in the evolution of amniotes into a variety of terrestrial lifestyles (Sawyer and Knapp 2003, Alibardi 2003, Prum 2005). The epidermal appendages of amniotes exhibit a wide range of physical properties which serve a variety of functions including but not limited to thermoregulation, camouflage and mating (Chuong et al. 2002). Generally, epidermal appendages form as the result of spatiotemporal interactions between cells of the epidermis and underlying dermis, and the process involves several different genetic elements (Haake et al. 1984, Alibardi 2016). While the specific elements and processes involved in the development of epidermal appendages vary, evidence suggests that they all evolved from a single or small number of conserved ancestral gene(s) (Strasser et al. 2014). In amniotes such as mammals and reptiles, many of the genes encoding proteins involved in the mechanically resilient structure of epidermal appendages are found at a specific genetic locus known as the Epidermal Differentiation Complex (EDC) (Kypriotou et al. 2012, Strasser et al. 2014, Holthaus et al. 2015, Holthaus et al. 2017, Holthaus et al. 2018).

One major reason for the evolutionary success of amniotic skin appendages is their unique and mechanically resilient physical properties (Eckhart et al. 2013). To serve their various purposes, skin appendages tend to have increased tensile, flexural and yield strengths relative to the epidermis proper or internal organs, all of which have significant impacts on the physical characteristics exhibited by skin appendages (Velasco et al. 2009). These unique properties are largely the result of the evolution of novel and complex developmental processes which make use of structural proteins capable of covalently crosslinking with themselves and one another, often through transglutamination and disulfide bonding (Hynes and Destree 1977, Sawyer and Knapp 2003). Studies have shown that differences in physical properties of different skin appendages can be correlated with differences in their respective amino acid contents. For example, Fujimoto et al. (2014) found that the number of disulfide bonds formed by keratin-associated proteins, enabled them to adhere to various structural proteins that they do not normally form associations with indicating that the number and positions of conserved cysteine residues would have a direct effect on the identity of the proteins involved in epidermal structure. These results suggest that differences between specific amino acid residues which are likely involved in protein crosslinking in structural genes could influence overall physical characteristics of the appendage in question.

The feathers of birds display a wide range of physical properties which have allowed birds to expand and survive in diverse environments across every continent including Antarctica (Li et al. 2014). Feathers were a critical adaptation in the evolution of avian flight, and the diversity observed across different species of birds' feathers are a major reason for their ecological success. Like other epidermal appendages, many the genes involved in the development and structure of feathers are located within the EDC locus and originated from a single or small number of ancestral genes (Strasser et al. 2014). The physical diversity observed across feathers is accompanied by the genetic diversity displayed by several differentially expressed avian EDC genes.

The avian EDC was first identified in the chicken (*Gallus gallus*) and was found to contain several genes which were characteristic of epidermal development and structure (Strasser et al. 2014). Several studies on the conservation of specific EDC genes identified in the chicken such as Epidermal Differentiation Cysteine Rich Protein (EDCRP), Epidermal differentiation protein containing DPCC motifs (EDDM) and Epidermal differentiation protein with an MTF motif rich in histidine (EDMTFH) have found that the EDC region as well as some specific genes are conserved across a broader range of avian species (Strasser et al. 2015, Alibardi et al. 2016, Lachner et al 2019). These studies found that while these genes were conserved across a broad range of avian species, there was significant sequence variation present. Moreover, studies on loricirns, a major component of the mammalian cornified envelope, in birds found that intragenic duplications of repetitive units has resulted in huge disparities in gene size, as well as a complex evolutionary history (Davis et al. 2019).

Both intragenic and whole gene duplication has been shown to play major roles in the evolution of genetic diversity as well as in that of novel form and function (Nam et al. 2010). The EDC locus has been found to have likely evolved through tandem gene duplication and diversification resulting in subfunctionalization, and ultimately neofunctionalization (Strasser et al. 2014). Furthermore, studies have found that β-keratins, the primary protein component of the barbs and barbules of mature feathers, have diversified into several distinctly conserved subfamilies which have expanded outside of the EDC to other parts of the genome however they likely originated from ancestral genes within the EDC locus (Greenwold et al. 2014).

In contrast to EDCRP, EDDM and loricrins which have evolved largely through intragenic duplications of repetitive units, other avian EDC genes represent members of conserved multigene families such as Epidermal Differentiation proteins containing cysteine histidine motifs (EDCHs) and Epidermal Differentiation proteins rich in aromatic amino acids and containing MTF motifs (EDAA/EDMTFs). These genes were originally identified and annotated by Strasser et al. 2014 as only EDMTFs, however, the conserved “MTF motif” identified does not infer any specific functional motif, rather that the amino acid sequence of M-T-F was highly conserved in these genes. We also found in many avian species the Phenylalanine (F) residue of the MTF motif was substituted with another aromatic amino acid, often Tyrosine (Y). Furthermore, related genes identified in crocodilians and turtles lack the MTF sequence found in birds but are also rich in aromatic amino acids, therefore we refer to these genes as both EDAA/EDMTFs to avoid confusion. The EDAA/EDMTF genes are short sequences of less than 125 amino acids which have been shown to be differentially expressed in developing feathers and scales of the chicken (Strasser et al. 2014, Alibardi et al. 2016). Previous studies have found that EDAA/EDMTF genes are conserved across a diverse set of avian species as well as in crocodilians and turtles, however little is known of their evolutionary histories, functions and conservation across a wider range of birds (Holthaus et al. 2015, Holthaus et al. 2018). It was also found that avian EDMTFH matches the sequence of the previously reported Histidine-rich or Fast-protien (HRP/fp) described by Barnes and Sawyer (1994) which plays an important role in epidermal development.

It is known that the evolution and expansion of the β-keratin multigene family, which originated within the EDC, was critical in the evolution of avian feathers (Greenwold et al. 2014). Studies focusing on other conserved multigene families within the avian EDC would likely provide greater insight into the evolution of large conserved groups of genes as well as their roles in the adaptation of novel structures such as feathers. In this study, we use phylogenetic and statistical analyses to more closely examine the evolution and conservation of the EDMTF genes in birds, as well as gain a better understanding of their possible functions in epidermal development. Furthermore, we provide a hypothesis that the evolution of novel structures such as feathers has largely been accompanied by the tandem duplication and diversification of EDC genes such as the EDAA/EDMTF gene family.

4.3 Methods

4.3.1 Identification and of EDAA/EDMTF gene families

Avian EDAA/EDMTF genes were identified by Blast+, specifically the *tblastn* command which searches a nucleotide database using amino acid sequences as queries (Altshcul et al. 1990, Gish and States 1993). The amino acid sequences of chicken EDAA/EDMTF genes were used as the initial blast query, however each identified sequence was added back to the query file and reciprocal rounds of blast searches were performed. In order to ensure no genes were missed, we used manual genomic screening methods which entailed extracting entire genomic regions between two identified genes, and manually scanning the nucleotides for evidence of EDC genes not found by blast.

Suspected EDAA/EDMTF sequences were extracted as nucleotide Fasta files and translated to amino acids using ExPasy Translate online analysis tool (Gasteiger et al. 2003). Translated amino acid sequences were characterized via multiple sequence alignment to chicken and other identified EDAA/EDMTF genes using ClustalW online anlysis tool (Thompson et al. 1997). To determine genomic orientation and total number of EDAA/EDMTF genes in birds, manual screening was performed of syntenic genomic regions was performed for all species. Supplemental table 4.1 details all identified

EMDTF genes using the chicken as reference. Genes were considered complete if both N and C termini with start and stop codons were present as well as the minimal presence of unknown nucleotides. Supplemental table 4.1 legend details the status and justification for all EDAA/EDMTF genes. Genes were considered incomplete if; 1 - there were persistent unknown nucleotides within the coding sequence, 2 – there was a frameshift present in the sequence that could be resolved by switching reading frames, 3 – no start codon was observed, 4 – no stop codon was observed, 5 – there was significant misalignment with reference sequences (i.e. no conserved elements of the gene in question were identified via alignment), and 6 – There was a stop codon interrupting the ORF. The scores in supplemental table 4.1 indicate the alignment score of each respective gene when aligned with that of the Chicken (*Gallus gallus*).

Figures 4.1, 4.2 and 4.3 were aligned using ClustalW online analysis tool and figures were created and annotated using Microsoft Powerpoint. Architecture and orientation of avian EDAA/EDMTF loci were analyzed using chicken genes identified by Strasser et al. 2014 as references. Identified genes were annotated based upon their position and genomic orientation corresponding to the chicken. Extra identified EDAA/EDMTF genes in addition to those in the chicken were also annotated based upon position and orientation. For example, the additionally identified EDAA/EDMTF gene in the chicken was annotated as EDMTF5 because it was located between EDMTF2 and EDMTF1 but presented different chromosomal orientation. In contrast, the additional genes identified in the Cuckoo were annotated as EDMTF1b and EDMTF1c because they were located adjacent to EDMTF1 and in the same chromosomal orientation suggesting they are recent tandem duplications.

4.3.2 Phylogenetic analysis of EDAA/EDMTF gene family

Phylogenetic analysis of avian EDAA/EDMTF genes was done using both Bayesian and Maximum Likelihood (ML) methods. Alignments of EDAA/EDMTF amino acid sequences were generated using ClustalW2 local alignment tools (Thompson et al. 1997) and the alignments were edited using Bioedit (Hall 1999). MEGA7 sequence analysis software (Kumar et al. 2016) was used and identified PROTGAMMAJTT as the best fit substitution model based on Bayesian Information Criterion (BIC), Akaike information Criterion corrected (AICc) and the substitution rate ($BIC^{JTT} = 3849.826$, $AICc^{JTT} = 2815.627$). Bayesian analysis was done using Mrbayes-v3.2 (Huelsenback and Ronquist 2001, Ronquist and Huelsenbeck 2003) and was run for 10,000,000 generations and was checked for convergence using the Potential Scale Reduction Factor Method (PSRF) (TL:PSRF=1.0 ; alpha: PSRF=1.0). ML analysis was performed using RAxML-v8.2.10 (Stamatakis 2014) utilizing MRE-based bootstrapping until convergence was detected, followed by inferring the best tree produced out of 1000 generated ML trees, and finally mapping the MRE bootstrap values on the identified best tree. Sequences of EDAA/EDMTF genes from crocodilians and turtles identified by Holthaus et al. (2015) and (2018) respectively were used as outgroups in both analyses. Avian sequences used in phylogenetic analyses were selected to represent a phylogenetically diverse group of bird species and lifestyles. All sequences used were considered complete and lacked unknown nucleotides. All sequences used in phylogenetic analysis are listed in supplemental table 4.2. Trees were edited and viewed using FigTree-v1.4.3 (Rambaut 2012).

4.3.3 Amino acid composition and principle component analysis (PCA)

Amino acid analyses of avian EDAA/EDMTF genes was done using ExPasy ProtParm online analysis tools (Gasteiger et al. 2005). The total numbers as well as overall percentages of each amino acid residue making up the ORFs of avian EDAA/EDMTF genes were calculated. The sequences used in amino acid analyses can be found in an accessory data table, supplemental table 4.1.1, since this is too much raw data to present in the dissertation it can be obtained upon request. To compensate for variation in the size of sequences across different species, we used the total percentage of each amino acid residue instead of the number. All sequences used were complete and contained no unknown nucleotides. Our overall amino acid composition analyses included 22 EDMTFH genes, 27 EDMTF4 genes and 62 EDMTF1-3/5 genes from 32 avian species.

Statistical analyses examining significant differences in amino acid contents of EDAA/EDMTF genes across different species, lifestyles and subfamilies was done using standard Single Factor Analysis of Variance (ANOVA) tests with the Microsoft excel data analysis ToolPak. This ANOVA test was selected due to the small sample size available in the analyses. Principle Component Analysis (PCA) was done in R using the BiocLite-pcaMethods package by BioConductor (R Core Team 2013, Smyth 2005) using the Singular Value Decomposition (SVD) method (Gerbrands 1981).

4.4 Results

4.4.1 The EDAA/EDMTF gene family is conserved in the avian EDC

To better understand the evolution and function of the EDAA/EDMTF gene family, we screened the genomes of 48 phylogenetically diverse avian species for their presence using BLAST+ and manual genomic screening methods. We identified 3 major groups of EDAA/EDMTF genes across the birds investigated, the previously investigated EDMTFH (HRP) genes, EDMTF4s and finally EDMTF1-3/5+. These genes are annotated as they are described by Strasser et al. (2014). As expected, several genes identified were either partial or contained unknown sequence artifacts. Incomplete or partially identified genes were only used as evidence for presence or absence of a specific genes and were excluded from amino acid and phylogenetic analyses. Each of the 3 major classes of EDAA/EDMTF genes are characterized by distinct conserved sequence elements, genomic orientations and amino acid contents, however there is considerable variation observed across different groups.

EDMTF4 is generally characterized by highly conserved aspartic acid (D) residues in the N-terminal and central domains as well as the presence of several conserved tyrosine (Y) and glycine (G) throughout the gene (Figure 4.1 A). While EDMTF4 is conserved across all birds investigated, we found that EDMTF4 of the chicken and turkey contain several conserved histidine residues not present in other species, resulting in much greater conservation in EDMTF4 sequence when the chicken and turkey are excluded from the analysis (Figure 4.1 B). We found evidence for EDMTF4 in all 48 species investigated, however we identified partial or incomplete copies in 9 species (supplemental table 4.1). The table shows the presence of EDAA/EDMTF genes across birds investigated, their alignment score relative to the corresponding gene in the chicken, as well as a descriptor if there was a problem or the gene was only partially found.

Previous studies had identified that the sequence of EDMTFH matched that of the previously reported Histidine-rich protein (HRP), and it was conserved across a wide range of avian species (Alibardi et al. 2016). Our results confirmed the presence of EDMTFH in all species investigated by Alibardi et al. (2016), however, we did not identify any EDMTFH genes in passerine birds except for the Golden-collared Manakin (*Manacus vitellinus*). Evidence of EDMTFH was identified in all the remaining 41 species, with 3 of those being partial or incomplete (supplemental table 4.1). As reported by Alibardi et al. (2016), only EDMTFH of the chicken and turkey was rich in histidine resulting in sequence variation, however all EDMTFH genes identified contained the highly conserved sequence '-PYGYRsFGsLYGNRG-' within their central domains (Figure 4.2 A - alignment). Outside of Galliformes, EDMTFH sequence was highly conserved across all species investigated, except for the passerines (Figure 4.2 B - alignment).

