A Comparative Study of Host Genome Evolution in Relation to Endogenous Retrovirus Load

Wanjing Zheng

DOCTOR OF PHILOSOPHY

Department of Evolutionary Studies of Biosystems

School of Advanced Sciences

The Graduate University for Advanced Studies

2019

Acknowledgement

I would like to express my deepest gratitude to my adviser, Professor Yoko, Satta, for her all-sided guidance on this thesis, and for her constant encouragement and great concern for my study and life.

I am very grateful to my sub-adviser, Dr. Jun Gojobori, for his guidance in computation, valuable advice on this thesis and encouragement to my study. I am also very grateful to Dr. Alexander Suh from Uppsala University for his critical reading and helpful comments on Chapter 3 of this thesis.

I greatly appreciate all the members of Satta Lab for their constructive discussions on my study during lab meetings. In addition, I thank Dr. Quintin Lau for his helpful advice in data analysis and English editing. I'm very thankful to people in the Department of Evolutionary Studies of Biosystems and the SOKENDAI headquarters for their academic or administrative support throughout my PhD program.

I thank the Ministry of Education, Culture, Sports, Science and Technology of Japan for offering me the Japanese Government (MEXT) Scholarship for Research Students. I thank the ROIS National Institute of Genetics (NIG) for offering the access to NIG supercomputer for my study.

In the end, I would like to offer my special thanks to my parents for their endless support.

Abstract

It is known that endogenous retroviruses (ERV) are present in all vertebrates investigated and that retrovirus infection in vertebrates has a history spanning hundreds of million years. The unique type of relationship between hosts and ERVs/retroviruses throughout the long history, which includes both conflict and co-option, may have shaped the host-parasite evolutionary interaction in vertebrates and this evolutionary interaction may differ between vertebrate groups. Mammals and birds differed largely in their ERV load, which is defined herein as the ERV copy number per giga base pairs (Gb) of the host genome, and host-ERV relationship may be related to this difference. This thesis will report a study aimed at contributing to understanding the host-ERV relationship during long-term evolution. This study consists of two parts.

Since the host immune system can take an important part in host-ERV evolutionary interaction, especially some innate immune receptors that have potential for recognizing retroviruses, the first part of this study is a case study of the functional evolution of innate immune receptors using the RIG-I-like receptors (RLRs) in birds. RLRs are pattern-recognition receptors for viral RNA and one of them, the retinoic acid-inducible gene I (RIG-I), is a potential sensor for retroviruses. Modes and intensity of natural selection of the coding genes of avian RLRs were examined to understand the roles of RLRs in bird evolution and bird-ERV evolutionary interaction. This part of my study provides results and discussion about the evolution of RLR genes in birds from aspects of conservation levels, positive selection modes, changes in selection intensity, and association between evolutionary rate of RLR genes and

endogenous retrovirus load; many of these results will be shown and discussed in comparison with those of mammals. In brief, the three RLR genes show distinct patterns of functional evolution but with possible influences to the evolution of each other and the gene encoding RIG-I evolved in correlation with endogenous retrovirus load in bird genomes. These findings suggest the possibility of interaction between host immunity and endogenous retroviruses in bird evolution.

The second part of my study takes a broader investigation at genome-wide scales on the evolutionary interaction between hosts and ERVs/retroviruses in mammals and birds. Phylogenetic gene-phenotype association analyses were applied to the gene evolutionary rate and ERV load, and gene set enrichment analyses (GSEA) based on the association results were performed to provide information about the relative weight of biological process in the evolutionary interaction of hosts with ERVs/retroviruses. From this study, I detected genes that evolved in association with ERV load in mammals and birds, separately, and revealed that the distribution of degrees of association between gene evolutionary rate and ERV load show a difference between mammals and birds, which indicate different levels of evolutionary interaction between mammals and birds. The genes that evolved in association with ERV load in both mammals and birds, as well as genes evolved in only one of the two groups, are reported. This part of my study also provides comparative insights into the evolutionary interaction between host genes and ERV loads in mammals and birds, with particular attention to the biological processes that have the highest potential for being host restrictions on ERV load. Such biological processes involve immune responses, gene silencing and DNA deletion. Genes showing high degrees of association between gene evolutionary rate and ERV load

and involved in these biological processes are also reported and discussed. Results of this part of my study suggest that gene silencing may play an important role in host-ERV evolutionary interaction, and that mammals and birds might evolve different strategies in immune responses to ERVs/retroviruses.

More detailed abstracts for the two parts of this study are present at the beginning of Chapter 2 and 3 of this thesis, respectively. Overall, this thesis provides evidence of host-ERV evolutionary interaction in mammals and birds, proposes explanations to the ERV load difference between mammals and birds, and supports the long history of host-ERV relationships comprising of a balance between host-parasite conflict, tolerance and co-option.

Table of Contents

Chapter 1

General Introduction

1.1 Host-Parasite Relationships in Change

The relationship between the host and the parasite has been a fascinating theme of biological studies. This relationship is considered as a war, where the parasite is the invader and the host the defender (Maizels 2009; Morand et al. 2015). Immunity is the defense reaction of a host's body against all threats to the body's integrity. Parasites constitute a substantial fraction of threatens and studies in immunology has accumulated tremendous knowledge about mechanisms of the defense against parasites since 1798 when Edward Jenifer published *An Inquiry into the Causes and Effects of the Variolae Vaccine* (Klein 1982). Nowadays, our knowledge about immunity has reached the molecular level in a variety of organisms as well as the genetics behind them. Immune systems in all kinds of organisms were established and developed through evolution. The immune system virtually exists in all cellular organisms and has a very ancient origin. Forms of immunity might have emerged and evolved independently and they can be grouped into two general categories: innate immunity and adaptive immunity. The innate immune system refers to the defense reaction to microorganisms triggered by a limited number of germline-encoded pattern-recognition receptors (PRRs) in eukaryotes (Akira et al. 2006). Besides the innate immune system, two types of adaptive immune system originated in vertebrates: jawed vertebrates generate a diverse repertoire of B and T cell antigen

receptors through the rearrangement of immunoglobulin V, D, and J gene fragments, whereas jawless fish assemble their variable lymphocyte receptors (VLR) through random combinations of leucine-rich repeat (LRR) modular units (Pancer and Cooper 2006). In prokaryotes (bacteria and archaea), clustered, regularly interspaced, short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems provide adaptive immunity against viruses and plasmids (Barrangou et al. 2007; Brouns et al. 2008).

The essence of immunity is discrimination of nonself from self. In ecological theories, parasitism is one form of symbiosis and the other two are mutualism and commensalism (Martin and Schwab 2013). All symbiotic partners are nonselves but the commensal gains the tolerance of the immune system of their partners. Evolutionary studies have shown that the relationship between two symbiotic organisms can shift (Nunes et al. 2018; Silknetter et al. 2019). In a symbiotic interaction undergoing a relationship shift, an organism can still be called a parasite while its actual role is ambiguous. More surprisingly, a parasite can even become ambiguous about being nonself. An example is the genomic parasite, endogenous viral elements (EVE), which are remnants of viral genomes inserted into host genomes. The most commonly found EVEs are derived from retroviruses, of which integration into the host genome is obligate for their replication cycle (Feschotte and Gilbert 2012; Weiss and Stoye 2013). These EVEs are endogenous retroviruses (ERV), which are present in vertebrates and believed to be the relics of historical retroviral infections. Some ERV-like elements are present in other eukaryotes including insects, yeasts and plants (Tanda et al. 1994; Britten 1995; Leblanc et al. 1997). Their origins vary and may be related to the origin of retroviruses (Malik et al. 2000). ERVs and ERV-like elements are also known as long terminal repeat (LTR)

retrotransposons (Wicker et al. 2007; Kapitonov and Jurka 2008) thus the boundary between the world of transposons and the world of viruses becomes nebulous (Kapusta et al. 2017).

Retroviruses present a complicated relationship with their hosts. On one hand they are pathogenic: retroviral infections can cause lethal diseases to vertebrates, such as the human immunodeficiency virus (HIV) and mouse mammary tumor virus (MMTV); insertion/transposition of ERVs may disrupt the host genome function in the same manner as other transposons (also named transposable elements, TE); and their expression may cause autoimmune diseases (Blomberg et al. 2013). On the other hand, certain properties of an ERV can be co-opted by the host for generating new functions (Frank and Feschotte 2017). Selection on the host may tend to eliminate the pathogenicity of ERVs while reserve opportunities for the host to co-opt them. There must exist a trade-off between the two forces driving the host evolution. No matter how, ERVs have influenced the evolution of their hosts in many aspects such as new function, transcriptional regulation, genome size and structure (Lowe et al. 2007; Slotkin and Martienssen 2007; Schmidt et al. 2012; Elliott and Gregory 2015; Wang et al. 2015; Frank and Feschotte 2017).

According to the literature presented above, the complicated relationship between hosts and ERVs represents a changeable host-parasite relationship and may shape their evolutionary interaction. In this thesis, evolutionary interaction refers to any kinds of association of changes between two subjects during evolution. This includes: (1) the one-way association, whereby the evolution of subject A is the cause or driving force of the evolution of subject B, which is to say, the evolution of subject

B is a consequence or response to the evolution of subject A; and (2) the two-way association, which is equivalent to co-evolution.

1.2 Host Gene Evolution and ERV load

The evolutionary interaction between hosts and ERVs can be indicated by the association between host gene evolutionary rate and ERV load during host evolution, which can be detected using phylogenetic association analyses (Pagel 1999). Phylogenetic association analysis tests the association between two traits in a monophyletic group of organisms. The point of this method is that the association resulting from the phylogenetic relationship will be excluded, and only the association resulting from functions will remain (Pagel 1999; Theodore Garland and Ives 2000). In a phylogenetic association analysis, host gene evolutionary rate can be presented by the ratio of non-synonymous substitution rate over synonymous substitution rate (dN/dS). Phylogenetic general least squares (PGLS) and phylogenetic independent contrasts (PIC) are two of the most commonly applied methods for phylogenetic association analyses. Though different in approaches, the two methods are highly consistent in results (Theodore Garland and Ives 2000). In the recent decade, high throughput sequencing has generated huge genome sequence data and promoted genomic research. Annotated coding gene sequences are available for a tremendous number of species and statistics of TEs, including ERVs, in vertebrate genomes are also available.

1.3 Aim of This Thesis

Overall, the relationship between hosts and ERVs may have led to a unique type of host-parasite evolutionary interaction. A question can be raised about how the evolution of immune systems and other biological functions of hosts interact with ERV load. This thesis will take advantage of existing genomic data to answer the above question with a case study of birds and mammals.

The rest part of my thesis consists of three chapters. Chapters 2 and 3 are research reports on two relatively independent topics, and Chapter 4 is a general discussion. Chapter 2 examines how selection has worked on RIG-I-like receptor genes in birds. As part of innate immunity, RIG-I-like receptors are intracellular nuclear acid sensors. Chapter 2 also serves as an exploratory study about the potential of association between host immunity and ERV loads. Chapter 3 reports a genome-wide evolutionary study on the host-ERV evolutionary interactions. The findings of Chapter 3 contribute the most to addressing the aim of this thesis.

References

Akira S, Uematsu S, Takeuchi O. 2006. Pathogen recognition and innate immunity. Cell 124(4):783–801.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science 315(5819):1709–1712.