The final group of EDAA/EDMTF genes identified were EDMTF1-3/5. These genes were highly conserved across closely related species, and in many cases appeared to represent species specific paralogs indicating a complex evolutionary history or possible concerted evolution. The most highly conserved elements of these genes across all species investigated were the presence of '-YQNQxED-' in the N-terminal region and '-RYSYGS-' in the C-terminal region, however there is variation present across different species in the exact amino acid content and gene lengths, specifically in those of the Galliformes (Figure 4.3). All species except for the Brown Mesite (*Mesitornis uniclolor*) contained at least a single copy of these genes. 36 of the 48 species were found to contain the genes EDMTF1 and EDMTF3, however were missing any additional copies. Specifically, these species were missing the gene annotated as EDMTF2 in the chicken. We did identify evidence of genes corresponding to the EDMTF2 position of the chicken in the Golden-collared Manakin (*Manacus vitellinus*), the Dalmatian Pelican (*Pelicanus crispus*), Common cuckoo (*Cuculus canorus*) and Ostrich (*Struthio camelus*). Furthermore, we identified an additional copy of EDMTF, annotated EDMTF5 in the chicken (*Gallus gallus*) and 2 additional copies in the Cuckoo annotated as EDMTF1b and EDMTF1c. These genes were annotated based on their sequence elements and genomic orientation and are indicated in the table as "+ genes" (supplemental table 4.1).

The overall conservation of the EDAA/EDMTF gene family in 5 phylogenetically diverse birds is presented in Figure 4.4. Our results demonstrate that the EDAA/EDMTF gene family is conserved across birds, but with considerable variation. We found that there is variation in the overall size of this region across different avian EDC loci that corresponds to the number of genes found. For example, in the chicken and cuckoo who contain additional copies of EDMTF genes, this region of the EDC contains 20,913 and 28,784 base pairs between EDMTF4 and EDMTF3 respectively. In contrast, this EDC region of the Bald Eagle, Adelie Penguin and Zebra Finch, which only possess EDMTF4/1/3 are 13,249, 14,562, and 12,422 base pairs in length respectively (Figure 4.4).

4.4.2 The EDAA/EDMTF gene family originated in a common archosaur ancestor

To investigate the evolutionary history of the EDAA/EDMTF gene family in birds and its role in the adaptation of complex appendages such as feathers and scales, we performed phylogenetic analyses using Bayesian and Maximum-Likelihood (ML) methods. Recent studies have identified homologous EDAA genes in the EDC loci of both crocodilians and turtles, and several of these genes were included in our analyses (Holthaus et al. 2015, Holthaus et al. 2018). In total we examined 149 EDAA/EDMTF genes including 108 avian genes from 28 different species, 22 from the Painted turtle (*Chrysemys picta*) as well as 19 from the two crocodilian species the American alligator (*Alligator mississppiensis* – 7 genes) and the Saltwater crocodile (*Crocodylus porosus* – 12 genes) (supplemental table 4.2).

In both ML and Bayesian analyses apart from EDAA10 of the Painted turtle, the EDAA genes of the crocodilians and turtles formed a large monophyletic clade with overall strong support and hence were selected as the outgroup. Our results confirmed the presence of 3 major groups of avian EDAA/EDMTF genes, EDMTFH, EDMTF4 and then the additional EDMTF1-3/5 genes (Figures 4.5 and 4.6). In both analyses, EDMTFH formed a monophyletic clade with strong support values. EDMTF4 and EDMTF1-3/5 genes form a large clade, with EDMTF4 representing a basal paraphyletic group and EDMTF1-3/5 making up a monophyletic subclade, however the support values associated with these groups are low. Interestingly, EDAA10 of the Painted turtle formed a monophyletic clade with EDMTFH in our Bayesian analysis, whereas in our ML analysis it was observed within the EDMTF4 paraphyletic group further highlighting the ambiguity associated with the low support values between the EDMTFH and EDMTF4 clades.

In both ML and Bayesian analyses, the EDMTF1-3/5 genes form a large monophyletic group (Figure 4.5 and 4.6). Within this group, the genes display a lineage-specific distribution like that observed in avian loricrins (Davis et al. 2018). The EDMTF genes of the Galliformes and Passerines form respective monophyletic groups within the major clade while all remaining avian EDMTF1-3/5 genes form a paraphyletic group. This distribution largely agrees with the currently accepted species phylogeny of birds and our own observations which show the sequences of EDMTF genes in Galliformes and Passeriformes contain unique amino acid contents relative to those of other species.

To better understand the origin of EDAA/EDMTF genes in birds as well as archosaurs in general, we further examined the evolutionary relationship of the avian EDMTFH and EDMTF4 genes once again using the EDAA genes of crocodilians as the outgroup (Figure 4.7 – EDMTF4/EDMTFH phylogeny). Interestingly 2 crocodilian genes, EDAA9 of the American alligator and EDAA12 of the Saltwater crocodile, were present within the EDMTF4 paraphyletic group. In the overall gene trees, these were the only crocodilian genes located outside of the crocodilian monophyletic group and instead were found within the turtle group (Figures 4.5 and 4.6). As in our previous analysis EDMTFH formed its own monophyletic group but was also part of a larger monophyletic clade with EDMTF4. This was in contrast with the previous analysis of all EDAA/EDMTF genes, where EDMTFH and EDMTF4 formed a paralogous clade which excluded EDMTF1-3/5 genes. The general support values for this tree were higher than those of the previous trees. All major branches contained values of 1.0 and the lowest support value observed was 0.5443 (Ach_EDMTF4) and was described for a terminal branch. All avian genes within the respective EDAA/EDMTF groups contained distinct groupings of the genes of Galliformes and Passeriformes respectively, and was largely in agreeance with the current avian species phylogeny proposed by Jarvis et al. (2012).

4.4.3 EDAA/EDMTF genes contain amino acid contents indicative of epidermal development and structure

Previous studies have demonstrated that the avian EDAA/EDMTF genes are differentially expressed in developing chicken epidermal tissues (Strasser et al. 2014). It is also known that the amino acid contents of several other avian EDC genes vary significantly across different species (Davis et al. 2018). Furthermore, amino acid composition of genes correlates directly with their general function. To gain a better understanding of their possible function or functions epidermal development of avian appendages, we analyzed the respective amino acid contents of the EDAA/EDMTF and performed statistical analyses such as principle component analyses (PCA). Like previous studies examining avian loricrins, due to the variation in overall size of the coding sequences of respective EDAA/EDMTF genes across different species their amino acid contents were analyzed as a percentage of specific residues instead of exact number. In order to ensure accuracy in our analyses, only complete genes containing no unknown residues (XXXs) were used. The raw data for amino acid analyses can be obtained upon request in a supplementary data file.

We analyzed the 3 main groups identified by phylogenetic analysis and found that all avian EDAA/EDMTF genes are rich in amino acid residues associated with epidermal structure and development processes (Strasser et al. 2014, Strasser et al. 2015, Alibardi et al. 2016, Lachner et al. 2019). The most abundant amino acid residues across all 3 groups were Tyrosine (Y), Glycine (G), Serine (S) and Cysteine (C) (supplemental table 4.3). EDMTFH and EDMTF4 contained similar amino acid contents with Tyrosine and Glycine making up 41.82% (Y=20.37%, σ=3.44; G=21.45%, σ=3.32) and 49.12% (Y=21.76%, σ=2.36; G=27.36%, σ=2.22) of each respective gene. The main difference between the amino acid contents of EDMTFH and EDMTF4 was the presence of increased Cysteine in EDMTF4 (EDMTF4:7.05%, σ=1.71 ; EDMTFH: 1.59%, σ=0.92). Both genes contained similar average Serine contents (EDMTF4=8.93%, σ=2.08 ; EDMTFH=8.47%, σ=2.16). EDMTF1-3/5 also was found to contain a very high Tyrosine content confirming all genes were indeed rich in aromatic amino acids (Y=22.19%, σ=4.3). In contrast to EDMTFH and EDMTF4 however, EDMTF1-3/5 was found to contain less Glycine (G=7.5%, σ=2.19), as well as higher amounts of Serine (S=15.48, σ=3.55) and Cysteine (C=15.17%, σ=2.67)(supplemental table 4.3).

Alibardi et al. (2016) found that the amino acid content of the EDMTFH gene was significantly different in the Galliformes, chicken and turkey, than in any other species. Specifically, Galliforme EDMTFH was rich in Histidine, whereas other avian EDMTFH genes contained little or no Histidine however all EDMTFH genes were rich in aromatic amino acids. We found that a similar difference is observed in the EDMTF4 amino acid composition of Galliformes relative to other avian species. Specifically we observed significant differences in amino acid contents of Cysteine (C)(Galliforme C=1.95%, σ=0.071, n=2 ; Other C=7.456%, σ=0.904, n=25 ; $F_{25,2}$=71.512, p<0.001), Histidine (H)(Galliforme H=8.75%, σ=1.77, n=2; Other H=0.172%, σ=0.43, n=25 ;

$F_{25,2}$=450.8799, p<0.001) and Glycine (G)(Galliforme G=23.25%, σ=3.182, n=2 ; Other G=27.77%, σ=1.84, n=25 ; $F_{25,2}$=9.89, p<0.005).

We also performed a principle component analysis to further examine the differences observed between the amino acid compositions of avian EDAA/EDMTF genes. In this analysis we also included the respective lengths of each gene as variables along with the amino acid residue percentages. The resulting PCA was graphed using 2 principle components which together described 52% of the total variation observed, however PC1 was considerably more significant than PC2 (R^2 PC1= 0.41, PC2=0.11) (Figure 4.8-PCA). Our results confirmed that the 3 major groups of avian EDAA/EDMTF genes contained unique amino acid compositions. Furthermore, they demonstrated that the amino acid contents of EDMTF1-3/5 genes are significantly different from those of EDMTFH and EDMTF4, who possess similar amino acid contents. We observed 10 data points across all genes which displayed significant variation and could be considered to deviate from their respective groups (Figure 4.8). All but 2 of these 10 data points can be attributed to the significant diversity observed in the EDAA/EDMTF genes of Galliformes. We did not identify any significant groupings of respective genes based on avian lifestyles, however due to the limited number of complete genes identified from aquatic and predatory birds, more data is needed to further examine the possibility of correlation between amino acid contents and lifestyle.

Finally, following the results of our phylogenetic analysis which demonstrated that EDMTFH and EDMTF4 are basal to other avian EDMTF genes, we looked for any trends in changes in their amino acid contents, specifically those associated with epidermal development and structure. We observed an interesting trend relating to increasing Cysteine content of EDAA/EDMTF genes (supplemental table 4.3) (Figure 4.9). While our results could not establish without a shadow of a doubt if EDMTFH or EDMT4 was basal, they clearly demonstrate that as the genes have evolved, their cysteine contents have increased.

4.5 Discussion

In this chapter we identified and characterized the EDAA/EDMTF gene family across a phylogenetically diverse set of avian species. Our results found that the EDAA/EDMTF gene family is conserved across birds and are rich in amino acid residues associated with epidermal structure and development. Furthermore, that it is likely to have originated in a common archosaur ancestor and since expanded in birds and other reptiles respectively. These results provide new insights into the function of specific EDC genes, as well as how the evolution and expansion of EDC genes has accompanied the adaptation of novel and complex skin appendages such as feathers.

Using genome screening, we identified EDAA/EDMTF homologs in every avian species investigated, however there was variation in the number and identity of EDAA/EDMTF genes present. Previous studies identified 5 total EDAA/EDMTF genes in the chicken annotated as EDMTFH, EDMTF4, EDMTF1, EDMTF2, and EDMTF3 (Stasser et al. 2014). We identified an additional duplicate of EDMTF1/2/3 in the chicken which was not previously reported which we annotated as EDMTF5. We found that of the 5 EDAA/EDMTF genes identified in the chicken by Strasser et al. (2014) and the additional EDMTF5, EDMTFH, EDMTF4 and EDMTF1/3 were conserved across birds, however we found that no passerine birds possessed and EDMTFH gene except for the Golden-collared Manakin. Given our phylogenetic results and the overall conservation of

EDMTFH in other avian species, it is likely that EDMTFH has been lost in passerines. Alternatively, EDMTFH may be present within the genomes of passerine birds but due to problems with genomic library preparation and sequencing associated with EDC genes could not be identified (Hron et al. 2016, Davis et al. 2018).

We found that both EDMTFH and EDMTF4 were more highly conserved across species that EDMTF1-3/5, which displayed more lineage-specific conservation where genes in each respective species appeared to be duplicates. For example, we identified at least a single copy of EDMTF1-3/5 in all species investigated except for the Brown Mesite, whereas EDMTF2 was only identified in 4 species other than the chicken indicating that the additional genes are the result of recent gene duplications and are not conserved across all birds. Furthermore, it is likely that EDMTF2 of the Cuckoo is not homologous with EDMTF2 of the chicken, but instead is the result of a recent duplication of another Cuckoo EMDTF gene. This is like the evolutionary history observed for avian loricrins where although LOR3 and LOR3b were conserved across birds, they appeared to be lineage-specific duplications (Davis et al. 2019). The identification of the EDMTF5 gene in the chicken, as well as additional EDMTF genes in the cuckoo, ostrich, pelican and manakin indicate that these genes are currently in the process of duplicating and expanding in many avian species.

To better understand the evolutionary history and origin of the avian EDAA/EDMTF gene family, we examined identified sequences of phylogenetically diverse birds using both Bayesian and ML methods. Previous studies examining the EDC loci in crocodilians and turtles have identified homologous EDAA genes in syntenic locations within their EDCs and we included these genes as outgroups in our analysis.

We found that there are 3 major groups of avian EDAA/EDMTF genes in birds (EDMTFH, EDMTF4, EDMTF1-3/5) and that they likely originated from a single ancestral archosaur EDC gene. Based upon these results, we hypothesize that the divergence of an ancestral archosaur gene resulted in EDMTFH in birds. Duplication and diversification of EDMTFH in birds resulted in EDMTF4, which was conserved across all species investigated. Further duplication and divergence of EDMTF4 resulted in an ancestral form of the EDMTF1-3/5 gene, which has continued to expand in some lineages such as the chicken and cuckoo. As mentioned previously, it is possible that at this point EDMTFH was lost in passerine birds except for the manakin which may have retained this gene.