Blomberg J, Ushameckis D, Jern P. 2013. Evolutionary aspects of human endogenous

retroviral sequences (HERVs) and disease. In: Madame Curie Bioscience Database [Internet]. Landes Bioscience.

Britten RJ. 1995. Active gypsy/Ty3 retrotransposons or retroviruses in Caenorhabditis elegans. Proc. Natl. Acad. Sci. 92(2):599–601.

Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin E V., Van Der Oost J. 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321(5891):960–964.

Elliott TA, Gregory TR. 2015. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos. Trans. R. Soc. B Biol. Sci. 370(1678):20140331.

Feschotte C, Gilbert C. 2012. Endogenous viruses: Insights into viral evolution and impact on host biology. Nat. Rev. Genet. 13(4):283.

Frank JA, Feschotte C. 2017. Co-option of endogenous viral sequences for host cell function. Curr. Opin. Virol. 25:81–89.

Kapitonov V V., Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat. Rev. Genet. 9(5):411–412.

Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. Proc. Natl. Acad. Sci. 114(8):E1460–E1469.

Klein J. 1982. Immunology: The science of self-nonself discrimination. Wiley.

Leblanc P, Desset S, Dastugue B, Vaury C. 1997. Invertebrate retroviruses: ZAM a new candidate in D. melanogaster. EMBO J. 16(24):7521–7531.

Lowe CB, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc. Natl. Acad. Sci. 104(19):8005–8010.

Maizels RM. 2009. Parasite immunomodulation and polymorphisms of the immune system. J. Biol. 8(7):62.

Malik HS, Henikoff S, Eickbush TH. 2000. Poised for contagion: Evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res. 10(9):1307–1318.

Martin BD, Schwab E. 2013. Current usage of symbiosis and associated terminology. Int. J. Biol. 5(1):32.

Morand S, Krasnov BR, Littlewood DTJ. 2015. Introduction. In: Morand S, Krasnov BR, Littlewood DTJ, editors. Parasite diversity and diversification. Cambridge, UK: Cambridge University Press. p. 1–8.

Nunes CEP, Maruyama PK, Azevedo-Silva M, Sazima M. 2018. Parasitoids turn herbivores into mutualists in a nursery system involving active pollination. Curr. Biol. 28(6):980–986.

Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature 401(6756):877–884.

Pancer Z, Cooper MD. 2006. The evolution of adaptive immunity. Annu. Rev. Immunol. 24:497–518.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonalves Â, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 148(1–2):335–348.

Silknetter S, Kanno Y, Kanapeckas Métris KL, Cushman E, Darden TL, Peoples BK. 2019. Mutualism or parasitism: Partner abundance affects host fitness in a fish reproductive interaction. Freshw. Biol. 64(1):175–182.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. 8(4):272–285.

Tanda S, Mullor JL, Corces VG. 1994. The Drosophila tom retrotransposon encodes an envelope protein. Mol. Cell. Biol. 14(8):5392–5401.

Theodore Garland J, Ives AR. 2000. Using the Past to Predict the Present: Confidence Intervals for Regression Equations in Phylogenetic Comparative Methods. Am. Nat. 155(3):346–364.

Wang J, Vicente-García C, Seruggia D, Moltó E, Fernandez-Miñán A, Neto A, Lee E, Gómez-Skarmeta JL, Montoliu L, Lunyak V V., et al. 2015. MIR retrotransposon sequences provide insulators to the human genome. Proc. Natl. Acad. Sci. 112(32):E4428–E4437.

Weiss RA, Stoye JP. 2013. Our viral inheritance. Science 340(6134):820–821.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8(12):973.

Chapter 2[1]

Functional Evolution of Avian RIG-I-Like Receptors

Abstract

RIG-I-like receptors (retinoic acid-inducible gene-I-like receptors, or RLRs) are family of pattern-recognition receptors for RNA viruses, consisting of three members: retinoic acid-inducible gene I (RIG-I), melanoma differentiation-associated gene 5 (MDA5) and laboratory of genetics and physiology 2 (LGP2). To understand the role of RLRs in bird evolution, I performed molecular evolutionary analyses on the coding genes of avian RLRs using filtered predicted coding sequences from 62 bird species. Among the three RLRs, conservation score and dN/dS (ratio of nonsynonymous substitution rate over synonymous substitution rate) analyses indicate that avian MDA5 has the highest conservation level in the helicase domain but a lower level in the caspase recruitment domains (CARDs) region, which differs from mammals; *LGP2*, as a whole gene, has a lower conservation level than *RIG-I* or *MDA5*. I found evidence of positive selection across all bird lineages in *RIG-I* and *MDA5* but only on the stem lineage of Galliformes in *LGP2*, which could be related to the loss of *RIG-I* in Galliformes. Analyses also suggest that selection relaxation

[1] Content of this chapter has been published as Zheng W, Satta Y. 2018. Functional Evolution of Avian RIG-I-Like Receptors. Genes (Basel). 9(9):456. Slight changes are made from the published article.

may have occurred in *LGP2* during the middle of bird evolution and the CARDs region of *MDA5* contains many positively selected sites, which might explain its conservation level. Spearman's correlation test indicates that root-to-tip dN/dS of *RIG-I* shows a negative correlation with endogenous retroviral load in bird genomes, suggesting the possibility of interaction between immunity and endogenous retroviruses during bird evolution.

2.1. Introduction

The innate immune system is the first-line defense of hosts when encountering infectious pathogens and it is phylogenetically ancient (Medzhitov and Janeway 1997). In the innate immune system, pattern-recognition receptors (PRRs) play a role in pathogen sensing by recognizing evolutionarily conserved molecular structures on pathogens (known as pathogen-associated molecular patterns, PAMPs) (Medzhitov and Janeway 1997). This pathogen recognition triggers the signaling pathways that eventually upregulate the expression of type I interferons, as well as proinflammatory cytokines and chemokines (Kawai and Akira 2010; Loo and Gale 2011). Examples of PRRs include Toll-like receptors (TLRs), NOD-like receptors (NLRs) and RIG-I-like receptors (retinoic acid-inducible gene-I-like receptors, or RLRs).

The RLRs are a family of three DExD/H box-containing RNA helicases and function as cytoplasmic PRRs sensing non-self RNA (Loo and Gale 2011). The RLRs are retinoic acid-inducible gene I (RIG-I), melanoma differentiation-associated gene 5 (MDA5), and laboratory of genetics and physiology 2 (LGP2) (Yoneyama and Fujita 2009). RIG-I and MDA5 consist of three functional domains (see the colored blocks

at the top of Figure 2.1): the DExD/H domain (or helicase domain) in the center is responsible for RNA recognition, the two caspase recruitment domains (CARDs) at the N-terminal are responsible for downstream signaling transduction, and the C-terminal domain (CTD) assists in pathogen recognition by binding specific viral RNA (Saito et al. 2007). Additionally, a repressor domain (RD) within the CTD is involved in the inhibition of RIG-I signaling in the absence of viruses, while MDA5 does not have an intact RD (Saito et al. 2007). MDA5 preferentially recognizes high-molecular-weight double-stranded RNA, while RIG-I preferentially recognizes shorter double-stranded RNA as well as single-stranded RNA (Kato et al. 2008; Yoneyama and Fujita 2009; Loo and Gale 2011). On the other hand, LGP2 lacks CARDs and therefore does not trigger immune responses but can up- or downregulate the signaling of RIG-I and MDA5 (Bamming and Horvath 2009; Satoh et al. 2010). The regulatory function of LGP2 is attributed to its retained helicase domain and RD (Venkataraman et al. 2007; Li et al. 2009).

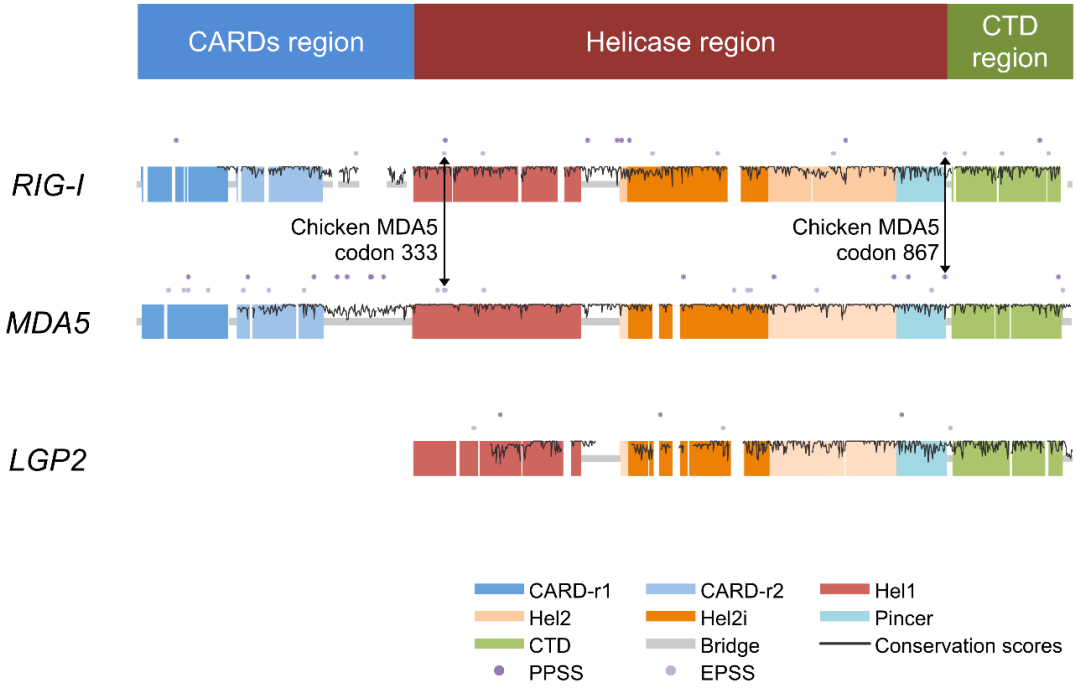


Figure 2.1. Overview of the three avian RIG-I-like receptors (RLRs). The double-sided black arrows indicate positively selected sites (PSSs) at identical positions in melanoma differentiation-associated gene 5 (*MDA5*) and retinoic acid-inducible gene I (*RIG-I*). The bar at the top indicates the locations of three regions of the RLR genes defined for convenience in this paper. Meanings of dots, colored blocks, and lines in the domain structure sketch of three avian RLRs are denoted in the bottom right of the figure. Conservation scores were calculated only for the sites containing <5% missing data and deletions in total.

RIG-I may have emerged prior to the appearance of vertebrates, while the other two RLR genes originated in vertebrates (Zou et al. 2009) by duplication of *RIG-I*. In recent years, the role of avian RLRs on major poultry diseases has been studied; they function as PRRs recognizing RNA viruses in the same manner as

mammalian RLRs (Barber et al. 2010; Liniger et al. 2012; Chen et al. 2013; Sun et al. 2013; Hayashi et al. 2014). It was reported that RIG-I expression increased upon avian influenza virus (AIV) infection in ducks (Barber et al. 2010) and geese (*Anser cygnoides*) (Sun et al. 2013), and in geese, RIG-I expression also increased upon Newcastle disease virus (NDV) infection (Sun et al. 2013). Chickens (*Gallus gallus*) showed weaker resistance to AIV and NDV infection than ducks and geese, and this could be attributed to the loss of RIG-I in chickens (Barber et al. 2010; Chen et al. 2013). In chickens, in which RIG-I has been lost, MDA5 was reported to function as the AIV sensor (Liniger et al. 2012). Chicken MDA5 was reported to preferentially sense short double-stranded RNA, which is usually done by RIG-I but not MDA5 (Hayashi et al. 2014). Studies also showed that in chicken cells, introduced duck RIG-I could trigger immune responses upon AIV infection (Shao et al. 2014) and introduced pigeon RIG-I could trigger immune responses upon AIV and infectious bursal disease virus (IBDV) infection (Xu et al. 2015). However, studies on the RLRs of other birds are still missing. In addition, the roles of RLRs in the long-term evolution of birds are not fully understood.

Since retroviruses have single-stranded RNA genomes and RIG-I can bind to single-stranded RNA, RIG-I has been considered among the candidate innate sensors of retroviruses (Hurst and Magiorkinis 2015). Recently, candidate sensors of retroviruses including RLRs were discussed in relation to their potential influence on endogenous retroviruses (ERVs) (Kassiotis and Stoye 2016). ERVs are sequences within the genome that are highly similar to retroviruses. After infecting a cell, retroviruses integrate into the host's genome and replicate through host cell machinery (Hayward and Katzourakis 2015; Saxena and Chitti 2016). If a retrovirus

invades and integrates into the germ-line and subsequently becomes transmitted vertically, it becomes an ERV (Hayward and Katzourakis 2015). ERVs can amplify their copy number via retro-transposition or re-infection (Belshaw et al. 2004; Bannert and Kurth 2006; Dewannieux et al. 2006; Magiorkinis et al. 2012); and as a result, 8% of the human genome (Lander et al. 2001) and 11% of the mouse genome (McCarthy and McDonald 2004) consists of ERVs. However, no studies have reported evidence that RIG-I triggers immune responses upon retroviral infection in mammalian cells. RIG-I-dependent pathways were reported to be inhibited in human immunodeficiency virus (HIV)-infected human cells, possibly by HIV proteases (Berg et al. 2012). In birds, an equivalent study is still lacking. Interestingly, however, birds have a much smaller amount of ERVs than mammals, ranging from 0.2% to 3.6% of the genome (Cui et al. 2014). Thus, it would be interesting to know whether RNA sensors are related to the variance of ERV load in hosts since RLRs may function against retrovirus integration during avian evolution.

Here, I report an evolutionary study of avian RLRs using the coding sequences from 62 bird species. I elucidate the evolutionary modes of avian RLRs and examine the evolutionary association of avian RLRs with ERV load. Our findings can provide a starting point for future evolutionary studies on the interaction between innate immunity and (endogenous) viruses. This interaction is also an important issue in studies of viral infection and inflammatory diseases. I believe that evolutionary perspectives, especially on organisms that play a role as reservoirs of human disease-causing viruses, such as birds, are informative to a wide range of studies aimed at improving human health.

2.2. Materials and Methods

2.2.1. Sequence Collection and Alignment

Because not all of the 62 bird genomes include RLRs that are fully or correctly annotated, I isolated coding sequences (CDSs) from the genomes of 62 bird species in the NCBI Reference Sequence Database (RefSeq) (O'Leary et al. 2016) using nucleotide BLAST (blastn). Exons of *Gallus gallus* Linnaeus, 1758 (chicken) *MDA5* and *LGP2*, and those of *Anser cygnoides* Linnaeus, 1758 (goose) *RIG-I* were used as queries for blastn searches. For each RLR gene, codon alignment was performed in Molecular Evolutionary Genetics Analysis (MEGA) version 7 (Kumar et al. 2016) (ClustalW) and profile codon alignment was performed in ClustalX version 2.1 (Larkin et al. 2007). Amino acid sequence alignment of the three avian RLR references was performed using MAFFT (strategy G-INS-I) (Katoh et al. 2002). In order to ensure the quality of predicted CDSs for analyses, I referred to the scaffolds where blastn hits were located to manually check the CDS starting, ending and exon-intron boundary regions so as to replace mistaken parts and delete unreliable parts. After this, I concatenated the edited blastn hits into predicted CDSs.

2.2.2. Six-Class Assessment of Sequence Face Quality

I assigned a predicted CDS into one of six classes from A to F indicating high to low sequence face quality according to the following criteria. Class A: the predicted CDS has both start and termination codons, no premature termination codon (PTC) or frame-shifting insertions and deletions (INDELs) and length ≥90% of the

alignment length (including INDELs). Class B: the predicted CDS is not in class A but has a sequence length ≥80% of the alignment length and a total fraction of PTCs and frame-shifting INDELs ≤0.05% of the alignment length. Class C: the predicted CDS is not in the above classes but with a length ≥70% of the alignment length. Class D: the predicted CDS is not in the above classes but with a length ≥50% of the alignment length. Class E: the predicted CDS is not in the above classes but consists of BLAST hits that cover ≥10% with an identity that is ≥70% of at least one query exon. Class F: the rest.

2.2.3. Datasets Preparation

According to the six-class assessment, I grouped the sequences into two datasets: dataset 1 of acceptable data quality (classes A–D; 54 RIG-I, 59 MDA5, and 59 LGP2 sequences) and dataset 2 of good data quality (classes A and B; 39 RIG-I, 57 MDA5, and 31 LGP2 sequences). If an analysis was susceptible to missing data, I applied dataset 2; otherwise, dataset 1 was applied. If an analysis could be susceptible to missing data, dataset 2 was applied; otherwise, dataset 1 was applied. I also generated dataset 3 by combining a subset of dataset 2 with mammal sequences retrieved from GenBank for the bird-mammal comparison. See Table 2.S1 for a detailed list of the datasets.

2.2.4. Molecular Evolutionary Analyses

Conservation scores of amino acid sites were calculated using Scorecons (Valdar 2002) with method valdar01. Average ratio of nonsynonymous substitution rate over synonymous substitution rate (dN/dS) values were estimated with the single likelihood ancestor counting (SLAC) method (Kosakovsky Pond and Frost 2005), and then calculated using the Nei–Gojobori (Nei and Gojobori 1986) method. Gene-wide positive selection was tested using the partitioning approach for robust inference of selection (PARRIS) method (Scheffler et al. 2006), a likelihood ratio test (LRT) of the alternative model in which a proportion of sites have evolved under an additional class of dN/dS > 1, against the null model in which the sites have evolved under a class of dN/dS ≤ 1. Episodic positive selection was tested using branch-site random effects likelihood (BSR), an LRT of the alternative model in which a proportion of sites have evolved under dN/dS > 1 on a specific branch (Figure 2.S1) (Kosakovsky Pond et al. 2011). The statistical significance was determined by the *p*-value corrected with the Holm–Bonferroni method (Holm 1979) of the LRT using the asymptotic distribution of a mixture of two χ^2 distributions (Kosakovsky Pond et al. 2011) and a corrected *p*-value < 0.05 was considered statistically significant.

Positively selected sites (PSSs) were detected using SLAC (Kosakovsky Pond and Frost 2005) and the mixed effects model of evolution (MEME) (Murrell et al. 2012). SLAC first reconstructs ancestral codons with maximum likelihood, and based on this reconstruction, the proportion of nonsynonymous substitutions to all substitutions is tested at each codon against the mean value. A significant excess to the mean value would indicate positive selection at that site. Statistical significance of

the excess is determined by the *p*-value using an extended binomial distribution (Kosakovsky Pond and Frost 2005). In contrast, MEME applies a branch-site random effects phylogenetic framework that allows the distribution of dN/dS to vary from site to site as well as from branch to branch, which allows MEME to identify instances of both episodic and pervasive positive selection. Among the parameters of MEME, there is a category containing an unrestricted dN parameter for an alternative model, while the null model has the dN parameter of this category restricted to being ≤dS. Statistical significance of the alternative model at a site will indicate positive selection, and the significance is determined by the *p*-value of the LRT using the asymptotic distribution of a mixture of three χ^2 distributions (Murrell et al. 2012). For both methods, a *p*-value < 0.05 can be basically considered as statistically significant evidence of PSSs. Further, I made an integrative determination of PSSs based on this basic criterion. Results from MEME analyses ($p < 0.05$) covers all the PSSs detected by SLAC ($p < 0.05$); thus, I categorized these PSSs detected by both methods as pervasive positively selected sites (PPSSs). I simultaneously categorized the sites detected with $p < 0.01$ in MEME and $p \in (0.05, 0.1)$ in SLAC as episodic positively selected sites (EPSSs). The criterion of EPSS was cautiously set due to the lack of alternative methods with sufficient power equivalent to MEME to detect EPSSs. To detect the PSSs of birds, dataset 2 was used.

RELAX (Wertheim et al. 2015) was used to detect changes of selection intensity. RELAX assumes that positive and/or negative (purifying) selection, if it exists, would be under the same relaxation or intensification. I denote positive and/or negative selection as 'selection' for short in this paper. In RELAX, the intensity of selection of two appointed groups of branches in a phylogenetic tree is compared. The

result of the comparison is expressed with an optimized parameter K, whereby $K > 1$ indicates an intensified and $K < 1$ indicates a relaxed selection of the test group relative to the reference group, as is in the alternative model, while the null model shows $K = 1$. Statistical significance of the alternative model is determined by the *p*-value of the LRT using standard χ^2 asymptotic distribution (Wertheim et al. 2015) and a *p*-value < 0.05 is considered statistically significant. Fifty million years ago (MYA) were chosen as the boundary of old and young branch groups because it is located approximately in the middle of the estimated time of ~102 MYA, tracing back to the ancestor of extant birds (Jarvis et al. 2014). If a branch ended no later than the boundary, it was assigned to the old branch group (test group); otherwise, it was assigned to the young branch group (reference group). For RELAX analysis of the comparison between birds and mammals, I made a joint alignment of bird and mammal CDSs for each of the three RLRs.

The necessary phylogenetic information of the 62 birds for analyses were comprised of the phylogeny reconstructed from 48 birds with whole-genome data (Jarvis et al. 2014), the suggested topological positions by BirdTree (BirdTree 2018) and the suggested divergence times by TIMETREE (Kumar et al. 2017)for other species. The phylogenetic information of 10 mammals was also cited from a reported phylogeny reconstruction with whole-genome data (Meredith et al. 2011). PARRIS, BSR, SLAC, MEME and RELAX analyses were performed with HyPhy (Kosakovsky Pond et al. 2005) on the webserver Datamonkey (Delport et al. 2010). Protein structure images were processed and exported using Chimera (Pettersen et al. 2004).

2.2.5. Correlation Analysis of Root-to-Tip dN/dS of Avian RLRs vs. ERV Load

Spearman's correlation tests were performed on root-to-tip dN/dS versus ERV load and a p-value < 0.05 was considered statistically significant. The dataset comprised the species in dataset 1 that intersected with the 48 species (Jarvis et al. 2014) with phylogenetic relationships reconstructed using whole-genome sequences (Table 2.S1). Here, root-to-tip dN/dS was used as an index of long-term average functional constraint. The ancestral sequences for calculating the root-to-tip dN/dS were inferred from dataset 1 with the maximum likelihood method in MEGA 7 (Kumar et al. 2016), and *Alligator mississippiensis* Daudin, 1802 (American alligator) was used as an outgroup. ERV load was defined as the ERV copy number divided by the genome size (Gb) of a species and was calculated based on published data (Cui et al. 2014). Since the significance may be biased if two biological traits were not taken independently from a common distribution but from a branching phylogeny, the data used in correlation tests had been transformed into normalized phylogenetic independent contrasts (PICs) (Felsenstein 1985) using DendroPy (Sukumaran and Holder 2010) with known tree topology and branch lengths.