It is known that different amino acid composition of genes correlates with different functions, and therefore can also correlate with differential expression of related genes (Misawa and Kikuno 2011). Strasser et al. (2014) found that avian EDAA/EDMTF genes exhibited differential expression in developing epidermal tissues such as feathers, scales and skin. We analyzed the amino acid contents of the EDAA/EDMTF genes to look for significant variation in amino acid composition which could correlate with different functions. We found that the amino compositions of EDMTFH and EDMTF4 were similar yet distinct, and significantly different from that of EDMTF1-3/5 (Figure 4.8). These results provide evidence that the differences in amino acid composition are significant enough to indicate differential function of avian EDAA/EDMTF genes. This is further supported by the results of Strasser et al. (2014) which showed that the expression profiles for EDMTFH and EDMTF4 were slightly different from one another, and significantly different from EDMTF1 suggesting differential functions. Together with the results of our phylogenetic analyses which indicate that the EDAA/EDMTF gene family originated from a single ancestral gene, these results provide further evidence that the tandem duplication and diversification of an ancestral EDC gene is sufficient to induce the divergence of gene sequences and differential expression.

Our amino acid analyses also identified the primary amino acid residues making up avian EDAA/EDMTF genes. Specifically, we found that the most prevalent amino acid residues across all EDAA/EDMTF genes are Tyrosine, Glycine, Cysteine and Serine, all residues which are known to be involved in epidermal development processes and mechanical structure. Tyrosine and Glycine are both heavily involved in transglutamination which has been demonstrated to play a major role in the mechanically resilient properties of the skin and appendages (Steinert et al. 1991). In chapter 3 of this study, we examined how avian loricrins are excellent candidates for transglutamination. Cysteine residues are known to facilitate disulfide bonding, which has been shown to be important in feather and scale structure (Hynes et al. 1977, Alibardi et al. 2008). Furthermore, several other avian EDC genes examined such as EDCRP, EDDM and loricrins have been found to contain high contents of cysteine as well as highly conserved cysteine residues, which are indicative of proteins involved in disulfide bonding (Steinert et al. 1991, Strasser et al. 2015, Lachner et al. 2019, Davis et al. 2019). Finally, Serine has been found to be essential in epidermal development processes by facilitating serine protease activity, which is essential for the development of epidermal permeability and indispensable for postnatal survival (Leyvraz et al. 2005; Elias et al. 2014).

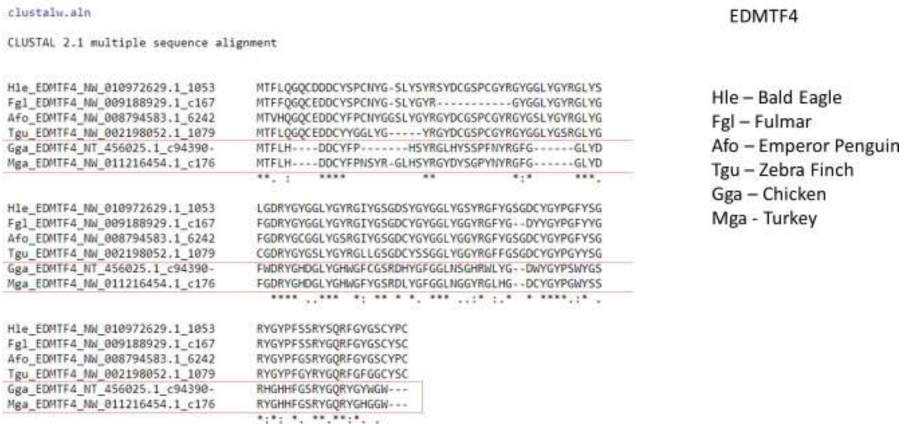
Alibardi et al. (2016) reported that EDMTFH of the chicken and the turkey contained a high amount of Histidine, whereas EDMTFH of all other species contained little to no Histidine, however all were rich in aromatic amino acids. We identified a similar discrepancy in EDMTF4, which was histidine rich in the chicken and turkey, but also contained much less cysteine relative to other species. Moreover, the EDMTF1-3/5 genes of the chicken contained an abundance of Tryptophan which was absent from other species, however this can potentially be explained by the lineage-specific duplication history suspected for these genes. Some studies have suggested domestication as a possible factor in differences associated with EDC genes such as EDMTFH and β-keratins, however we find it unlikely domestication is the reason for this due to the lack of evolutionary time required for the degree of change observed (Nam et al. 2010). Future studies which compare specific physical properties of feathers across different groups of birds such as Galliformes may identify correlation with amino acid differences in EDC genes. Further research is required to better understand the significance of the increased Histidine contents of Galliformes' EDMTFH and EDMTF4.

Our phylogenetic analysis of the EDAA/EDMTF gene family highlighted a similar pattern of evolution with other avian EDC genes. Primarily, evolution through tandem duplications and divergence ultimately resulting in neofunctionalization, however there are 2 major contrasting "types" of evolution observed. The first is what is observed primarily in the EDC genes EDDM and EDCRP which are single genes conserved within the EDC of all birds investigated which have evolved primarily through tandem intragenic duplications (Strasser et al. 2015, Lachner et al. 2019). This has resulted in genes of highly variable size and exact amino acid composition across different species. The avian EDAA/EDMTF gene family, in contrast has evolved largely through tandem gene duplication entire genes of much smaller size. It is likely that the EDCH gene family described by Strasser et al. (2014) also follows this method of evolution. Interestingly, we found that evolution of avian loricrins constitutes both of these methods of evolution, where they have expanded into multiple conserved genes with differential expression, but they have also evolved through significant intragenic gene duplications resulting in variation between species (Davis et al. 2018).

These results highlight the overall evolutionary history of the EDAA/EDMTF gene family and show that there are several similarities to the proposed evolutionary history of the β-keratin gene family. β-keratins are the primary protein component of mature barbs and barbules of feathers and their genetic components have evolved into multiple conserved subfamilies (Greenwold and Sawyer 2011). Evidence suggests that all β-keratin subfamilies originated from a single or few β-keratin gene(s) located within the EDC of an ancestral archosaur and have since expanded to multiple parts of the genome and diversified (Greenwold et al. 2014). It is this diversification and expansion of differentially expressed β-keratin genes which is thought to have played a major role in the adaptation of birds to diverse lifestyles (Sawyer et al. 2000). Our results show that the avian EDAA/EDMTF gene family also likely evolved from a single or small number of ancestral genes and has since expanded and diversified within the EDC locus into multiple conserved subgroups which are differentially expressed. While there is no evidence that the EDAA/EDMTF genes have expanded outside the EDC, these results demonstrate that tandem duplication and divergence of genes has taken place and in light of the results of other studies which show differential expression of EDAA/EDMTF genes, this may be enough to induce neofunctionalization resulting in differentially expressed genes. Furthermore, we suggest that the avian EDAA/EDMTF gene family may continue expanding via gene duplication and divergence in birds therefore it is possible they may translocate and expand to other parts of the genome like β-keratins. Further research is needed, however, to better speculate about the specific function of the EDAA/EDMTF genes in feather structure and development, as well as the selective pressures driving their evolution.

4.6 Figures

Figure 4.1 A

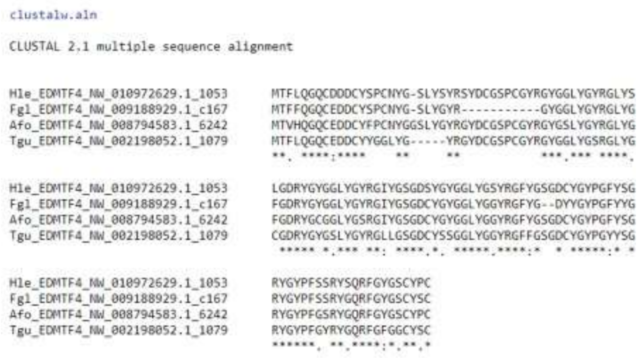


```
clustalw.aln

CLUSTAL 2.1 multiple sequence alignment

Hle_EDMTF4_NW_010972629.1_1053      MTFLQGQCDDDCYSPCNYG-SLYSYRSYDCGSPCGYRGYGGLYGYRGLYS
Fgl_EDMTF4_NW_009188929.1_c167      MTFFQGQCEDDCYSPCNYG-SLYGYR-----------GYGGLYGYRGLYG
Afo_EDMTF4_NW_008794583.1_6242      MTVHQGQCEDDCYFPCNYGGSLYGYRGYDCGSPCGYRGYGSLYGYRGLYG
Tgu_EDMTF4_NW_002198052.1_1079      MTFLQGQCEDDCYYGGLYG-----YRGYDCGSPCGYRGYGGLYGSRGLYG
Gga_EDMTF4_NT_456025.1_c94390-      MTFLH----DDCYFP-------HSYRGLHYSSPFNYRGFG------GLYD
Mga_EDMTF4_NW_011216454.1_c176      MTFLH----DDCYFPNSYR-GLHSYRGYDYSGPYNYRGFG------GLYD
                                    **. :    ****           **           *:*      ***.

Hle_EDMTF4_NW_010972629.1_1053      LGDRYGYGGLYGYRGIYGSGDSYGYGGLYGSYRGFYGSGDCYGYPGFYSG
Fgl_EDMTF4_NW_009188929.1_c167      FGDRYGYGGLYGYRGIYGSGDCYGYGGLYGGYRGFYG--DYYGYPGFYYG
Afo_EDMTF4_NW_008794583.1_6242      FGDRYGCGGLYGSRGIYGSGDCYGYGGLYGGYRGFYGSGDCYGYPGFYSG
Tgu_EDMTF4_NW_002198052.1_1079      CGDRYGYGSLYGYRGLLGSGDCYSSGGLYGGYRGFFGSGDCYGYPGYYSG
Gga_EDMTF4_NT_456025.1_c94390-      FWDRYGHDGLYGHWGFCGSRDHYGFGGLNSGHRWLYG--DWYGYPSWYGS
Mga_EDMTF4_NW_011216454.1_c176      FGDRYGHDGLYGHWGFYGSRDLYGFGGLNGGYRGLHG--DCYGYPGWYSS
                                      **** ..***  *: ** * *. *** ..:* :.*  * ****.:* .

Hle_EDMTF4_NW_010972629.1_1053      RYGYPFSSRYSQRFGYGSCYPC
Fgl_EDMTF4_NW_009188929.1_c167      RYGYPFSSRYGQRFGYGSCYSC
Afo_EDMTF4_NW_008794583.1_6242      RYGYPFGSRYGQRFGYGSCYPC
Tgu_EDMTF4_NW_002198052.1_1079      RYGYPFGYRYGQRFGFGGCYSC
Gga_EDMTF4_NT_456025.1_c94390-      RHGHHFGSRYGQRYGYWGW---
Mga_EDMTF4_NW_011216454.1_c176      RYGHHFGSRYGQRYGHGGW---
                                    *:*: *. **.**:*. .
```

EDMTF4

Hle – Bald Eagle
Fgl – Fulmar
Afo – Emperor Penguin
Tgu – Zebra Finch
Gga – Chicken
Mga - Turkey

Figure 4.1 B



```
clustalw.aln

CLUSTAL 2.1 multiple sequence alignment

Hle_EDMTF4_NW_010972629.1_1053      MTFLQGQCDDDCYSPCNYG-SLYSYRSYDCGSPCGYRGYGGLYGYRGLYS
Fgl_EDMTF4_NW_009188929.1_c167      MTFFQGQCEDDCYSPCNYG-SLYGYR-----------GYGGLYGYRGLYG
Afo_EDMTF4_NW_008794583.1_6242      MTVHQGQCEDDCYFPCNYGGSLYGYRGYDCGSPCGYRGYGSLYGYRGLYG
Tgu_EDMTF4_NW_002198052.1_1079      MTFLQGQCEDDCYYGGLYG-----YRGYDCGSPCGYRGYGGLYGSRGLYG
                                    **. ****:****    **     **           ***.*** ****.

Hle_EDMTF4_NW_010972629.1_1053      LGDRYGYGGLYGYRGIYGSGDSYGYGGLYGSYRGFYGSGDCYGYPGFYSG
Fgl_EDMTF4_NW_009188929.1_c167      FGDRYGYGGLYGYRGIYGSGDCYGYGGLYGGYRGFYG--DYYGYPGFYYG
Afo_EDMTF4_NW_008794583.1_6242      FGDRYGCGGLYGSRGIYGSGDCYGYGGLYGGYRGFYGSGDCYGYPGFYSG
Tgu_EDMTF4_NW_002198052.1_1079      CGDRYGYGSLYGYRGLLGSGDCYSSGGLYGGYRGFFGSGDCYGYPGYYSG
                                     ***** *.*** **: ****.*. *****.****:*  * *****:* *

Hle_EDMTF4_NW_010972629.1_1053      RYGYPFSSRYSQRFGYGSCYPC
Fgl_EDMTF4_NW_009188929.1_c167      RYGYPFSSRYGQRFGYGSCYSC
Afo_EDMTF4_NW_008794583.1_6242      RYGYPFGSRYGQRFGYGSCYPC
Tgu_EDMTF4_NW_002198052.1_1079      RYGYPFGYRYGQRFGFGGCYSC
                                    ******. **.****:*.**.*
```

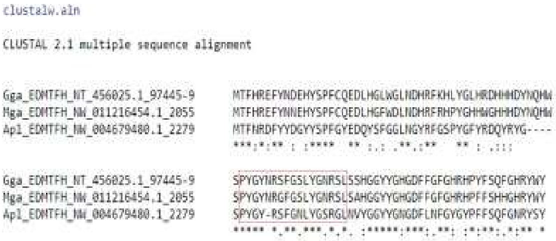
Without Galliformes (Chicken + Turkey) High conservation of EDMTF4

Figure 4.1 A +B : (A) Alignment of EDMTF4 sequences from phylogenetically diverse group of birds. Red lines indicate Galliforme EDMTF4 which has a higher Histidine (H) content not conserved in other species. However, all EDMTF4 contain conserved aromatic amino acid residues in these positions. (B) Alignment of non-Galliforme EDMTF4 genes. When Galliformes are removed from the alignment, there is much higher conservation of EDMTF4.