2.3. Results

2.3.1. Face Quality of Predicted CDSs

Since genome annotations are not completely satisfactory, predicted CDSs were isolated from the genome data of 62 bird species (Table 2.S1) using BLAST.

Subsequently, manual procedures were applied to optimize the quality of data for evolutionary analyses (see Materials and Methods for details).

I assessed the face quality of each predicted CDS, namely, the extent of completeness as a practical CDS used in the analysis. According to a series of subjective criteria (see Materials and Methods for details), I assigned each predicted CDS to one of the six classes (A–F; high to low face quality). *MDA5* had the largest number of cases in which a species has a predicted CDS of good face quality in class A or B (n = 57) and *LGP2* had the smallest (n = 31), while *RIG-I* had the largest number of cases of the worst class F (n = 8) and *LGP2* had the largest number of cases of the second worst class E (n = 12). The six-class assessment also showed variety among species (Figure 2.S1). Even though the *Meleagris gallopavo* Linnaeus, 1758 (turkey), chicken, *Coturnix japonic*a Temminck and Schlegel, 1849 (Japanese quail) and *Chaetura pelagica* Linnaeus, 1758 (chimney swift) had finely assembled genomes, their *RIG-I* was assigned to class F, which suggest putative loss of *RIG-I*.

2.3.2. Gene-Wide Conservation and dN/dS Levels of Avian RLRs

To evaluate the conserved mode of evolution of the three avian RLRs, average conservation scores and average dN/dS ratios were calculated.

I first calculated the conservation scores of amino acid sites of the three RLRs (dataset 2). Between the two signaling receptors, MDA5 (0.933 ± 0.096, 43%) showed a slightly higher average conservation score and proportion of invariant sites than those of RIG-I (0.913 ± 0.114, 37%) (Table 2.1). I then excluded the CARDs region (defined as the two CARDs with in-between or flanking nondomain regions in

this study), such that the scores became comparable with LGP2, and found that MDA5 (0.953 ± 0.080, 50%) showed a much higher average conservation score and proportion of invariant sites than RIG-I (0.915 ± 0.112, 37%) and LGP2 (0.881 ± 0.139, 36%). Notably, MDA5 had a lower conservation level than RIG-I in CARDs; therefore, the exclusion of the CARDs resulted in a further lifted average conservation score for MDA5 (Table 2.1). When I analyzed the helicase region only (helicase domain with the pincer in this study), this region contributed the most to the leading conservation level of MDA5 over the other two. For both the helicase and CTD regions (here, this refers to the C-terminal domain and its flanking nondomain regions), LGP2 showed a slightly lower conservation level than RIG-I and MDA5 (Table 2.1). Each of the three avian RLRs had an average conservation score of nearly or over 0.9 and a proportion of invariant sites of over 30%, which indicates an overall high conservation level in the three avian RLRs. Conservation scores across sites are indicated in Figure 2.1.

Table 2.1. Conservation Scores of The Three Avian RIG-I-Like Receptors (Retinoic Acid-Inducible Gene-I-Like Receptors, or RLRs).

Domain regions	RIG-I	MDA5	LGP2
Three regions			
Avg. ± S.D.	0.913 ± 0.114	0.933 ± 0.096	N/A
Invariant/all	37%	43%	
Without CARDs region			
Avg. ± S.D.	0.915 ± 0.112	0.953 ± 0.080	0.881 ± 0.139
Invariant/all	37%	50%	36%
CARDs region			
Avg. ± S.D.	0.903 ± 0.121	0.861 ± 0.113	N/A

Invariant/all	35%	15%	
Helicase region			
Avg. ± S.D.	0.914 ± 0.111	0.955 ± 0.079	0.883 ± 0.138
Invariant/all	36%	53%	35%
CTD region			
Avg. ± S.D.	0.916 ± 0.116	0.942 ± 0.085	0.877 ± 0.141
Invariant/all	41%	39%	39%

RIG-I: retinoic acid-inducible gene I; MDA5: melanoma differentiation-associated gene 5; LGP2: laboratory of genetics and physiology 2. SD: standard deviation; CARDs: caspase recruitment domains; CTD: C-terminal domain.

I checked dataset 2 for the conservation level of two ubiquitinated sites in duck CARDs, Lys167 and Lys193 (duck site number) (Miranzo-Navarro and Magor 2014). Polyubiquitin chains that are attached to lysine residues by the ubiquitin ligase TRIM25 are necessary for CARDs activation in humans but seem unnecessary in ducks; instead, noncovalent ubiquitin chains may play more important roles (Miranzo-Navarro and Magor 2014). I found that Lys167 is invariant among the alignment of dataset 2, while Lys193 shows substitutions in the two crow species (Thr) and the carmine bee-eater (Glu).

Next, I estimated the average dN/dS ratio of the three avian RLRs (dataset 2) with two methods: the SLAC (Kosakovsky Pond and Frost 2005) and Nei–Gojobori methods (Nei and Gojobori 1986). With full sequences of bird RLRs, *MDA5* shows a similar dN/dS to that of *RIG-I*, with their 95% confidence intervals (CI) based on the SLAC method largely overlapping (Table 2.2). This result shows consistency with that of a smaller dataset (dataset 3, see Figure 2.S2 and Materials and Methods). I

looked into dataset 3 for some more details. When the CARDs region is excluded, dN/dS of *RIG-I* does not change much, while that of *MDA5* decreases and becomes lower than that of *RIG-I* or *LGP2*. This tendency is consistent with the observation of conservation scores. The dN/dS values of all three avian RLRs are higher than the average level of avian protein coding genes (Figure 2.S3) (Zhang et al. 2014). However, the dN/dS values of the RLRs are much smaller than 1, which is supported by the 95% CIs of SLAC (Table 2.2). This indicates that purifying selection is still the dominant mode of selection acting on the three avian RLRs.

Table 2.2. Mean dN/dS of RLR genes in birds and mammals.

Domain regions and datasets		No. of Species *RIG-I/MDA5/LGP2*		dN/dS (95% Confidence Interval) *RIG-I*	*MDA5*	*LGP2*
Dataset 2 and literature						
All regions	Birds (dataset 2)	39/57/31	†	0.385 (0.368, 0.401)	0.369 (0.355, 0.383)	0.222 (0.211, 0.234)
	Mammals (Cagliani et al. 2014)	42/46/46	‡	0.403 (0.390, 0.416)	0.293 (0.284, 0.302)	0.221 (0.213, 0.230)
Dataset 3						
Three regions	Birds	9/12/7	†	0.350 (0.325, 0.376)	0.384 (0.357, 0.411)	—
			‡	0.237	0.237	—
	Mammals	7/10/8	†	0.408 (0.374, 0.443)	0.330 (0.306, 0.356)	—
			‡	0.283	0.225	—
Without CARDs region	Birds	9/12/7	†	0.352 (0.323, 0.383)	0.284 (0.257, 0.312)	0.200 (0.180, 0.220)
			‡	0.237	0.185	0.253
	Mammals	7/10/8	†	0.370 (0.334, 0.410)	0.260 (0.234, 0.288)	0.240 (0.219, 0.264)
			‡	0.262	0.180	0.254

†: SLAC method. ‡: Nei–Gojobori method.

Conservation score and dN/dS results together suggest that for avian MDA5, the CARDs region might have experienced weaker functional constraint or stronger positive selection while the helicase and CTD regions may have experienced stronger functional constraint or weaker positive selection compared with the other two RLRs. Conservation score results suggest that avian LGP2 may have faced slightly weaker functional constraint or slightly stronger positive selection compared with the helicase-CTD regions of the other RLRs.

From the comparison of contrasting patterns of dN/dS between birds and mammals without the CARDs region, I found that birds and mammals share similar dN/dS levels for each of the three RLRs (dataset 3, Table 2.2). However, different from the pattern observed in birds, with the CARDs region, dN/dS values of *RIG-I* and *MDA5* in mammals both increase and thus *RIG-I* still has higher dN/dS than *MDA5* (reported mammal data (Cagliani et al. 2014) and dataset 3). Such contrasting patterns are consistent between the results of the two methods (SLAC and Nei–Gojobori) and are supported by the 95% CI of SLAC.

2.3.3. Positive Selection and Positively Selected Sites in Avian RLRs

To evaluate the extent of adaptive evolution, I examined gene-wide positive selection. Using a *p*-value of 0.05 as the significance level, positive selection was detected in *RIG-I* ($p = 1.6 \times 10^{-4}$) and *MDA5* ($p = 4.7 \times 10^{-10}$) but not in *LGP2* ($p = 0.999$). However, episodic (lineage-specific) positive selection (corrected $p = 0.014$) was detected in *LGP2* in the stem lineage of the Galliformes (the ancestral branch of the chicken, Japanese quail, and turkey) (Figure 2.S4).

To determine the sites that are vital for adaptive evolution, I performed positively selected sites (PSSs) detection of two categories, PPSSs and EPSSs (see Materials and Methods), based on the results of two distinct detecting methods, SLAC and MEME (Table 2.3) (Murrell et al. 2012). *MDA5* has the highest number and density of PPSSs and EPSSs, while *LGP2* has the lowest (Figures 1 and 2). This again suggests that the level of positive selection acting on *LGP2* is lower than the other RLRs, which cannot explain the lower conservation level of *LGP2* compared with the

other two. When comparing *MDA5* and *RIG-I*, the density of PSSs (number of PSSs over total number of sites, expressed as %) differed the most in the CARDs region: *MDA5* had 15 PSSs (5.05% of the region), while *RIG-I* had two PSSs (0.82% of the region), differences which were not marked in other regions (Figure 2.1). This suggests that positive selection may be the cause of the lower conservation level of the *MDA5* CARDs region rather than that of *RIG-I.*

Table 2.3. Number of Positively Selected Sites Identified in the Three Avian RLR genes.

Sites	No. Codons (Proportion to the Alignment)		
	RIG-I	*MDA5*	*LGP2*
Alignment	933 (100%)	1025 (100%)	677 (100%)
PPSS	8 (0.9%)	14 (1.4%)	3 (0.4%)
EPSS	10 (1.1%)	17 (1.7%)	3 (0.4%)
Total	18 (1.9%)	31 (3.0%)	6 (0.9%)

PPSS: pervasive positively selected sites; EPSS: episodic positively selected sites.