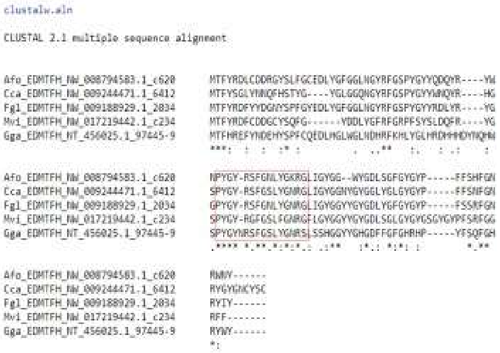
Figure 4.2 A

Galliformes EDMTFH

Avian EDMTFH



```
clustalw.aln

CLUSTAL 2.1 multiple sequence alignment


Gga_EDMTFH_NT_456025.1_97445-9      MTFHREFYNDEHYSPFCQEDLHGLWGLNDHRFKHLYGLHRDHHHDYNQHW
Mga_EDMTFH_NW_011216454.1_2055      MTFHREFYNNEHYSPFCQEDLHGFWDLNDHRFRHPYGHHWGHHHDYNQHW
Apl_EDMTFH_NW_004679480.1_2279      MTFNRDFYYDGYYSPFGYEDQYSFGGLNGYRFGSPYGFYRDQYRYG----
                                    ***:*:** : :****  ** :.: .**.:**   ** : .:::

Gga_EDMTFH_NT_456025.1_97445-9      SPYGYNRSFGSLYGNRSLSSHGGYYGHGDFFGFGHRHPYFSQFGHRYWY
Mga_EDMTFH_NW_011216454.1_2055      SPYGYNRGFGSLYGNRSLSAHGGYYGHGDFFGFGHRHPFFSHHGHRYWY
Apl_EDMTFH_NW_004679480.1_2279      SPYGY-RSFGNLYGSRGLNVYGGYYGNGDFLNFGYGYPFFSQFGNRYSY
                                    ***** *.**.***.*.*. :*****:***:.**: :*:**:.*:** *
```



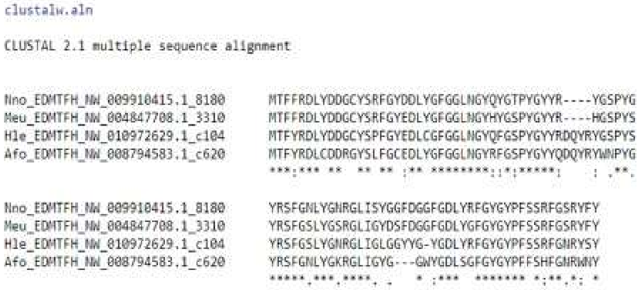
```
clustalw.aln

CLUSTAL 2.1 multiple sequence alignment

Afo_EDMTFH_NW_008794583.1_c620      MTFYRDLCDDRGYSLFGCEDLYGFGGLNGYRFGSPYGYYQDQYR----YW
Cca_EDMTFH_NW_009244471.1_6412      MTFYSGLYNNQFHSTYG----YGLGGQNGYRFGSPYGYYWNQYR----HG
Fgl_EDMTFH_NW_009188929.1_2034      MTFYRDFYYDGNYSPFGYEDLYGFGGLNGYRFGSPYGYYRDLYR----YG
Mvi_EDMTFH_NW_017219442.1_c234      MTFYRDFCDDGCYSQFG------YDDLYGFRFGRPFSYSLDQFR----YG
Gga_EDMTFH_NT_456025.1_97445-9      MTFHREFYNDEHYSPFCQEDLHGLWGLNDHRFKHLYGLHRDHHHDYNQHW
                                    ***:  :  :  :* :          .  ..**   :.   : .:    :

Afo_EDMTFH_NW_008794583.1_c620      NPYGY-RSFGNLYGKRGLIGYGG--WYGDLSGFGYGYP-----FFSHFGN
Cca_EDMTFH_NW_009244471.1_6412      SPYGY-RSFGSLYGNRGLIGYGGHYGYGGLYGLGYGYP-----FFSNFGN
Fgl_EDMTFH_NW_009188929.1_2034      GPYGY-RSFGNLYGNRGLIGYGGYYGYGDLYGFGYGYP-----FSSRFGN
Mvi_EDMTFH_NW_017219442.1_c234      SPYGY-RGFGSLFGNRGFLGYGGYYGYGDLSGLGYGYGSGYGYPFSRFGG
Gga_EDMTFH_NT_456025.1_97445-9      SPYGYNRSFGSLYGNRSLSSHGGYYGHGDFFGFGHRHP-----YFSQFGH
                                    .**** *.**.*:*:*.:  .:**    :*.: *:*: :        *.**

Afo_EDMTFH_NW_008794583.1_c620      RWWY------
Cca_EDMTFH_NW_009244471.1_6412      RYGYGNCYSC
Fgl_EDMTFH_NW_009188929.1_2034      RYIY------
Mvi_EDMTFH_NW_017219442.1_c234      RFF-------
Gga_EDMTFH_NT_456025.1_97445-9      RYWY------
                                    *:
```

Gga – Chicken
Mga – Turkey
Apl – Duck
Mvi - Manakin

Afo – Emperor Penguin
Cca – Will's Widow
Fgl – Fulmar

'51 PYGYRSFGSLYGNRG '65

Figure 4.2 B

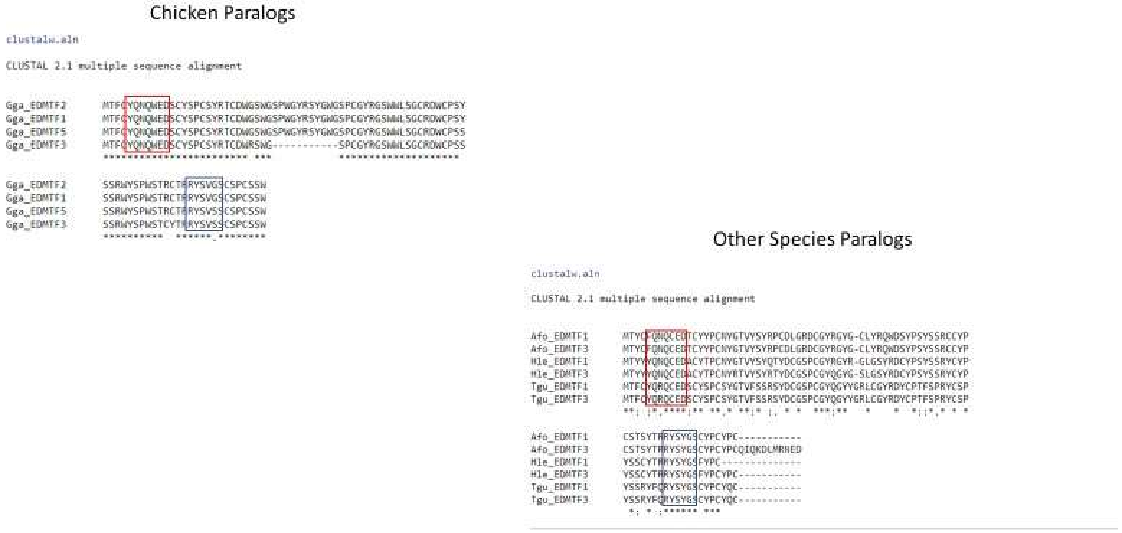


```
clustalw.aln

CLUSTAL 2.1 multiple sequence alignment


Nno_EDMTFH_NW_009910415.1_8180      MTFFRDLYDDGCYSRFGYDDLYGFGGLNGYQYGTPYGYYR----YGSPYG
Meu_EDMTFH_NW_004847708.1_3310      MTFFRDLYDDGCYSRFGYEDLYGFGGLNGYHYGSPYGYYR----HGSPYS
Hle_EDMTFH_NW_010972629.1_c104      MTFYRDLYDDGCYSPFGYEDLCGFGGLNGYQFGSPYGYYRDQYRYGSPYS
Afo_EDMTFH_NW_008794583.1_c620      MTFYRDLCDDRGYSLFGCEDLYGFGGLNGYRFGSPYGYYQDQYRYWNPYG
                                    ***:*** **  ** ** :** ********::*:*****:    : .**.

Nno_EDMTFH_NW_009910415.1_8180      YRSFGNLYGNRGLISYGGFDGGFGDLYRFGYGYPFSSRFGSRYFY
Meu_EDMTFH_NW_004847708.1_3310      YRSFGSLYGSRGLIGYDSFDGGFGDLYGFGYGYPFSSRFGSRYFY
Hle_EDMTFH_NW_010972629.1_c104      YRSFGSLYGNRGLIGLGGYYG-YGDLYRFGYGYPFSSRFGNRYSY
Afo_EDMTFH_NW_008794583.1_c620      YRSFGNLYGKRGLIGYG---GWYGDLSGFGYGYPFFSHFGNRWNY
                                    *****.***.****. .    * :***  ******* *:**.*: *
```

Without Galliformes (Chicken + Turkey) High conservation of EDMTFH

Figure 4.2 A + B : (A) Alignment of EDMTFH sequences from Galliformes (Chicken and turkey) + Duck on left ; Emperor penguin, Will's Widow, Fulmar, Manaking, and Chicken on right. Red box highlights the highly conserved sequence shown in the green box. (B) Alignment of EDMTFH sequences minus the galliformes. Indicates that like EDMTF4, there are differences in the amino acid contents of EDMTFH genes, however aromatic amino acid residues are conserved.

Figure 4.3



Chicken Paralogs

```
clustalw.aln

CLUSTAL 2.1 multiple sequence alignment


Gga_EDMTF2      MTFCYQNQMEDSCYSPCSYRTCDWGSWGSPWGYRSYGWGSPCGYRGSWWLSGCRDWCPSY
Gga_EDMTF1      MTFCYQNQMEDSCYSPCSYRTCDWGSWGSPWGYRSYGWGSPCGYRGSWWLSGCRDWCPSY
Gga_EDMTF5      MTFCYQNQMEDSCYSPCSYRTCDWGSWGSPWGYRSYGWGSPCGYRGSWWLSGCRDWCPSS
Gga_EDMTF3      MTFCYQNQMEDSCYSPCSYRTCDWRSWG-----------SPCGYRGSWWLSGCRDWCPSS
                ************************ ***           *******************

Gga_EDMTF2      SSRWYSPWSTRCTRRYSVGSCSPCSSW
Gga_EDMTF1      SSRWYSPWSTRCTRRYSVGSCSPCSSW
Gga_EDMTF5      SSRWYSPWSTRCTRRYSVSSCSPCSSW
Gga_EDMTF3      SSRWYSPWSTCYTRRYSVSSCSPCSSW
                **********  ******.********
```

Other Species Paralogs

```
clustalw.aln

CLUSTAL 2.1 multiple sequence alignment


Afo_EDMTF1      MTYCFQNQCEDTCYYPCNYGTVYSYRPCDLGRDCGYRGYG-CLYRQWDSYPSYSSRCCYP
Afo_EDMTF3      MTYCFQNQCEDTCYYPCNYGTVYSYRPCDLGRDCGYRGYG-CLYRQWDSYPSYSSRCCYP
Hle_EDMTF1      MTYYYQNQCEDACYTPCNYGTVYSYQTYDCGSPCGYRGYR-GLGSYRDCYPSYSSRYCYP
Hle_EDMTF3      MTYYYQNQCEDACYTPCNYRTVYSYRTYDCGSPCGYQGYG-SLGSYRDCYPSYSSRYCYP
Tgu_EDMTF1      MTFCYQRQCEDSCYSPCSYGTVFSSRSYDCGSPCGYQGYYGRLCGYRDYCPTFSPRYCSP
Tgu_EDMTF3      MTFCYDRQCEDSCYSPCSYGTVFSSRSYDCGSPCGYQGYYGRLCGYRDYCPTFSPRYCSP
                **: :*.****:** **.* **:* :. * *  ***:**  .*    *  *::*.* * *

Afo_EDMTF1      CSTSYTRRYSYGSCYPCYPC------------
Afo_EDMTF3      CSTSYTRRYSYGSCYPCYPCQIQKDLMRHED
Hle_EDMTF1      YSSCYTRRYSYGSFYPC---------------
Hle_EDMTF3      YSSCYTRRYSYGSFYPCYPC------------
Tgu_EDMTF1      YSSRYFQRYSYGSCYPCYQC------------
Tgu_EDMTF3      YSSRYFQRYSYGSCYPCYQC------------
                 *: * .:****** ***
```

Figure 4.3 : Alignment of EDMTF1-3/5 paralogs. On left alignment of all chicken paralogs EDMTF1, EDMTF2, EDMTF3 and the newly identified EDMTF5. Alignment shows that with exception of small deletion in Chicken EDMTF3, these genes represent duplicate genes. (B) Alignment of EDMTF paralogs from additional species demonstrate high lineage-specific conservation. Red and Blue boxes indicate highly conserved sequences found across avian EDMTF genes.

Figure 4.4

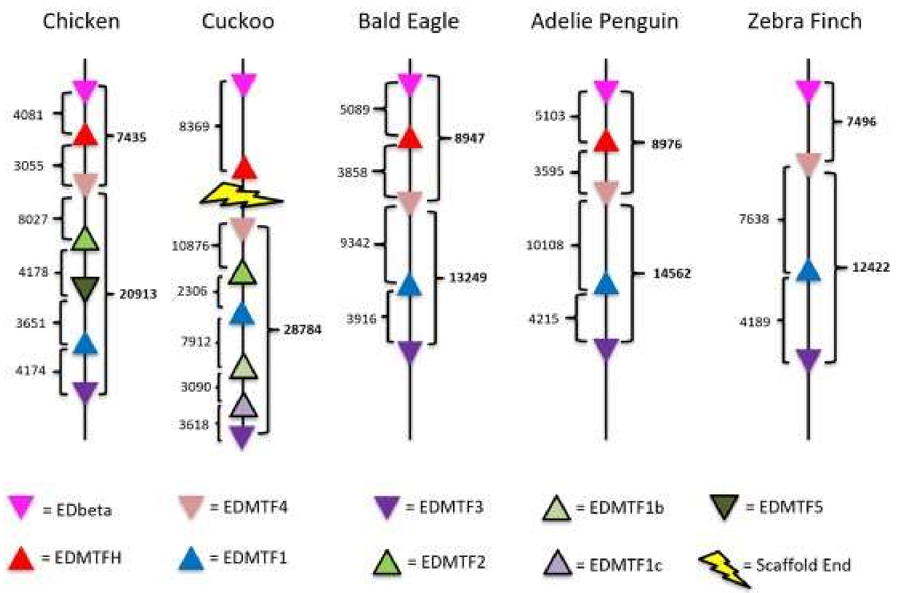


Figure 4.4 : Overall conservation of genomic organization of EDAA/EDMTF gene family. The region of the EDC containing EDAA/EDMTF genes from 5 diverse bird species. The conserved β-keratin gene, EDbeta, is included for reference. Figure depicts variation in number of EDAA/EDMTF genes across different species as well as the variation in the overall size of this region. Brackets with numbers indicate the number of nucleotide residues between EDAA/EDMTF ORFs. The Cuckoo was the only species presented here where this entire region was not found on a single genomic scaffold.

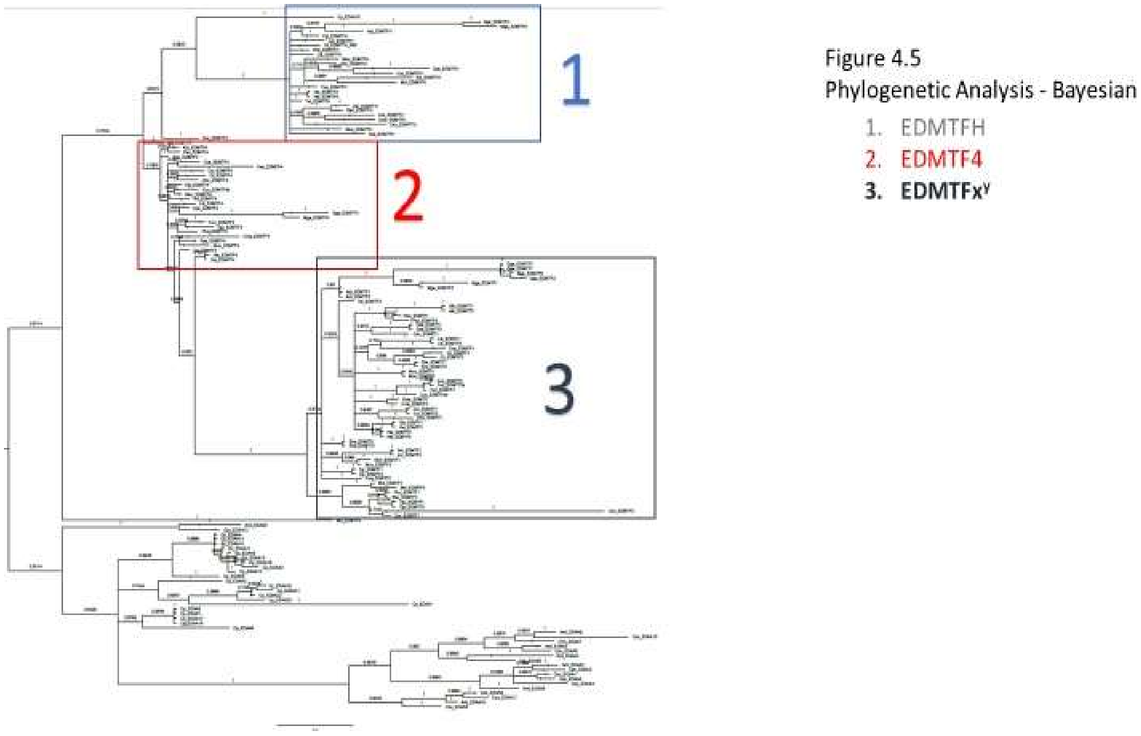


Figure 4.5 : Bayesian Phylogenetic Analysis of EDAA/EDMTF gene family. Figure depicts Bayesian Phylogenetic analysis of avian EDAA/EDMTF genes, using the EDAA genes of crocodilians and testudines as outgroups. The results demonstrate there are 3 conserved groups of avian EDAA/EDMTF genes. Group 1 contains avian EDMFH genes, Group 2 contains EDMTF4 genes and Group 3 contains the remaining EDMTF1-3/5 genes. Group 3 genes display a lineage specific organization like that of LOR3 and LOR3B genes in Davis et al. 2019. The turtle gene cp_EDAA10 was located within the avian EDMTFH group and was the only non-avian species present in the 3 EDAA/EDMTF groups.

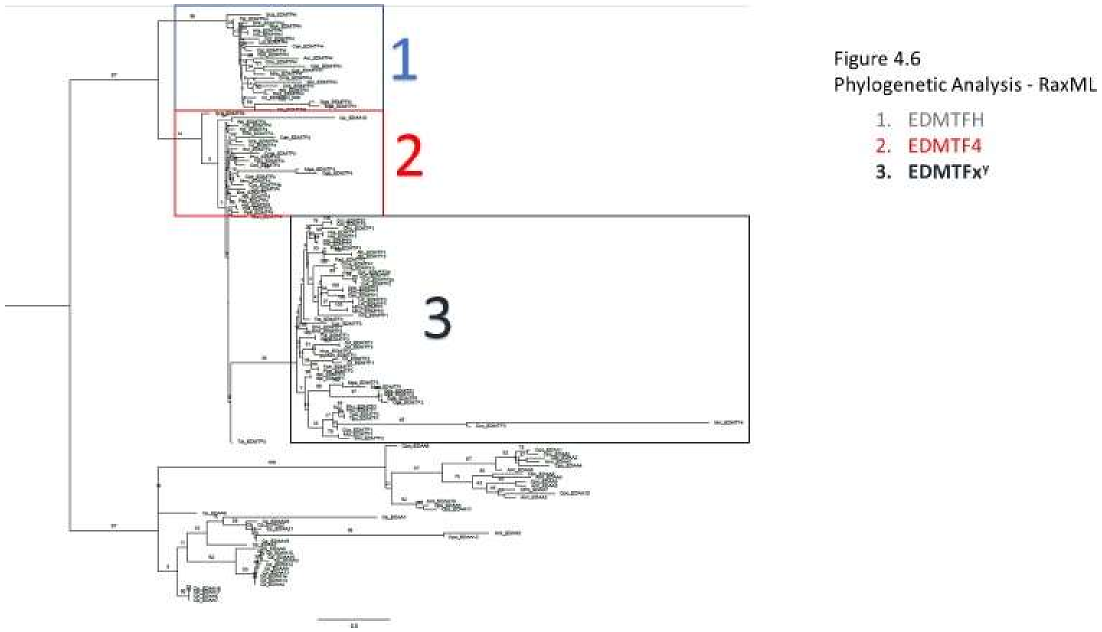


Figure 4.6 : Maximum Likelihood (ML) phylogenetic analysis of EDAA/EDMTF gene family. ML results display similar phylogenetic organization as Bayesian results confirming conservation of 3 distinct groups of avian EDAA/EDMTF genes. The turtle gene cp_EDAA10 was in the avian EDMTF4 group. This contrasted with the Bayesian analysis which placed this gene in the avian EDMTFH group.

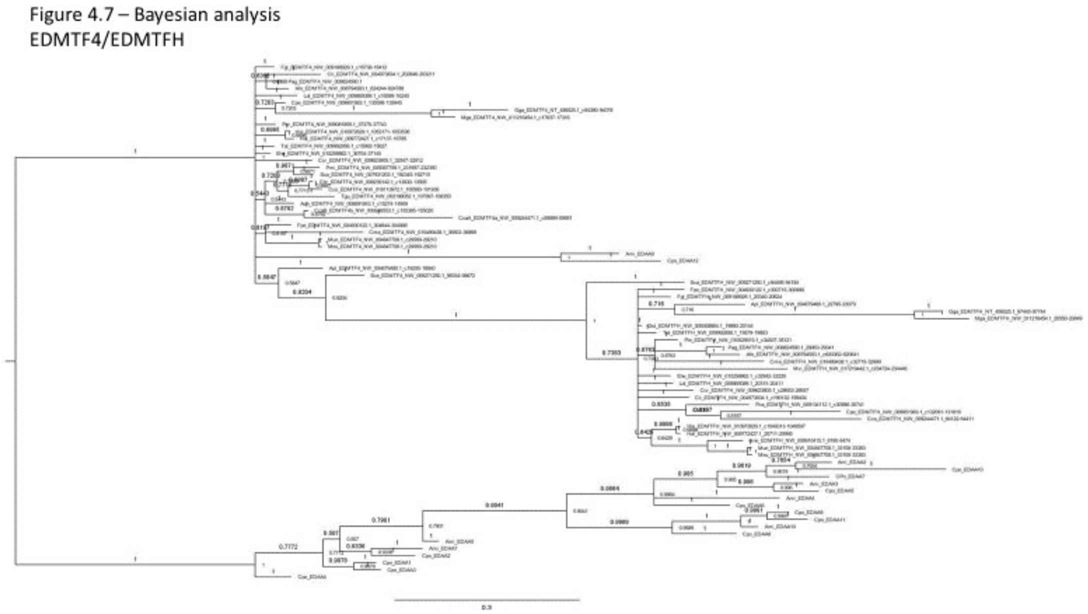


Figure 4.7 : Bayesian phylogenetic analysis of EDMTF4 and EDMTFH genes. Previously identified crocodilian and testudine EDAA genes were used as outgroups. In contrast with complete phylogenetic analyses, here avian EDMTF4 is basal to EDMTFH. Interestingly, the crocodilian genes, Ami_EDAA9 and Cpo_EDAA12 were found in the avian EDMTF4 group.

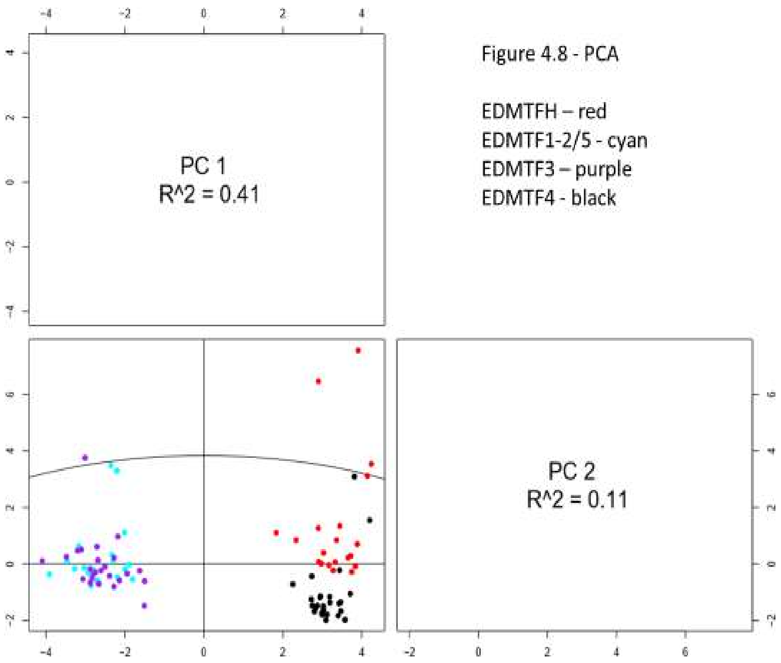


Figure 4.8 : Principle Component Analysis (PCA) of avian EDAA/EDMTF gene amino acid contents. Results demonstrate that the amino acid contents of EDMTFH and EDMTF4 are significantly distinct from those of EDMTF1-3/5. Also that EDMTF4 and EDMTFH have conserved amino acid differences, though not as significant as compared with EDMTF1-3/5. Outliers likely represent the sequences of Galliformes which have unique amino acid compositions, but are still rich in aromatic amino acids.

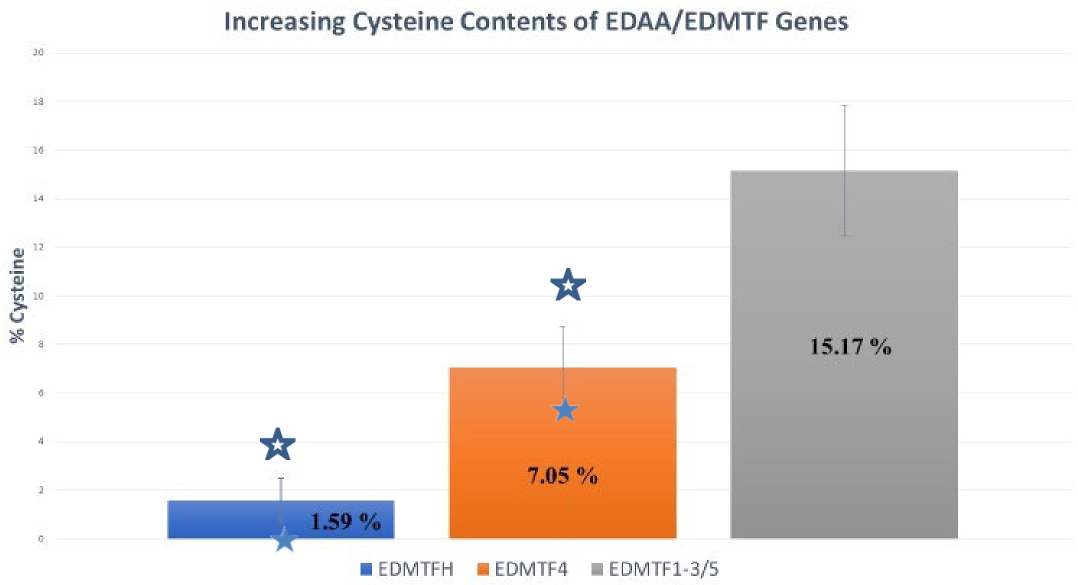


Figure 4.9 : Increasing Cysteine content of avian EDAA/EDMTF genes. Bars indicate the percentage of the total coding sequence which is made up of cysteine residues across avian EDAA/EDMTF genes. Based upon these and phylogenetic results, there appears to be a trend toward increasing cysteine content. The stars indicate that our phylogenetic results did not clearly distinguish without a doubt which gene (EDMTFH or EDMTF4) is basal.

4.7 References

1. Alibardi L, Dalla Valle L, Nardi A, Toni M. 2009. Evolution of hard proteins in the sauropsid integument in relation to the cornification of skin derivatives in amniotes. *J. Anat.* 214, pp560-586. doi:10.1111/j.1469-7580.2009.01045.x.
2. Alibardi L. 2016. Review: cornification, morphogenesis and evolution of feathers. *Protoplasma*. doi:10.1007/s00709-016-1019-2
3. Alibardi L, Holthaus KB, Sukseree S, Hermann M, Tschachler E, Eckhart L. 2016. Immunolocalization of a histidine-rich epidermal differentiation protein in the chicken supports the hypothesis of an evolutionary developmental link between the embryonic subperiderm and feather barbs and barbules. *PLOS one*. doi:10.1037/journal.pone.0167789.
4. Barnes GL, Sawyer RH. 1995. Histidine-rich protein B of embryonic feathers is present in the transient embryonic layers of scutate scales. *J. Exp. Zool.* 271:307-314.
5. Bornelöv S, Seroussi E, Yosefi S, Pendavis K, Curgess SC, Grabherr M, Friedman_einat M, Andersson L. 2017. Correspondence on Levell et al.: Identification of chicken genes previously assumed to be evolutionarily lost. *Genome Biology.* 18: 1 12. Doi:10.1186s13059-017-1231-1.