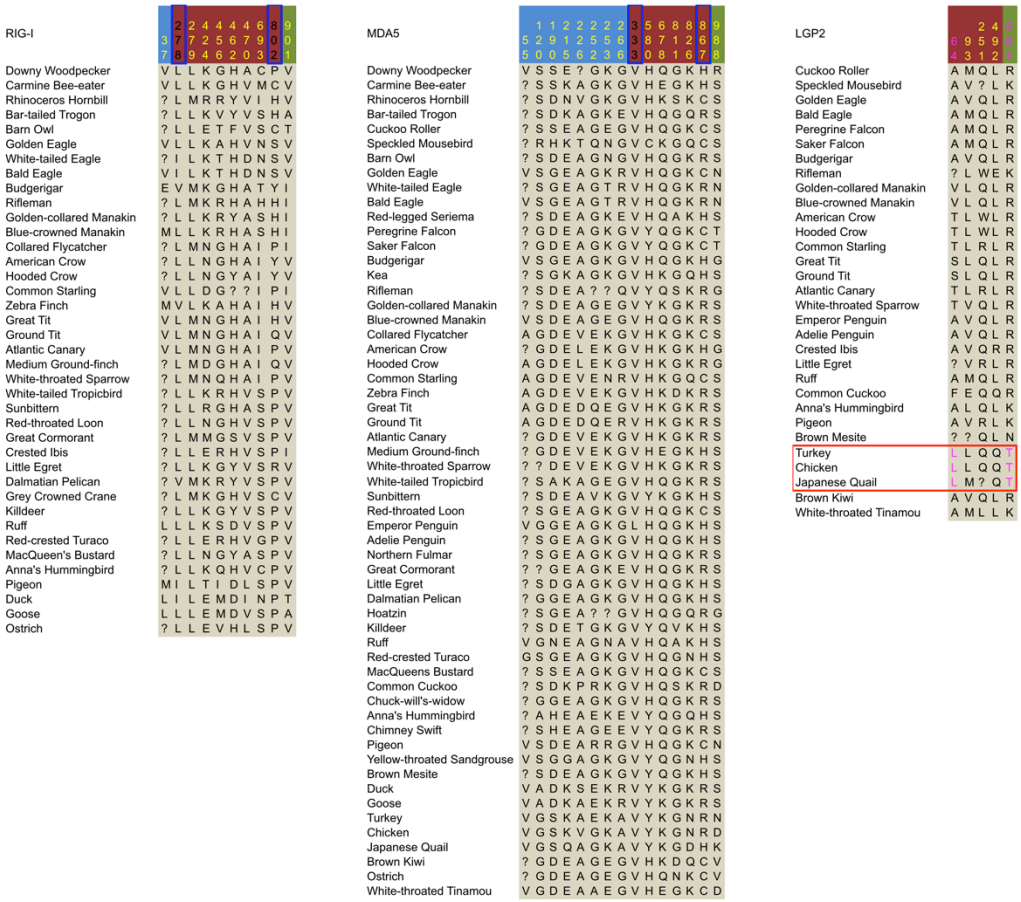


Figure 2.2. Summary of positively selected sites identified in the three avian RIG-I-Like receptors. Colors of site numbers indicate the type of positively selected sites (PSSs): yellow corresponds to pervasive positively selected sites (PPSSs), magenta corresponds to the episodic positively selected sites (EPSSs) showing specific substitutions in the LGP2 of Galliformes, and black corresponds to other EPSSs. The LGP2 of Galliformes is indicated by a red rectangular frame. The blue frames on the site numbers indicate PSSs located identically in RIG-I and MDA5. Site numbers are according to chicken sequences for MDA5 and LGP2, and the goose sequence for RIG-I. Background colors of site numbers represent coding sequence (CDS) regions defined for convenience in this

paper: sky blue for the caspase recruitment domains (CARDs) region, dark red for the helicase region, and green for the C-terminal domain (CTD) region.

I further examined EPSSs showing specific amino acid substitutions in the *LGP2* of Galliformes, since *LGP2* displayed episodic positive selection in the stem lineage of the Galliformes, and such EPSSs can be important for the functional adaptation of LGP2 in Galliformes. I found two EPSSs showing specific substitutions in Galliforme LGP2, with one substitution (Ala64Leu, chicken site number used) located in the CARDs region and the other (Arg587Thr) in the helicase region (Figure 2.2). Among all the substitutions on either of the two sites in the alignment, the amino acid replacement that could lead to the Galliforme-specific substitution has a relatively small rate in the Dayhoff matrix (Dayhoff 1978). Particularly for Arg587Thr in Galliformes, not only is the corresponding amino acid replacement rate much lower than other substitutions occurring at this site, but also the chemical nature is uniquely changed from acidic to neutral nonpolar. Since chicken Thr587 is located in the 3′ end-binding loop (Uchikawa et al. 2016), this substitution may have changed the double strain RNA (dsRNA) end-binding function of LGP2 in Galliformes.

Three-dimensional images of duck RIG-I (Kowalinski et al. 2011), chicken MDA5 and LGP2 with bound RNA (Uchikawa et al. 2016) showed that almost all of the PPSSs and abovementioned EPSSs are exposed on the surface of the proteins, and none of them are located within close contact to the bound RNA (Figure 2.3). This suggests that these PPSSs and EPSSs may play a role in minor adjustments to the

process of pathogen recognition and CTD regulation rather than direct changes to the core function of pathogen binding.

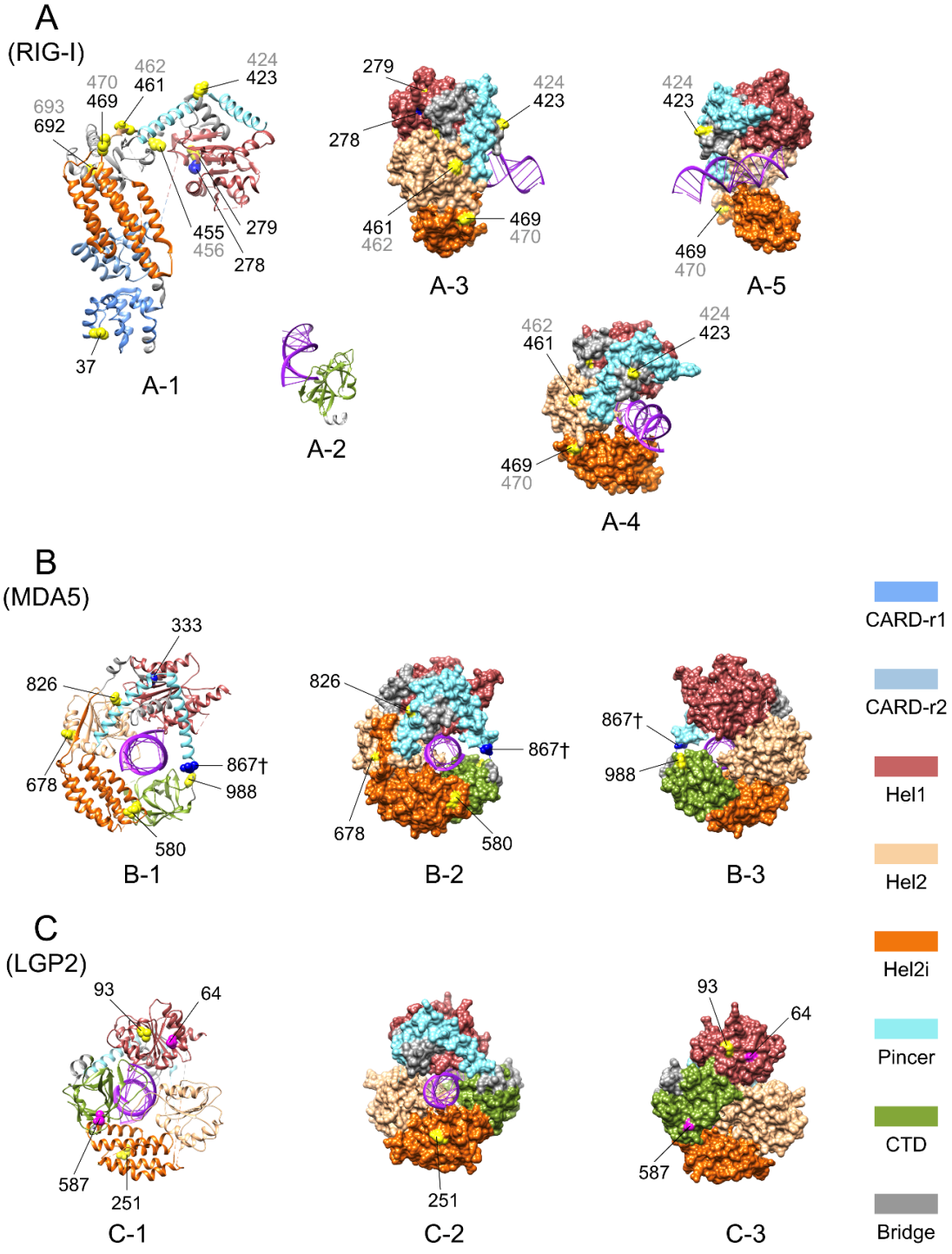


Figure 2.3. Three-dimensional (3d) structures of avian RIG-I-like receptors with some positively selected sites (PSSs) marked. Pervasive positively selected sites (PPSSs), the episodic positively selected sites

(EPSSs) with specific substitutions in LGP2 of Galliformes, and the PSSs located identically in MDA5 and RIG-I are mapped on available 3D protein structure images of duck RIG-I, partial chicken MDA5, and LGP2 (the MDA5 image does not include the caspase recruitment domains (CARDs) region; some PSSs are within the missing regions of the images and thus cannot be shown). (A) Duck RIG-I. Original image PDB IDs: 4a2w for A-1, 4a2x for A-2 and 4a36 for A-3, A-4 and A-5. In (A), A-1 shows the ribbon diagram of domains except the C-terminal domain (CTD) and A-2 shows the CTD; A-3 to A-5 show surface displays of the helicase domain from various angles. (B) Chicken MDA5. Original image PDB ID: 5JCH. (C) Chicken LGP2. Original image PDB ID: 5JBJ. In (B) and (C), from left to right, the ribbon diagram, surface display diagram and the surface display diagram rotated 180° on the vertical axis are shown. Colors of domains are denoted in the bottom right of the figure. Sites marked in yellow or with † appended to the site number represent PPSSs; sites in magenta represent EPSSs with specific substitutions in LGP2 of Galliformes; sites in blue represent PSSs located identically in RIG-I and MDA5. Marked sites are shown with the sphere display of side chains for the purpose of highlighting. Site numbers pointed to marked sites correspond to the chicken sequence for MDA5 and LGP2, to the duck sequence in black, and to the goose sequence in grey if different from that of duck for RIG-I.

Other PSSs may also have interesting functional effects. Two PSSs are located identically in both MDA5 and RIG-I (Figures 1 and 2) and are located separately at each of the two tails of the helicase region. Additionally, the only PSS (a PPSS) in RIG-I CARDs is located within the spliced-out sequence (namely the skipped exon 2 that encodes part of the first CARD) of a splice variant observed in duck (Miranzo-Navarro and Magor 2014). This splice variant also exists in mammalian RIG-I (Gack et al. 2008) and may also exist in other birds. Since the incomplete CARDs of the splice variant cannot interact with TRIM25, its CARDs cannot be ubiquitinated covalently or noncovalently and thus are not activated (Miranzo-Navarro and Magor 2014). The reason why these sites experienced positive selection needs future investigation.

2.3.4. Changes of Selection Intensity in Avian RLR Evolution

To examine whether relaxation of functional constraint occurred and contributed to the contrasting patterns of conservation level among the three RLRs in birds, I tested selection intensity change using RELAX (Wertheim et al. 2015). RELAX compares the selection (positive and/or negative, see Materials and Methods) intensity between two groups of branches in a phylogeny. In our study, each RELAX analysis was performed independently in each of the RLRs.

First I considered the possibility that selection relaxation on *LGP2* or *RIG-I* started in the ancestor of birds, by comparing the selection intensity between the groups of birds and mammals of dataset 3. If selection relaxation were detected in the bird group compared with the mammal group, it would suggest that selection

relaxation began in the ancestor of birds. The results show that no significant (0.05 level) difference of selection intensity was detected on *LGP2* or *RIG-I*, suggesting that selection relaxation did not occur in the ancestor of birds and thus is not a contributor to the lower conservation level of *LGP2* or *RIG-I* relative to *MDA5*.

I then examined whether relaxation could occur during the middle of bird evolution on *LGP2* or *RIG-I* by comparing the intensity of selection between old and young branches of *LGP2* and *RIG-I* in dataset 2. I divided the branches into old branch (test) group and young branch (reference) group. The old branch group represents the earlier stages of bird evolution, while young branch group represents the later stages. When all regions of each gene were used, selection relaxation on the young branch group compared with the old branch group was detected on *LGP2* (K = 0.82, $p = 0.042$) but not on *RIG-I*. This supports the notion that selection relaxation on *LGP2* might have occurred in the middle of bird evolution and contributed to the lower conservation level of *LGP2* relative to *MDA5*.

Furthermore, I examined whether the CARDs region of *MDA5* has experienced selection relaxation during bird evolution. I compared selection intensity on the *MDA5* CARDs region between birds and mammals, as well as between the old and young branches. No significant (0.05 level) relaxation was detected in either test, suggesting that selection relaxation at the beginning or middle of bird evolution did not occur and thus could not contribute to the lower conservation level of the CARDs region of avian *MDA5*.

2.3.5. Correlation between Functional Constraint of Avian RLR Genes and ERV Load

The small ERV load of birds (compared with mammals) and the RNA sensing ability of RLRs raised our query about whether an association exists between the functional constraint of avian RLR genes and ERV load during bird evolution. To address this, I performed Spearman's correlation test on root-to-tip dN/dS of avian RLRs versus ERV load. ERV information was retrieved from a reported study in which endogenous viral elements including ERVs in 48 birds were mined using a library of representative viral protein sequences derived from a known species list (Cui et al. 2014). To avoid effects of phylogenetic relationships among samples, phylogenetic independent contrasts (PICs) of root-to-tip dN/dS of avian RLRs and ERV load were used in the correlation tests. With a *p*-value of 0.05 as the significance level, a negative correlation was found between the ERV load and the root-to-tip dN/dS of *RIG-I* ($\rho = -0.3698$, $p = 0.019$) (Figure 2.4) but not *MDA5* or *LGP2*. Since low dN/dS values indicate high functional constraint, this result suggests that a positive association may exist between ERV load and functional constraint on *RIG-I* during avian evolution.

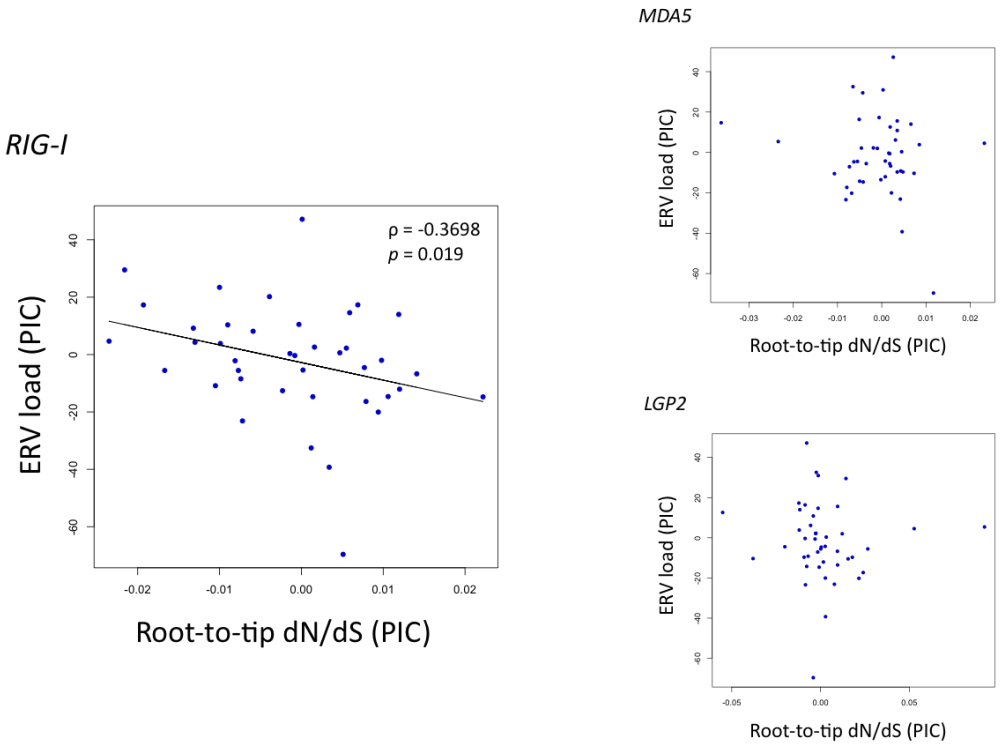


Figure 2.4. Relationship between the phylogenetic independent contrasts (PICs) of endogenous retroviruses (ERV) load and root-to-tip dN/dS of three avian RIG-I-like receptors. When Spearman's correlation has a $p < 0.05$, a trend line is shown. Note that the slope is not equal to Spearman's rank correlation coefficient ("ρ" in the figure).

Furthermore, I investigated the location of the nodes with a difference larger than 30 or 20 of the ranks between root-to-tip dN/dS and ERV load out of the 40 nodes in the tree (Figure 2.4). I found that these nodes are located widely across the tree without gathering in specific clades (Figure 2.S5), indicating that the correlation is contributed from various bird lineages and reflects a pervasive evolutionary mode in birds.

2.4. Discussion

The conservation and the dN/dS levels identified in our study suggest that purifying selection is the major driving force in the evolution of avian RLRs. However, evolutionary modes differ among the three avian RLR genes. Avian *MDA5* may be more functionally conserved and may have been under the strongest purifying selection among the RLRs, since *MDA5* shows the lowest level of dN/dS and its encoded amino acid sequences show the highest level of conservation, especially in the helicase-CTD region. *MDA5* may have also been under the strongest positive selection among the three, since: (1) it has the highest number and density of PSSs and (2) it exhibits a higher PSS density but not selection relaxation, which may explain the lower conservation level in its CARDs region. The above suggestions imply that *MDA5* may not only bear constant functional importance but may also be a hotspot of genetic adaptation in immune pathways during bird evolution. The exceeded PSS density of *MDA5* over *RIG-I* in the CARDs region (especially the nondomain region following CARDs) suggests that signal transduction behavior can be an important aspect in MDA5-related adaptation apart from pathogen recognition.

On the other hand, avian *RIG-I* seems to have undergone a lesser degree of natural selection than *MDA5*. However, its evolution has unique characteristics. Since the *RIG-I* sequences of the three Galliformes in our datasets (turkey, chicken, Japanese quail) and the chimney swift were from high-quality genome assembles but were assigned to the lowest face quality class F, *RIG-I* loss is suggested to have occurred in these species. Since *RIG-I* loss has been reported in chickens (Barber et al. 2010), it would be interesting to also confirm *RIG-I* loss in turkeys and Japanese

quails. If confirmed, *RIG-I* loss might have occurred in the stem lineage of Galliformes. The *RIG-I* loss in chimney swifts also needs to be confirmed. If those losses were confirmed, an interesting question would be raised: whether or not those losses of genes were adaptive, and if not, how the losses were compensated. More interestingly, a positive association was identified between the functional constraint on *RIG-I* and ERV load. This association suggests a certain interaction between *RIG-I* and ERVs during bird evolution. For example, *RIG-I* might respond to expressed ERVs. This possibility is worthy of further study since previous studies in mammals showed that RIG-I is the only RLR that can detect single-stranded RNA (Kato et al. 2008; Yoneyama and Fujita 2009; Loo and Gale 2011), which is the exact component of transcribed ERVs as well as genomes of retroviruses. RIG-I was also reported to have induced immune responses against the introduced genome of the retrovirus HIV in human cells (Berg et al. 2012). Since knowledge of the innate immunity against retroviruses/ERVs is still limited, our preliminary study here can inform future studies on innate immunity against retroviruses and the evolutionary role of ERVs in vertebrates.

As for avian *LGP2*, I found a heterogeneous evolutionary mode, including episodic positive selection on *LGP2* in the stem lineage of Galliformes. This could be related to the putative loss of *RIG-I* in the Galliforme ancestor as mentioned above. Since MDA5 does not have an intact RD and can be regulated by the RD of LGP2, loss of *RIG-I* might change the regulatory manner of LGP2 on MDA5 through evolution with a footprint of positive selection. In addition, selection (positive and/or negative) intensity might have intensified in the origin of birds and relaxed in the

middle of bird evolution. This relaxation might have led to the relatively low conservation level of *LGP2* among the three RLRs.

I also found evidence supporting the idea that evolutionary modes of RLRs may partially differ between birds and mammals. In mammals, *RIG-I* has higher dN/dS than *MDA5*; when the CARDs region was excluded, all dN/dS values decrease but such contrasting patterns remain. However, in birds, *RIG-I* has a similar dN/dS to that of *MDA5* and only the dN/dS of *MDA5* decreases after excluding the CARDs region (that is, the dN/dS of *MDA5* becomes lower than that of *RIG-I*). The difference between birds and mammals could be related to the differences in the evolutionary modes among gene regions and the differences between birds and mammals in the functional balance among the three RLRs during evolution.

This study reports the first evidence of an evolutionary association between RIG-I and ERV load. Future work is needed to investigate whether such an association also occurs in mammals and other vertebrates. The expansion of ERVs is one of the characteristics of mammalian genome evolution and is related to the complicated evolutionary interaction between host and ERVs/retroviruses (Kassiotis and Stoye 2016). Although most ERVs are defective due to mutations, some of them retain the capacity for expression and even infection. Moreover, the expressed ERVs could be related to numerous inflammatory diseases (Hurst and Magiorkinis 2015; Kassiotis and Stoye 2016). On the other hand, ERVs are sometimes co-opted by the host and become an important source of functional or regulatory innovation, as is known in mammalian evolution (Chuong et al. 2016; Kassiotis and Stoye 2016). Regarding the role of the immune system in the evolution of ERV load, a number of questions could be asked: Can evolutionary interactions between hosts and

ERVs/retroviruses restrict the innate immune mechanisms against retroviruses? Could evolutionary interactions between hosts and ERVs/retroviruses in birds be less complex than that in mammals since avian genomes contain a smaller proportion of ERVs? Do some innate immune sensors and pathways function more effectively against ERVs/retroviruses in birds than in mammals as a consequence? I believe that the further exploration of molecular evolutionary signals is one of the important approaches to answering such questions.

References

Bamming D, Horvath CM. 2009. Regulation of signal transduction by enzymatically inactive antiviral RNA helicase proteins MDA5, RIG-I, and LGP2. J. Biol. Chem. 284(15):9700–9712.

Bannert N, Kurth R. 2006. The evolutionary dynamics of human endogenous retroviral families. Annu. Rev. Genomics Hum. Genet. 7:149–173.

Barber MRW, Aldridge JR, Webster RG, Magor KE. 2010. Association of RIG-I with innate immunity of ducks to influenza. Proc. Natl. Acad. Sci. 107(13):5913–5918.

Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M. 2004. Long-term reinfection of the human genome by endogenous retroviruses. Proc. Natl. Acad. Sci. 101(14):4894–4899.