6. Chuong CM, Nickoloff BJ, Elias PM, Goldsmith LA, Macher E, Maderson PA, Sundberg JP, Tagami H, Plonka PM, Thestrup-pederson K, Bernard BA, Schroder JM, Dotto P, Chang CM, Williams ML, Feingold KR, King LE, Kligman LE, Kligman AM, Rees JL, Chistophers E. 2002. What is the 'true' function of skin? *Exp. Derm.* 11(2):159-187. Doi:10.1034/j.1600-0625.2002.00112.x.
7. Eckhart L. Lippens S, Tschachler E, Declercq W. 2013. Cell Death by Cornification. *Biochemica et Biophysica Acta.* 1833:3471-3480.
8. Elias PM, Gruber R, Crumrine D, Menon G, William ML, Wakefield JS, Holleran WM, Uchida Y. 2014. Formation and functions of the corneocyte lipid envelope. *Biochem Biophys Acta.* 1841(3): 314-318. doi:10.1016/j.bbalip.2013.09.011.
9. Fujimoto S, Takase T, Kadono N, Maekubo K, Hirai Y. 2014. Krtap11-1, a hair keratin-associated protein, as a possible crucial element for the physical properties of hair shafts. *J. Derm. Sci.* Vol. 74 (1): 39-47. Doi:10.1016/j.jdermsci.2013.12.006.
10. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 31(13): 3784-3788.
11. Gerbrands JJ. 1981. On the Relationships Between SVD, KLT and PCA. *Pattern Recognition.* Vol. 14, pp 375-381.
12. Greenwold M, Sawyer RH. 2011. Linking the molecular evolution of avian beta (β) keratins to the evolution of feathers. *J. exp. Zool. B Mol. Dev. Evol.* 15:316(8):609-16. doi:10.1002/jez.b.21436.

13. Greenwold MJ, Bao W, Jarvis ED, Hu H, Li C, Gilbert MTP, Zhang G, Sawyer RH. 2014. Dynamic evolution of the alpha (α) and beta (β) keratins has accompanied integument diversification the adaptation of birds into novel lifestyles. *BMC Evol Biol.* 14:249. Doi:10.1186/s12862-014-0249-1.

14. Haake AR, Konig G, Sawyer RH. 1984. Avian Feather Development: Relationships between Morphogenesis and Keratinization. *Developmental Biology*. 106, 406-413.

15. Hall TA. 1999. BioEdit: a user-friendly biological sequence lignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp.* 41: 95-98.

16. Holthaus KB, Strasser B, Sipos W, Schmidt HA, Mlitz V, Sukseree S, Weissenbacher A, Tschaler E, Alibardi L, Eckhart L. 2015. Comparative genomics Identifies Epidermal Proteins Associated with the Evolution of the Turtle Shell. *Mol. Biol. Evol.* 33(3):726-737. doi: 10.1093/molbev/msv265.

17. Holthaus KB, Strasser B, Lacher J, Sukseree S, Sipos W, Weissenbacher A, Tshcachler E, ALibardi L, Eckhart L. 2018. Comparative analysis of epidermal differentiation genes of crocodilians suggests new models for evolutionary origin of avian feather proteins. *Genome Biol. Evol.* 10(2):694-704. doi:10.1093/gbe/evy035.

18. Holthaus KB, Miltz V, Strasser B, Tchachler E, Alibardi L, Eckhart L. 2017. Identification and comparative analysis of the epidermal differentiation complex in snakes. *Nature Sci. Rep.* 7:45338. doi:10.1038/srep45338.

19. Hron T, Pajer P, Paces J, Bartunek P, Elleder D. 2015. Hidden genes in birds. *Genome Biol.* 16:164. doi:10.1186/s13059-015-0724-z.

20. Huelsonbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics*. 17: 754-755.

21. Hynes R and Destree A. 1977. Extensive disulfide bonding at the mammalian cell surface. *Proc. Natl. Acad. Sci USA Cell Bio.* Vol 74, No. 7, pp. 2855-2859.

22. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33: 1870-1874. doi: 10.1093/molbev/msw054.

23. Kypriotou M, Huber M, Hohl D. 2012. The human epidermal differentiation complex: cornified envelope precursors, S100 proteins and the 'fused genes' family. *Exp. Dermatol.* 21(9):643-649. doi: https://doi.org/10.1111/j.1600-0625.2012.01472.x.

24. Lachner J, Ehrlich F, Miltz V, Hermann M, ALibardi L, Tschaler E, Eckhart L. 2019. Immunolocalization and phylogenetic profiling of the feather protein with the highest cysteine content. *Protoplasma.* doi:https://doi.org/10.1007/s00709-019-01381-3.

25. Leyvraz C, Charles RP, Rubera I, Guitard M, Rotman S, Breiden B, Sandhoff K, Hummler E. 2005. The epidermal barrier function is dependent on the serine protease CAP1/*Prss8*. *J. Cell Bio.* Vol 170, No. 3, 487.496. doi:10.1083/jcb200501038.

26. Li et al. 2014. Two Antarctic penguin genomes reveal insights into their evolutionary history and molecular changes related to the Antarctic environment. *Gigascience.* 3:27. doi:10.1186/2047-217X-3-27.

27. Misawa K and Kikuno RF. 2011. Relationship between amino acid composition and gene expression in the mouse genome. *BMC Res Notes* . (4):20. doi:10.1186/1756-0500-4-20.

28. Nam K, Mugal C, Nabholz B, Scheilzeth H, Wolf JBW, Backstrom N, Kunstner A, Balakrishnan CN, Heher A, Ponting CP, Claton DF, Ellegren H. 2010. Molecular evolution of genes in avian genomes. *Genome Biol. Evol.* 11:R68.

29. Prum RO. 2005. Evolution of the Morphological Innovations of Feathers. *J. Exp. Zoo. MDE.* 304B:570-579. doi:10.1002.jez.b.21073.

30. R Core Team. 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna Austria.

31. Rambaut A. 2012. FigTree Phylogenetic viewing software. http://tree.bio.ed.ac.uk/software/figtree/.

32. Robinson NA, Lapic S, Welter JF, Eckert RL. 1997. S100A11, S100A10, Annexin I, Desmosomal Protiens, Small Proline-rich Protiens, Plasminogen Activator Inhibitor-2, and Involucrin are components of the cornified envelope of cultured human epidermal keratinocytes. *J. Biol. Chem.* Vol 272, No. 18: 12035-12046.

33. Ronquist F, Huelsenbeck JP. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 19: 1572-1574.

34. Sawyer RH, Abbott UK, Fry GN. 1974. Avian Scale Development III: Ultrastructure of the keratinizing cells of the outer and inner epidermal surfaces of the scale ridge. *J. Exp. Zool.* Vol. 190, No.1: 57-70.

35. Sawyer RH, Glenn T, French JO, Mays B, Shames RB, Barnes GL, Rhodes W, Ishikawa Y. 2000. The Expression of Beta (β) keratins in the epidermal appendages of reptiles and birds. *American Zool.* Vol. 40, Issue 4:530-539. Doi:10.1093/icb/40.4.530.

36. Sawyer RH, Knapp LW. 2003. Avian skin development and the evolutionary origin of feathers. *J. Exp. Zool. MDE.* 298B:57-72. doi:10.1002/jez.b.00026.

37. Saathoff M, Blum B, Quast T, Kirfel G, Herzog V. 2004. Simultaneous cell death and desquamation of the embryonic diffusion barrier during epidermal development. *Exp. Cell Res.* 299:415-426. doi:10.1016/j.yexcr.2004.06.009.

38. Smyth GK. 2005. Limma: linear models for microarray data. In Bioinformatics and Computational Biology solutions using R and Bioconductor. *Springer New York,* pages 397-420.

39. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 30(9):1312-1313. doi: 10.1093/bioinformatics/btu033.

40. Steinert P, Mack J, Korge B, Gan SQ, Haynes S, Steven A. 1991. Glycine Loops in Proteins: their occurrence in certain intermediate filament chains, loricrins and single-stranded RNA binding protiens. *Int. J. Biol. Macromol.* 13(3):130-139. doi: https://doi.org/10.1016/0141-8130(91)90037-U.

41. Strasser B, Miltz V, Hermann M, Rice RG, Eigenheer RA, Alibardi L, Tschaler E, Eckhart L. 2014. Evolutionary origin and diversification of epidermal barrier proteins in amniotes. *Mol. Biol. Evol.* 31(12):3194-3205. doi:10.1093/molbev/msu251.

42. Strasser B, Miltz V, Hermann M, Tschachler E, Eckhart L. 2015. Convergent evolution of cysteine-rich proteins in feathers and hairs. *BMC Evol. Biol.* 15:82. doi:10.1186/s12862-015-0360-y.

43. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25(24): 4876-4882.

44. Velasco MVR, de Sa Dias TC, de Freitas AZ, Junior NDV, de Oliveira Pinto CAS, Kaneko TM, Baby AR. 2009. Hair fiber characteristics and methods to evaluate hair physical and mechanical properties. *Braz. J. Pharm. Sci.* Vol. 45. Doi: 10.1590/S1984-82502009000100019.

CHAPTER 5

GENERAL CONCLUSION

5.1 General Conclusion

In this dissertation, we characterized and examined the EDC locus across 48 diverse bird species in order to gain a better understanding of the role it has played in the evolution of feathers. This was accomplished using observational studies, bioinformatics techniques and statistical analyses. Specifically, we utilized the EDC gene sequences of the chicken identified by Strasser et al. (2014) to identify, analyze and compare the avian EDC across 48 diverse species of birds. Since the amino acid composition of a protein has a direct influence on its physical properties and therefore its function, in depth statistical analyses of the amino acid contents of avian EDC genes can provide insights into their functions in epidermal development and structure, as well as provide clues to how they interact with one another and other elements known to be involved in the development process. In this dissertation we utilized amino acid and phylogenetic analyses to better investigate the evolutionary origins and possible functions of avian EDC genes.

In Chapter 2, we used the sequences of the EDC genes identified in the chicken by Strasser et al. (2014) as queries to identify and characterize the conservation of the EDC across 48 diverse species of birds. We found that the architecture of the EDC identified in the chicken is generally conserved across all birds examined, with some exceptions. We did not identify any avian EDC outside of the EDC locus indicating. We also examined the amino acid contents of identified avian EDC genes and found that many of them contain high amounts of residues which are known to be involved in epidermal development and structure such as cysteine, tyrosine and glycine. Furthermore, several EDC genes were only partially identified, contained frameshifts, and some were not identified at all however were not believed to be missing since they were identified in a closely related species. In order to examine the possibility that many of the avian EDC genes that were either partially or not identified at all were actually complete/present, we investigated the nucleotide sequences of identified avian EDC genes for biased G/C contents which could have an impact on their ability to be detected by sequences algorithms or assembled in genomic libraries. We found that many avian EDC genes did in fact contain biased G/C contents which could result in failure to be identified in some species. We propose this as a more likely explanation for the absence or partial identification of many avian EDC genes, as opposed to species-specific gene loss events.

Previous studies have demonstrated that in the chicken, the EDC genes are differentially expressed in several tissues both during epidermal development as well as in mature appendages such as feathers and scales (Strasser et al. 2014). Given the amount of physical diversity observed across avian feathers, a characterization of the conservation of the EDC as whole across birds is critical in further evaluating the specific roles played by EDC genes in feather structure and development. We found that the general organization and architecture of the EDC observed in the chicken, is highly conserved across all birds. We did however identify significant intragenic variation in several EDC genes, as reported by Strasser et al. (2015) in EDCRP, in several genes. These results indicate that avian EDC genes are likely important in the development and structure of feathers due to their high conservation across birds. Furthermore, the presence of significant intragenic variation, even between closely related species indicates that the diversification of the avian EDC has accompanied the evolution of feathers with diverse structures and physical properties. Further research is needed to explore correlation between variations in specific avian EDC genes with the structure and physical properties of feathers.

In Chapter 3 we examined a specific group of conserved avian EDC genes which are homologous to mammalian loricrin, an important structural component of the mammalian cornified envelope and appendages such as hair (Steinert et al. 1991, Candi et al. 2005). We used phylogenetic and statistical analyses techniques to discern the unique and complex evolutionary history exhibited by avian loricrins. Our results demonstrate that in contrast to the single loricrin gene observed in mammals, avian loricrins have expanded into multiple conserved orthologues which are differentially expressed. Furthermore, our phylogenetic analyses showed that the gene reported as LOR2 by Strasser et al. 2014 in the chicken, of all non-passerine birds was actually the result of a more recent gene duplication of LOR3 and was therefore annotated as LOR3B whereas true LOR2 had been lost in all lineages except for passerine birds (Figure 3.4).

Our results also indicate that avian loricrins are important aspects of epidermal appendage development and structure due to the presence of several amino acids associated with epidermal processes. We also show that avian loricrins likely take on the same specialized protein conformation proposed for mammalian loricrins known as the glycine loop, which results from the abundant presence of stretches of glycine residues followed conserved aromatic amino acids results that take on large loop formations. These structures play a significant role in the mechanically resilient properties of epidermal appendages such as flexural and elastic strength.

In chapter 4, we examined another family of avian EDC genes which is rich in aromatic amino acids and which begin with MTF motifs (EDAA/EDMTFs). This group of genes was first identified and annotated in the chicken by Strasser et al. (2014) and contained 5 genes named EDMTFH, and EDMTF1 through EMDTF4. It was soon found that EDMTFH was the previously reported Histidine-rich/fast protein (HRP/fp). Our results showed that the EDAA/EDMTF gene family is conserved across birds. Interestingly, we did not identify EDMTFH in the passerine birds suggesting that it may have been lost. Furthermore, we identified and additional duplicate gene which we annotated EDMTF5 in the chicken as well as additional copies of EDAA/EDMTF genes in the cuckoo. These results suggest, that like several other avian EDC genes, the EDAA/EDMTF gene family has evolved through a series of complex gene loss and duplication events. In contrast to the evolution of other avian EDC genes such as Epidermal Differentiation Protein Rich in Cysteine (EDCRP) and Epidermal Differentiation Protein containing DPCC motifs (EDDM), which have undergone significant intragenic duplication events resulting in long, highly repetitive genes, the EDAA/EDMTF genes have evolved largely through whole gene duplication and divergence. Further studies which focus on the implications of these two contrasting evolutionary histories are needed to further understand how they influence the evolution of novel structures such as feathers.

Further studies are needed to better examine the functions of specific avian EDC genes in epidermal development and feather structure. Preliminary studies investigating the expression of EDMTFH throughout feather development of both wild type and scaleless mutant (sc/sc) chickens have demonstrated that there are potential differences in expression. Further expression studies focusing on EDC genes in the sc/sc chicken may provide insight into the developmental pathways EDC genes are involved in. Overall, these results demonstrate that the avian EDC is conserved across a diverse set of bird species and it contains genes which are involved in the development and structure of epidermal appendages such as feathers. Moreover, these studies demonstrate that the evolution and diversification of the genes in the avian EDC has accompanied the adaptation of birds to a wide variety of diverse habitats and lifestyles via involvement in the development and structure of the feathers and scales of birds. The genes of the avian EDC also represent a model evolutionary system that demonstrates that the tandem duplication and diversification of genes is enough to induce sub/neofunctionalization capable of the adaptation of novel and complex form and function such as that observed epidermal appendages like feathers.