Berg RK, Melchjorsen J, Rintahaka J, Diget E, Søby S, Horan KA, Gorelick RJ, Matikainen S, Larsen CS, Ostergaard L, et al. 2012. Genomic HIV RNA induces innate immune responses through RIG-I-dependent sensing of secondary-

structured RNA. PLoS One 7(1):e29291.

BirdTree. 2018. BirdTree. Available from: https://birdtree.org

Cagliani R, Forni D, Tresoldi C, Pozzoli U, Filippi G, Rainone V, De Gioia L, Clerici M, Sironi M. 2014. RIG-I-like receptors evolved adaptively in mammals, with parallel evolution at LGP2 and RIG-I. J. Mol. Biol. 426(6):1351–1365.

Chen S, Cheng A, Wang M. 2013. Innate sensing of viruses by pattern recognition receptors in birds. Vet. Res. 44(1):82.

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 351(6277):1083–1087.

Cui J, Zhao W, Huang Z, Jarvis ED, Gilbert MTP, Walker PJ, Holmes EC, Zhang G. 2014. Low frequency of paleoviral infiltration across the avian phylogeny. Genome Biol. 15(12):539.

Dayhoff MO. 1978. Observed frequencies of amino acid replacements between closely related proteins. In: Atlas of Protein Sequence and Structure. Vol. 5. Washington D.C.: National Biomedical Research Foundation.

Delport W, Poon AFY, Frost SDW, Kosakovsky Pond SL. 2010. Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26(19):2455–2457.

Dewannieux M, Harper F, Richaud A, Letzelter C, Ribet D, Pierron G, Heidmann T. 2006. Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res. 16(12):1548–1556.

Felsenstein J. 1985. Phylogenies and the comparative method. Am. Nat. 125(1):1–15.

Gack MU, Kirchhofer A, Shin YC, Inn K-S, Liang C, Cui S, Myong S, Ha T, Hopfner K-P, Jung JU. 2008. Roles of RIG-I N-terminal tandem CARD and splice variant

in TRIM25-mediated antiviral signal transduction. Proc. Natl. Acad. Sci. 105(43):16743–16748.

Hayashi T, Watanabe C, Suzuki Y, Tanikawa T, Uchida Y, Saito T. 2014. Chicken MDA5 senses short double-stranded RNA with implications for antiviral response against avian influenza viruses in chicken. J. Innate Immun. 6(1):58–71.

Hayward A, Katzourakis A. 2015. Endogenous retroviruses. Curr. Biol. 25(15):R644–R646.

Holm S. 1979. A Simple Sequentially Rejective Multiple Test Procedure. Scand J Stat. 6:65–70.

Hurst TP, Magiorkinis G. 2015. Activation of the innate immune response by endogenous retroviruses. J. Gen. Virol. 96(pt 6):1207–1218.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346(6215):1320–1331.

Kassiotis G, Stoye JP. 2016. Immune responses to endogenous retroelements: Taking the bad with the good. Nat. Rev. Immunol. 16(4):207–219.

Kato H, Takeuchi O, Mikamo-Satoh E, Hirai R, Kawai T, Matsushita K, Hiiragi A, Dermody TS, Fujita T, Akira S. 2008. Length-dependent recognition of double-stranded ribonucleic acids by retinoic acid–inducible gene-I and melanoma differentiation–associated gene 5. J. Exp. Med. 205(7):1601–1610.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30(14):3059–3066.

Kawai T, Akira S. 2010. The role of pattern-recognition receptors in innate immunity:

Update on toll-like receptors. Nat. Immunol. 11(5):373–384.

Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22(5):1208–1222.

Kosakovsky Pond SL, Frost SDW, Muse S V. 2005. HyPhy: Hypothesis testing using phylogenies. Bioinformatics 21(5):676–679.

Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol. Biol. Evol. 28(11):3033–3043.

Kowalinski E, Lunardi T, McCarthy AA, Louber J, Brunel J, Grigorov B, Gerlier D, Cusack S. 2011. Structural basis for the activation of innate immune pattern-recognition receptor RIG-I by viral RNA. Cell 147(2):423–435.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A resource for timelines, timetrees, and divergence times. Mol. Biol. Evol.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33(7):1870–1874.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860–921.

Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, McWilliam H, Valentin F, Wallace I, Wilm A, Lopez R, et al. 2007. ClustalW and ClustalX version 2. Bioinformatics 23(21):2947–2948.

Li X, Ranjith-Kumar CT, Brooks MT, Dharmaiah S, Herr AB, Kao C, Li P. 2009. The RIG-I-like receptor LGP2 recognizes the termini of double-stranded RNA. J.

Biol. Chem. 284(20):13881–13891.

Liniger M, Summerfield A, Zimmer G, McCullough KC, Ruggli N. 2012. Chicken Cells Sense Influenza A Virus Infection through MDA5 and CARDIF-signaling Involving LGP2. J. Virol. 86(2):705–717.

Loo YM, Gale M. 2011. Immune Signaling by RIG-I-like Receptors. Immunity 34(5):680–692.

Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R. 2012. Env-less endogenous retroviruses are genomic superspreaders. Proc. Natl. Acad. Sci. 109(19):7385–7390.

McCarthy EM, McDonald JF. 2004. Long terminal repeat retrotransposons of Mus musculus. Genome Biol. 5(3):R14.

Medzhitov R, Janeway CA. 1997. Innate immunity: The virtues of a nonclonal system of recognition. Cell 91(3):295–298.

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TLL, Stadler T, et al. 2011. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334(6055):521–524.

Miranzo-Navarro D, Magor KE. 2014. Activation of duck RIG-I by TRIM25 is independent of anchored ubiquitin. PLoS One 9(1):e86968.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8(7):e1002764.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3(5):418–426.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera - A visualization system for exploratory research and analysis. J. Comput. Chem. 25(13):1605–1612.

Saito T, Hirai R, Loo Y-M, Owen D, Johnson CL, Sinha SC, Akira S, Fujita T, Gale M. 2007. Regulation of innate antiviral defenses through a shared repressor domain in RIG-I and LGP2. Proc. Natl. Acad. Sci. U. S. A. 104(2):582–587.

Satoh T, Kato H, Kumagai Y, Yoneyama M, Sato S, Matsushita K, Tsujimura T, Fujita T, Akira S, Takeuchi O. 2010. LGP2 is a positive regulator of RIG-I- and MDA5-mediated antiviral responses. Proc. Natl. Acad. Sci. 107(4):1512–1517.

Saxena SK, Chitti S V. 2016. Molecular biology and pathogenesis of retroviruses. In: Advances in Molecular Retrovirology. London: InTech.

Scheffler K, Martin DP, Seoighe C. 2006. Robust inference of positive selection from recombining coding sequences. Bioinformatics 22(20):2493–2499.

Shao Q, Xu W, Yan L, Liu J, Rui L, Xiao X, Yu X, Lu Y, Li Z. 2014. Function of duck RIG-I in induction of antiviral response against IBDV and avian influenza virus on chicken cells. Virus Res. 191(1):184–191.

Sukumaran J, Holder MT. 2010. DendroPy: A Python library for phylogenetic computing. Bioinformatics 26(12):1569–1571.

Sun Y, Ding N, Ding SS, Yu S, Meng C, Chen H, Qiu X, Zhang S, Yu Y, Zhan Y, et al. 2013. Goose RIG-I functions in innate immunity against Newcastle disease

virus infections. Mol. Immunol. 53(4):321–327.

Uchikawa E, Lethier M, Malet H, Brunel J, Gerlier D, Cusack S. 2016. Structural analysis of dsRNA binding to anti-viral pattern recognition receptors LGP2 and MDA5. Mol. Cell 62(4):586–602.

Valdar WSJ. 2002. Scoring residue conservation. Proteins Struct. Funct. Genet. 48(2):227–241.

Venkataraman T, Valdes M, Elsby R, Kakuta S, Caceres G, Saijo S, Iwakura Y, Barber GN. 2007. Loss of DExD/H box RNA helicase LGP2 manifests disparate antiviral responses. J. Immunol. 178(10):6444–6455.

Wertheim JO, Murrell B, Smith MD, Pond SLK, Scheffler K. 2015. RELAX: Detecting relaxed selection in a phylogenetic framework. Mol. Biol. Evol. 32(3):820–832.

Xu W, Shao Q, Zang Y, Guo Q, Zhang Y, Li Z. 2015. Pigeon RIG-I function in innate immunity against H9N2 IAV and IBDV. Viruses 7(7):4131–4151.

Yoneyama M, Fujita T. 2009. RNA recognition and signal transduction by RIG-I-like receptors. Immunol. Rev. 227(1):54–65.

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346(6215):1311–1320.

Zou J, Chang M, Nie P, Secombes CJ. 2009. Origin and evolution of the RIG-I like RNA helicase gene family. BMC Evol. Biol. 9(1):85.

Supplementary

Table 2.S1 Datasets of RLR CDSs.

Common name	Species name	*RIG-I*				*MDA5*				*LGP2*			
Birds													
Downy Woodpecker	*Picoides pubescens*	1	2	3	†	3	3	3	†	3			†
Carmine Bee eater	*Merops nubicus*	1	2	3	†	3	3	3	†	3			†
Rhinoceros Hornbill	*Buceros rhinoceros*	1	2		†	3	3		†	3			†
Bar tailed Trogon	*Apaloderma vittatum*	1	2		†	3	3		†	3			†
Cuckoo Roller	*Leptosomus discolor*	1			†	3	3		†	3	3		†
Speckled Mousebird	*Colius striatus*	1			†	3	3		†	3	3		†
Barn Owl	*Tyto alba*	1	2		†	3	3		†	3			†
Golden Eagle	*Aquila chrysaetos*	1	2			3	3			3	3		
Turkey Vulture	*Cathartes aura*				†				†				†
White tailed Eagle	*Haliaeetus albicilla*	1	2		†	3	3		†	3			†
Bald Eagle	*Haliaeetus leucocephalus*	1	2	3	†	3	3	3	†	3	3	3	†
Red legged Seriema	*Cariama cristata*	1			†	3	3		†	3			†
Peregrine Falcon	*Falco peregrinus*	1			†	3	3	3	†	3	3	3	†
Saker Falcon	*Falco Cherrug*	1				3	3			3	3		
Budgerigar	*Melopsittacus undulatus*	1	2		†	3	3		†	3	3		†
Kea	*Nestor notabilis*	1			†	3	3		†	3			†
Rifleman	*Acanthisitta chloris*	1	2		†	3	3		†	3	3		†
Golden collared Manakin	*Manacus vitellinus*	1	2	3	†	3	3	3	†	3	3	3	†
Blue crowned Manakin	*Lepidothrix coronata*	1	2			3	3			3	3		
Collared Flycatcher	*Ficedula albicollis*	1	2			3	3			3			
American Crow	*Corvus brachyrhynchos*	1	2		†	3	3		†	3	3		†
Hooded Crow	*Corvus cornix*	1	2			3	3			3	3		
Common Starling	*Sturnus vulgaris*	1	2			3	3			3	3		
Zebra Finch	*Taeniopygia guttata*	1	2	3	†	3	3	3	†	3			†
Great Tit	*Parus major*	1	2	3		3	3	3		3	3	3	
Ground Tit	*Pseudopodoces humilis*	1	2			3	3			3	3		
Atlantic Canary	*Serinus canaria*	1	2	3		3	3	3		3	3	3	
Medium Ground finch	*Geospiza fortis*	1	2		†	3	3		†	3			†
White throated	*Zonotrichia*	1	2			3	3			3	3		