5.2 References

1. Alibardi L, Valle LD, Nardi A, Toni M. 2009. Evolution of hard proteins in the sauropsid integument in relation to the cornification of skin derivatives in amniotes. *J. Anat.* 214. 560-586. doi:10.1111/j.1469-7580.2009.01045.
2. Alibardi L, Holthaus KB, Sukseree S, Hermann M, Tschaler E, Eckhart L. 2016. Immunolocalization of a Histidine-Rich Epidermal Differentiation Protein in the Chicken Supports the Hypothesis of an Evolutionary Developmental Link between the Embryonic Subperiderm and Feather Barbs and Barbules. *PLOS One* 11(12): e0167789. doi:10.1371/journal.pone.0167789.

3. Alibardi L. 2017. Review: cornification, morphogenesis and evolution of feathers. *Protoplasma*. 254(3):1259-1281. doi: https://doi.org/10.1007/s00709-016-1019-2

4. Greenwold MJ, Bao W, Jarvis ED, Hu H, Li C, Gilbert MTP, Zhang G, Sawyer RH. 2014. Dynamic Evolution of the alpha and beta keratins has accompanied integument diversification and the adaptation of birds into novel lifestyles. *BMC Evo. Biol.* 14:249. doi: 10.1186/s12862-014-0249-1.

5. Holthaus KB, Strasser B, Sipos W, Schmidt HA, Mlitz V, Sukseree S, Weissenbacher A, Tschaler E, Alibardi L, Eckhart L. 2015. Comparative genomics Identifies Epidermal Proteins Associated with the Evolution of the Turtle Shell. *Mol. Biol. Evol.* 33(3):726-737. doi: 10.1093/molbev/msv265.

6. Holthaus KB et al. 2018. Comparative analysis of epidermal differentiation genes of crocodilians suggests new models for the evolutionary origin of avian feather proteins. *Gen. Biol. Evol.* 10(2): 694-704. doi: 10.1093/gbe/evy035.

7. Jarvis ED et al. 2014. Whole genome analyses resolve the early branches in the tree of life of modern birds. *Science.* 346(6215):1320-1331. doi: 10.1126/science.1253451.

8. Lachner J, Ehrlich F, Mlitz V, Harmann M, ALibardi L, Tschaler E, Eckhart L. 2019. Immunolocalization of phylogenetic profiling of the feather protein with the highest cysteine content. *Protoplasma*. doi:https://doi.org/10.1007/s00709-019-01381-3.

9. Sawyer RH, Knapp LW. 2003. Avian skin development and the evolutionary origin of feathers. *J. Exp. Zool. MDE.* 298B:57-72. doi:10.1002/jez.b.00026.

10. Strasser B et al. 2014. Evolutionary Origin and Diversification of epidermal barrier proteins in amniotes. *Mol Biol Evol*. 31(12): 3194-3205. doi: 10.1093/molbev/msu251.

11. Strasser B, Miltz V, Hermann M, Tschachler E, Eckhart L. 2015. Convergent evolution of cysteine-rich-proteins in feathers and hair. *BMC Evol Biol*. 15:82. doi: https://doi.org/10.1186/s12862-015-0360-y.

COMPLETE REFERENCES

1. Alibardi L, Valle LD, Nardi A, Toni M. (2009). Evolution of hard proteins in the sauropsid integument in relation to the cornification of skin derivatives in amniotes. *J. Anat.* 214. 560-586. doi:10.1111/j.1469-7580.2009.01045.
2. Alibardi L, Holthaus KB, Sukseree S, Hermann M, Tschaler E, Eckhart L. (2016). Immunolocalization of a Histidine-Rich Epidermal Differentiation Protein in the Chicken Supports the Hypothesis of an Evolutionary Developmental Link between the Embryonic Subperiderm and Feather Barbs and Barbules. *PLOS One* 11(12): e0167789. doi:10.1371/journal.pone.0167789.
3. Alibardi L. (2017). Review: cornification, morphogenesis and evolution of feathers. *Protoplasma*. 254(3):1259-1281. doi: https://doi.org/10.1007/s00709-016-1019-2
4. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.
5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389-3402.
6. Backström N et al. (2010). The recombination landscape of the zebra finch Taeniopygia guttata genome. *Genome Res.* doi: 10.1101/gr.101410.109.
7. Barati D, Kader S, Shariati SRP, Moeinzadeh S, Sawyer RH, Jabbari E. (2017). Synthesis and Characterization of Phot-Cross-Linkable Keratin Hydrogels for Stem Cell Encapsulation. *Biomacromolecules*. 18(2):398-412. doi: 10.1021/acs.biomac.6b01493.

8. Barnes GL, Sawyer RH. 1995. Histidine-rich protein B of embryonic feathers is present in the transient embryonic layers of scutate scales. *J. Exp. Zool.* 271:307-314.

9. Bornelöv S, Seroussi E, Yosefi S, Pendavis K, Curgess SC, Grabherr M, Friedman_einat M, Andersson L. (2017). Correspondence on Levell et al.: Identification of chicken genes previously assumed to be evolutionarily lost. *Genome Biology.* 18: 1 12. Doi:10.1186s13059-017-1231-1.

10. Burley SK, Petsko GA. (1985). Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science*. Vol. 229, 4708; 23-28. doi: 10.1126/science.3892686.

11. Cai Q, Qian X, Lang Y, et al. (2013). Genome sequence of ground tit Pseudopodoces humilis and its adaptation to high altitude. *Genome Biol*. 14(3):R29. doi:10.1186/gb-2013-14-3-r29.

12. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. (2009). Basic local alignment search tool. *BMC Bioinformatics.* 10:42. doi: 10.1186/1471-2105-10-421.

13. Candi E, Schmidt R, Melino G. (2005). The cornified envelope: a model of cell death in the skin. *Nat. Rev. Mol. Cell Bio*. 6(4): 328-340. doi:10.1038/nrm1619

14. Chuong CM et al. (2002). What is the 'true' function of skin? *Exp Dermatol.* 11:159-187. doi: https://doi.org/10.1034/j.1600-0625.2002.00112.x.

15. Contzler R, Favre B, Huber M, Hohl D. (2005). Cornulin, a New Member of the “Fused Gene” Family, is Expressed During Epidermal Differentiation. *J. Invs. Derm.* Vol. 124, Issue 5: 990-997. Doi:https://doi.org/10.1111/j.0022-202X.2005.23694.x.

16. Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biol.* Lett. 8, 783-786. Doi: 10.1098/rsbl.2012.0331.

17. Daiquing, Liao. (1999). Concerted Evolution: Molecular Mechanism and Biological Implications. *Am. J. Hum. Genet.* 64(1): 24-30. doi: 10.1086/302221.

18. Dawson DA et al. (2007). Gene order and recombination rate in homologous chromosome regions of the chicken and a passerine bird. *Mol. Biol. Evol.* 24(7): 1537-1552. doi: 10.1093/molbev/msm07.

19. Elder JT, Zhao X. (2002). Evidence for local control of gene expression in the epidermal differentiation complex. *Exp. Derm.* Vol. 11, Issue 5: 406-412. doi:https://doi.org/10.1034/j.1600-0625.2002.110503.

20. Eckhart L, Lippens S, Tschaler E, Declercq W. (2013). Cell death by cornification. *BBA-Mol. Cell Res.* 1833(12): 3471-3480. doi: https://doi.org/10.1016/j.bbamcr.2013.06.010.

21. Elias PM, Gruber R, Crumrine D, Menon G, William ML, Wakefield JS, Holleran WM, Uchida Y. 2014. Formation and functions of the corneocyte lipid envelope. *Biochem Biophys Acta.* 1841(3): 314-318. doi:10.1016/j.bbalip.2013.09.011.

22. Ericson PGP, Christidis L, Cooper A, Irestedt M, Jackson J, Johansson US, Norman JA. (2002). A Gondwanan origin of passerine birds supported by DNA sequences of the endemic New Zealand wrens. *P. Roy. Soc. B-Biol. Sci.* 269(1488): 234-241. doi: 10.1098/rspb.2001.1877.

23. Fankl C, Kuhl H, Weber M, Ralser M, Timmermann B, Gahr M. (2014). NCBI Serinus canaria Annotation Release 100. MPI Molgen, NCBI Eukaryotic Genome Annotation Pipeline. *NCBI Genome.* GCF_000534875.

24. Fujimoto S, Takase T, Kadono N, Maekubo K, Hirai Y. 2014. Krtap11-1, a hair keratin-associated protein, as a possible crucial element for the physical properties of hair shafts. *J. Derm. Sci.* Vol. 74 (1): 39-47. Doi:10.1016/j.jdermsci.2013.12.006.

25. Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D., Bairoch A. (2003) *ExPASy: the proteomics server for in-depth protein knowledge and analysis* Nucleic Acids Res. 31:3784-3788

26. Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. (2002). *Protein Identification and Analysis Tools on the ExPASy Server;*
(In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press pp. 571-607

27. Gelman A, Rubin DB. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science.* 7(4): 457-511.

28. Gerbrands JJ. 1981. On the Relationships Between SVD, KLT and PCA. *Pattern Recognition.* Vol. 14, pp 375-381.

29. Gilbert SF. 2014. Developmental Biology, 10th edition. Pgs. 79-83.

30. Gish, W. & States, D.J. (1993) "Identification of protein coding regions by database similarity search." Nature Genet. 3:266-272.

31. Greenwold MJ, Sawyer RH. (2013). Molecular Evolution and Expression of Archosaurian β-Keratins: Diversification and Expansion of Archosaurian β-Keratins and the Origin of Feather β-keratins. *J. Exp. Zool. Part* B. 320(6):393-405. doi: 10.1002/jez.b.22514.

32. Greenwold MJ, Bao W, Jarvis ED, Hu H, Li C, Gilbert MTP, Zhang G, Sawyer RH. (2014). Dynamic Evolution of the alpha and beta keratins has accompanied integument diversification and the adaptation of birds into novel lifestyles. *BMC Evo. Biol.* 14:249. doi: 10.1186/s12862-014-0249-1.

33. Gremillet D, Chauvin C, Wilson RP, Meho YL, Wanless S. (2005). Unusual feather structure allows partial plumage wettability in diving great cormorants Phalacrocorax carbo. *J. Avian Biol.* 36(1): 57-63. doi: https://doi.org/10.1111/j.0908-8857.2005.03331.x.

34. Haake AR, Konig G, Sawyer RH. 1984. Avian Feather Development: Relationships between Morphogenesis and Keratinization. *Developmental Biology*. 106, 406-413.

35. Hall TA. (1999). BioEdit: a user-friendly biological sequence lignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp.* 41: 95-98.

36. Hohl D, Mehrel T, Lichti U, Turner ML, Roop DR, Steinert P. (1991). Characterization of Human Loricrin. *J. Biol. Chem.* 266(10) :6626-6636

37. Holland PWH, Garcia-Fernàndez J, Williams NA, Sidow A. (1994). Gene duplications and the origins of vertebrate development. *Development*. 125-133.

38. Holthaus KB, Mlitz V, Strasser B, Tschaler E, Alibardi L, Eckhart L. (2017). Identification and comparative analysis of the epidermal differentiation complex in snakes. *Sci. Rep.* 7, 45338 doi: 10.1038/srep45338.

39. Holthaus KB, Strasser B, Sipos W, Schmidt HA, Mlitz V, Sukseree S, Weissenbacher A, Tschaler E, Alibardi L, Eckhart L. (2015). Comparative genomics Identifies Epidermal Proteins Associated with the Evolution of the Turtle Shell. *Mol. Biol. Evol.* 33(3):726-737. doi: 10.1093/molbev/msv265.

40. Holthaus KB et al. (2018). Comparative analysis of epidermal differentiation genes of crocodilians suggests new models for the evolutionary origin of avian feather proteins. *Gen. Biol. Evol.* 10(2): 694-704. doi: 10.1093/gbe/evy035.

41. Hron t, Pajer P, Paces J, Bartunek P, Elleder D. (2015). Hidden genes in birds. *Genome Biology*. 16:164. Doi:10.1186/s13059-015-0724-z.

42. Hunter S. (2002). Feathers: Whats Flight got to do – got to do with it. http://ncsce.org/pages/feathers.html.

43. Huelsonbeck JP, Ronquist F. (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics*. 17: 754-755.

44. Hynes RO, Destree A. (1977). Extensive disulfide bonding at the mammalian cell surface. *Proc. Natl. Acad. Sci.* Vol. 74, No. 7: 2855-2859.

45. Ishida-Yamamoto A., Takahashi H., Iizuka H. (1998). Loricrin and human skin diseases: molecular bases of loricrin keratodermas. *Histol. Histopathol.* 13(3):819-826. doi: 10.14670/HH-13.819

46. Jarvis ED et al. (2014). Whole genome analyses resolve the early branches in the tree of life of modern birds. *Science.* 346(6215):1320-1331. doi: 10.1126/science.1253451.

47. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. (1998). Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* 23(10): 403-405.

48. John JA et al. (2012). Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biol.* 13(1): 415. doi:10.1186/gb-2012-13-1-415.

49. Kalinin AE, Kajava AV, Steinert PM. (2002). Epithelial barrier function: assembly and structural features of the cornified cell envelope. *BioEssays.* 24(9); 789-800. doi: https://doi.org/10.1002/bies.10144.

50. Kazilek CJ. (2009). Feather Biology. ASU – Ask A Biologist. Retrieved April 16, 2020 from https://askabiologist.asu.edu/explore/feather-biology

51. Kischer CW. (1965). Fine structure of the developing down feather 1. 1963. *J. Ultstr. Res.* Vol. 8, (3-4): 305-321.

52. Kondo M, Sekine T, Miyakoshi T, Kitajima K, Egawa S, Seki R, Abe G, Tamura K. (2018). Flight feather development: its early specialization during embryogenesis. *Zoological Lett.* 4:2. doi: 10.1186/s40851-017-0085-4

53. Kumar S, Stecher G, Tamura K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33: 1870-1874. doi: 10.1093/molbev/msw054.

54. Kypriotou M, Huber M, Hohl D. (2012). The human epidermal differentiation complex: cornified envelope precursors, S100 proteins and the ‘fused genes’ family. *Exp. Dermatol.* 21(9):643-649. doi: https://doi.org/10.1111/j.1600-0625.2012.01472.x.

55. Lachner J, Ehrlich F, Mlitz V, Harmann M, ALibardi L, Tschaler E, Eckhart L. (2019). Immunolocalization of phylogenetic profiling of the feather protein with the highest cysteine content. *Protoplasma*. doi:https://doi.org/10.1007/s00709-019-01381-3.

56. Leyvraz C, Charles RP, Rubera I, Guitard M, Rotman S, Breiden B, Sandhoff K, Hummler E. 2005. The epidermal barrier function is dependent on the serine protease CAP1/*Prss8*. *J. Cell Bio.* Vol 170, No. 3, 487.496. doi:10.1083/jcb200501038.

57. Li, C., Zhang, Y., Li, J., Kong, L., Hu, H., Pan, H., Xu, L., Deng, Y., Li, Q., Jin, L., Yu, H., Chen, Y., Liu, B., Yang, L., Liu, S., Zhang, Y., Lang, Y., Xia, J., He, W., Shi, Q., … Zhang, G. (2014). Two Antarctic penguin genomes reveal insights into their evolutionary history and molecular changes related to the Antarctic environment. *GigaScience*, *3*(1), 27. https://doi.org/10.1186/2047-217X-3-27

58. Leyvraz C, Charles RP, Rubera I, Guitard M, Rotman S, Breiden B, Sandhoff K, Hummerl E. (2005). The epidermal barrier function is dependent on the serine protease CAP1/Prss8. *J. Cell Bio.* 170(3): 487-496. doi:https://www.jcb.org/cgi/doi/10.1083/jcb.200501038.

59. Milinkovitch MC, Helaers R, Depiereux E, Tzika AC, Gabaldon T. (2010). 2x genomes – depth does matter. *Genome Biol.* 11:R16. doi: https://doi.org/10.1186/gb-2010-11-2-r16.

60. Miller HC, Biggs PJ, Voelckel C, Nelson NJ. (2012). De novo sequence assembly and characterization of a partial transcriptome for an evolutionarily distinct reptile, the tuatara (Sphenodon punctatus). *BMC Genomics*. 13:439. doi: 10.1186/1471-2164-13-439.

61. Misawa K and Kikuno RF. 2011. Relationship between amino acid composition and gene expression in the mouse genome. *BMC Res Notes* . (4):20. doi:10.1186/1756-0500-4-20.

62. Nam K, Mugal C, Nabholz B, Scheilzeth H, Wolf JBW, Backstrom N, Kunstner A, Balakrishnan CN, Heher A, Ponting CP, Claton DF, Ellegren H. 2010. Molecular evolution of genes in avian genomes. *Genome Biol. Evol.* 11:R68.

63. Norberg UM. (1985). Evolution of Vertebrate Flight: An Aerodynamic Model for the Transistion from Gliding to Active Flight. *American Naturalist*. Vol. 126, No. 3. doi:10.1086/284419

64. Notredame C, Higgins DG, Heringa J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302:205-217. doi: 10.1006/jmbi.2000.4042.

65. Ostmann OW, Ringer RK, Tetzlaff M. (1963). The anatomy of the feather follicle and its immediate surroundings. *Poultry Sci.*, 42957-969.

66. Peona V, Weissensteiner MH, Suh A. (2018). How complete are ‘complete’ genome assemblies? – An avian perspective. *Mol. Ecol. Resour.* doi: 10.1111/1755-0998.12933.

67. Pierard G, Goffin V, Hermanns-Le T, Pierard-Franchimont C. (2000). Corneocyte desquamation. *Int. J. Mol. Med.* 6(2):217-238. doi: https://doi.org/10.3892/ijmm.6.2.217.

68. Prum OR, Berv JS, Dorburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature.* 534, S7-8. doi: 10.1038/nature15697.

69. Rambaut A. (2012). FigTree Phylogenetic viewing software. http://tree.bio.ed.ac.uk/software/figtree/.

70. Rice RH, Winters BR, Durbin-Johnson BP, Rocke DM. (2013). Chicken Corneocyte Cross-Linked Proteome. *J. Proteome Res.* 12(2): 772-776. doi: 10.1021/pr301036k.

71. Robinson NA, Lapic S, Welter JF, Eckert RL. (1997). S100A11, S100A10, Annexin I, Desmosomal Protiens, Small Proline-rich Protiens, Plasminogen Activator Inhibitor-2, and Involucrin are components of the cornified envelope of cultured human epidermal keratinocytes. *J. Biol. Chem.* Vol 272, No. 18: 12035-12046.

72. Ronquist F, Huelsenbeck JP. (2003). MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 19: 1572-1574.

73. Rose GD, Roy S. (1980). Hydrophobic basis of packing in globular proteins. *Proc. Natl. Acad. Sci.* 77(8): 4643-4647.

74. RStudio Team. (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

75. Sawyer S. (1989). Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* 6(5): 526–538. doi: https://doi.org/10.1093/oxfordjournals.molbev.a040567.

76. Sawyer RH, Abbott UK, Fry GN. (1974). Avian Scale Development III: Ultrastructure of the keratinizing cells of the outer and inner epidermal surfaces of the scale ridge. *J. Exp. Zool.* Vol. 190, No.1: 57-70.

77. Sawyer RH, Knapp LW. (2003). Avian skin development and the evolutionary origin of feathers. *J. Exp. Zool. MDE.* 298B:57-72. doi:10.1002/jez.b.00026.

78. Saathoff M, Blum B, Quast T, Kirfel G, Herzog V. (2004). Simultaneous cell death and desquamation of the embryonic diffusion barrier during epidermal development. *Exp. Cell Res.* 299:415-426. doi:10.1016/j.yexcr.2004.06.009.

79. Science Learning Hub – The University of Waikato. (2007). *Feathers and Flight*. Retrieved from https://www.sciencelearn.org.nz/resources/308-feathers-and-flight

80. Shames RB, Knapp LW, Barnes GL, Sawyer RH. (1993). Region-Specific Patterns of Beta Keratin Expression During Avian Skin Development. *Developmental Dynamics* 196:283-290.

81. Singh J, Thornton JM. (1985). The interaction between phenylalanine rings in proteins. *FEBS Lett.* 191(1): 1-6. doi: https://doi.org/10.1016/0014-5793(85)80982-0.

82. Smyth GK. 2005. Limma: linear models for microarray data. In Bioinformatics and Computational Biology solutions using R and Bioconductor. *Springer New York,* pages 397-420.

83. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. (2007). pcaMethods – a Bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23(9):1164–1167. doi: 10.1093/bioinformatics/btm069.

84. Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 30(9):1312-1313. doi: 10.1093/bioinformatics/btu033.

85. Steinert P, Mack J, Korge B, Gan SQ, Haynes S, Steven A. (1991). Glycine Loops in Proteins: their occurrence in certain intermediate filament chains, loricrins and single-stranded RNA binding protiens. *Int. J. Biol. Macromol.* 13(3):130-139. doi: https://doi.org/10.1016/0141-8130(91)90037-U.

86. Strasser B et al. (2014). Evolutionary Origin and Diversification of epidermal barrier proteins in amniotes. *Mol Biol Evol*. 31(12): 3194-3205. doi: 10.1093/molbev/msu251.

87. Strasser B, Miltz V, Hermann M, Tschachler E, Eckhart L. (2015). Convergent evolution of cysteine-rich-proteins in feathers and hair. *BMC Evol Biol*. 15:82. doi: https://doi.org/10.1186/s12862-015-0360-y.

88. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25(24): 4876-4882.

89. Treangen TJ. Salzberg S. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*; 13(1): 36-46. doi:10.1038/nrg3117.

90. Vanhoutteghem A, Djian P, Green H. (2008). Ancient origin of the gene encoding involucrin, a precursor of the cross-linked envelope of epidermis and related epithelia. *P. Natl. Acad. Sci-Biol.* 105(40): 15481-15486. doi: https://doi.org/10.1073/pnas.0807643105.

91. Velasco MVR, de Sa Dias TC, de Freitas AZ, Junior NDV, de Oliveira Pinto CAS, Kaneko TM, Baby AR. 2009. Hair fiber characteristics and methods to evaluate hair physical and mechanical properties. *Braz. J. Pharm. Sci.* Vol. 45. Doi: 10.1590/S1984-82502009000100019.

92. Veltri A, Lang C, Lein WH. (2017). Concise Review: Wnt Signaling Pathways in Skin Development and Epidermal Stem Cells. *Stem Cells.* 36:22-35. http://dx.doi.org/10.1002/stem.2723.

93. Völker M et al. (2010). Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome Res.* doi: 10.1101/gr.103663.109.

94. Wang H, Parry D, Jones LN, Idler WW, Marekov LN, Steinert P. (2000). In Vitro Assembly and Structure of Trichocyte Keratin Intermediate Filaments: A Novel Role for Stabilization by Disulfide Bonding. *J. Cell Biol.* 151(7): 1459-1468.

95. Wang S, Yang Z, Gong G, Wang J, Wu J, Yang S, Jiang L. (2016). Icephobicity of penguins Spheniscus humboldti and and artificial replica of penguin feather with air-infused hierarchical rough structures. *J. Phys. Chem.* 120. 15923-15929. doi:10.1021/acs.jpcc.5b12298.

96. Yang F, Zhao G, Zhou L, Li B. (2015). Complete mitochondrial genome of White-rumped Munia Lonchura striata swinhoei (Passeriformes: Estrildidae). *Mitochondrial DNA.* 27(4): 3028-3029. doi: 10.3109/19401736.2015.1063052.

97. Yu M, Wu P, Widelitz RB, Chuong CM. (2002). The morphogenesis of feathers. *Nature*. 420(6913):308-312. doi:10.1038/nature01196.

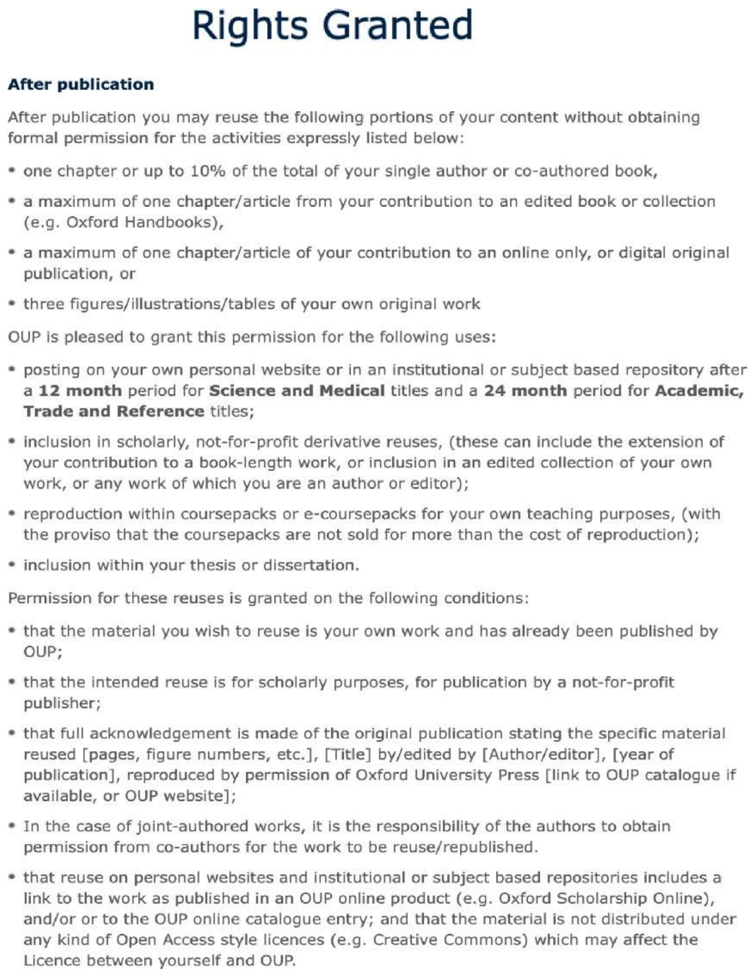
98. Zhang G et al. (2014) Comparative Genomics Reveals Insights into Avian Genome Evolution and Adaptation. *Science*. 346(6215): 1311-1320. doi: 10.1126/science.1251385.

99. Zhang J. (2003). Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18(6): 292-298. doi: 10.1016/S0169-5347(03)00033-8.

100. Zhu BY, Zhou NE, Kay CM, Hodges RS. (1993). Packing and hydrophobicity effects on protein folding and stability: Effects of β-branched amino acids, valine and isoleucine, on the formation and stability of two-stranded α-helical coiled coils/leucine zippers. *Protein Sci.* 2(3): 383-394. doi: 10.1002/pro.5560020310.

APPENDIX A

PERMISSIONS TO PUBLISH



Rights Granted

After publication

After publication you may reuse the following portions of your content without obtaining formal permission for the activities expressly listed below:

- one chapter or up to 10% of the total of your single author or co-authored book,
- a maximum of one chapter/article from your contribution to an edited book or collection (e.g. Oxford Handbooks),
- a maximum of one chapter/article of your contribution to an online only, or digital original publication, or
- three figures/illustrations/tables of your own original work

OUP is pleased to grant this permission for the following uses:

- posting on your own personal website or in an institutional or subject based repository after a **12 month** period for **Science and Medical** titles and a **24 month** period for **Academic, Trade and Reference** titles;
- inclusion in scholarly, not-for-profit derivative reuses, (these can include the extension of your contribution to a book-length work, or inclusion in an edited collection of your own work, or any work of which you are an author or editor);
- reproduction within coursepacks or e-coursepacks for your own teaching purposes, (with the proviso that the coursepacks are not sold for more than the cost of reproduction);
- inclusion within your thesis or dissertation.

Permission for these reuses is granted on the following conditions:

- that the material you wish to reuse is your own work and has already been published by OUP;
- that the intended reuse is for scholarly purposes, for publication by a not-for-profit publisher;
- that full acknowledgement is made of the original publication stating the specific material reused [pages, figure numbers, etc.], [Title] by/edited by [Author/editor], [year of publication], reproduced by permission of Oxford University Press [link to OUP catalogue if available, or OUP website];
- In the case of joint-authored works, it is the responsibility of the authors to obtain permission from co-authors for the work to be reuse/republished.
- that reuse on personal websites and institutional or subject based repositories includes a link to the work as published in an OUP online product (e.g. Oxford Scholarship Online), and/or or to the OUP online catalogue entry; and that the material is not distributed under any kind of Open Access style licences (e.g. Creative Commons) which may affect the Licence between yourself and OUP.