Sparrow	*albicollis*													
White tailed Tropicbird	*Phaethon lepturus*	1	2		†		3	3		†		3		†
Sunbittern	*Eurypyga helias*	1	2		†		3	3		†		3		†
Red throated Loon	*Gavia stellata*	1	2		†		3	3		†		3		†
Emperor Penguin	*Aptenodytes forsteri*	1			†		3	3		†		3	3	†
Adelie Penguin	*Pygoscelis adeliae*	1			†		3	3		†		3	3	†
Northern Fulmar	*Fulmarus glacialis*	1			†		3	3		†		3		†
Great Cormorant	*Phalacrocorax carbo*	1	2		†		3	3		†		3		†
Crested Ibis	*Nipponia nippon*	1	2		†		3			†		3	3	†
Little Egret	*Egretta garzetta*	1	2		†		3	3		†		3	3	†
Dalmatian Pelican	*Pelecanus crispus*	1	2		†		3	3		†		3		†
Hoatzin	*Opisthocomus hoazin*	1			†		3	3		†		3		†
Grey Crowned Crane	*Balearica regulorum*	1	2		†		3			†		3		†
Killdeer	*Charadrius vociferus*	1	2		†		3	3		†		3		†
Ruff	*Calidris pugnax*	1	2				3	3				3	3	
Red crested Turaco	*Tauraco erythrolophus*	1	2		†		3	3		†		3		†
MacQueen's Bustard	*Chlamydotis macqueenii*	1	2		†		3	3		†		3		†
Common Cuckoo	*Cuculus canorus*	1			†		3	3		†		3	3	†
Chuck will's widow	*Caprimulgus carolinensis*	1			†		3	3		†		3		†
Anna's Hummingbird	*Calypte anna*	1	2		†		3	3		†		3	3	†
Chimney Swift	*Chaetura pelagica*				†		3	3		†		3		†
American Flamingo	*Phoenicopterus ruber*				†					†				†
Great Crested Grebe	*Podiceps cristatus*				†					†				†
Pigeon	*Columba livia*	1	2		†		3	3		†		3	3	†
Yellow throated Sandgrouse	*Pterocles gutturalis*	1			†		3	3		†		3		†
Brown Mesite	*Mesitornis unicolor*	1			†		3	3		†		3	3	†
Duck	*Anas platyrhynchos*	1	2		†		3	3		†		3		†
Goose	*Anser cygnoides*	1	2	3			3	3	3			3		
Turkey	*Meleagris gallopavo*				†		3	3	3	†		3	3	3	†
Chicken	*Gallus gallus*				†		3	3	3	†		3	3	3	†
Japanese Quail	*Coturnix japonica*						3	3				3	3		
White throated Tinamou	*Tinamus guttatus*				†		3	3		†		3	3		†
Brown Kiwi	*Apteryx australis*	1					3	3				3	3		
Ostrich	*Struthio camelus*	1	2	3	†		3	3	3	†		3			†
Mammals														
Domestic Cat	*Felis catus*													
Pacific Walrus	*Odobenus rosmarus divergens*			3					3					†
Polar Bear	*Ursus maritimus*			3					3					†
Alpaca	*Vicugna pacos*								3					†

Sperm Whale	*Physeter catodon*		3	
Bottlenose dolphin	*Tursiops truncatus*		3	†
White-faced sapajou	*Cebus capucinus imitator*	3	3	†
Human	*Homo sapiens*	3	3	†
Philippine Tarsier	*Carlito syrichta*	3	3	†
African Savanna Elephant	*Loxodonta africana africana*	3	3	

†: among the 48 species in the literature (Jarvis et al. 2014) with reconstructed phylogeny using whole-genome data.

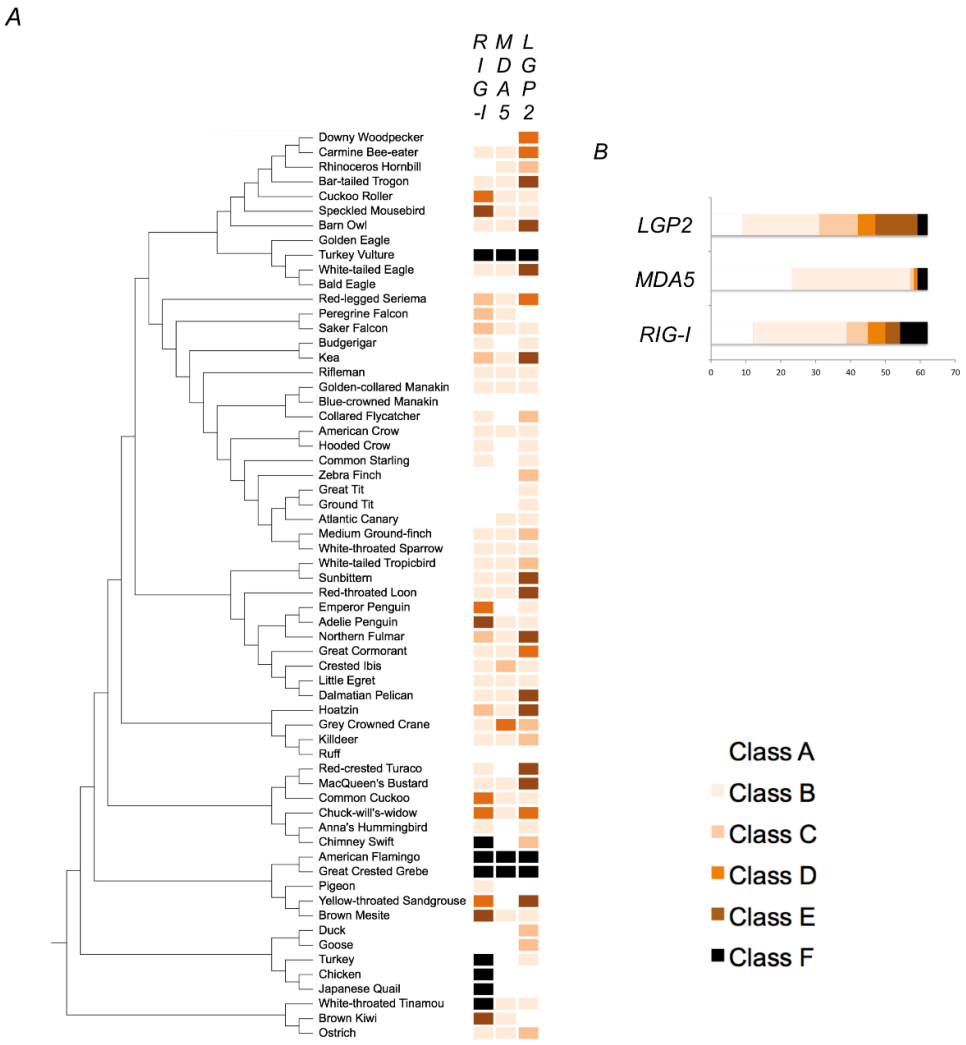


Figure 2.S1. Six-class face quality assessment of the predicted avian RLR CDSs. The six colors of the blocks are indicative of the assigned classes as denoted in lower right corner of the figure. (A) Six-class assessment of the three RLR CDSs for each of the 62 species. The tree shows the phylogenetic relationships of the birds without being scaled on divergence time. (B) Summary of six-class assessment for each avian RLR.

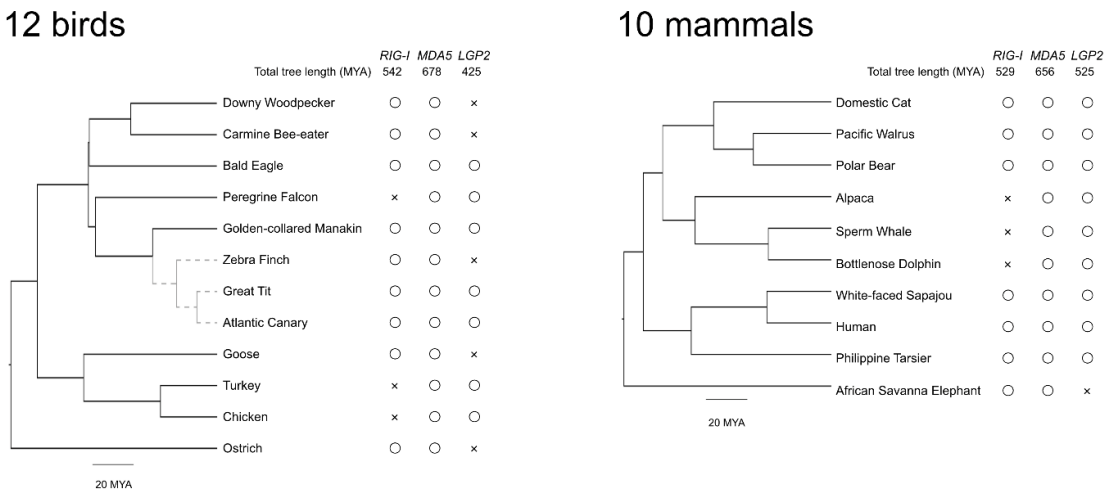


Figure 2.S2. Phylogeny of the species used for comparative analyses of RLRs between birds and mammals (dataset 3). The tree topology shows the phylogenetic relationship. A circle or a cross indicates the presence or absence, respectively, of a reliable CDS record from our dataset for birds or from GenBank for mammals. Branch lengths are in the scale of species divergence time. Branches in dotted line have an unknown divergence time and thus are not scaled. Total branch lengths denoted for each RLR gene are calculated with the species with a reliable CDS record of that gene. Total branch lengths involving branches of unknown divergence time take the average of possible maximum and minimum.

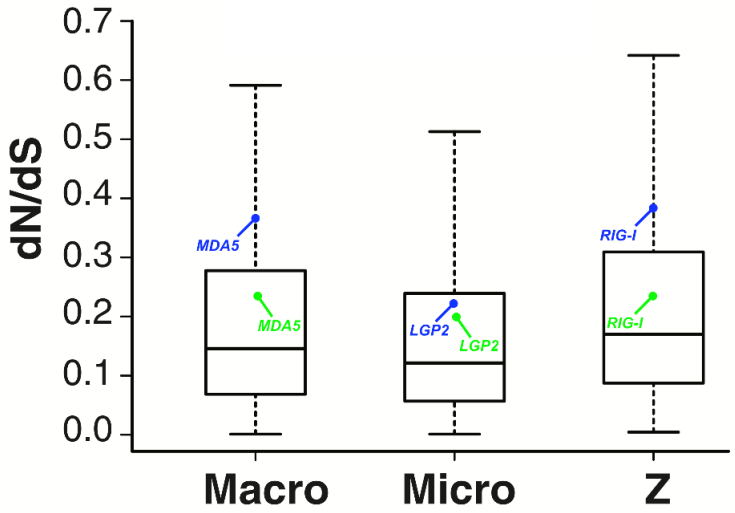


Figure 2.S3. dN/dS of the three avian RLRs compared with average level of coding genes in birds. The figure is adapted from one in the literature (Zhang et al. 2014[2]). Blue represents the result of SLAC using dataset 2 and green represents the result of Nei-Gojobori method using dataset 3.

[2] Zhang et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science 2014, 346, 1311–1320.

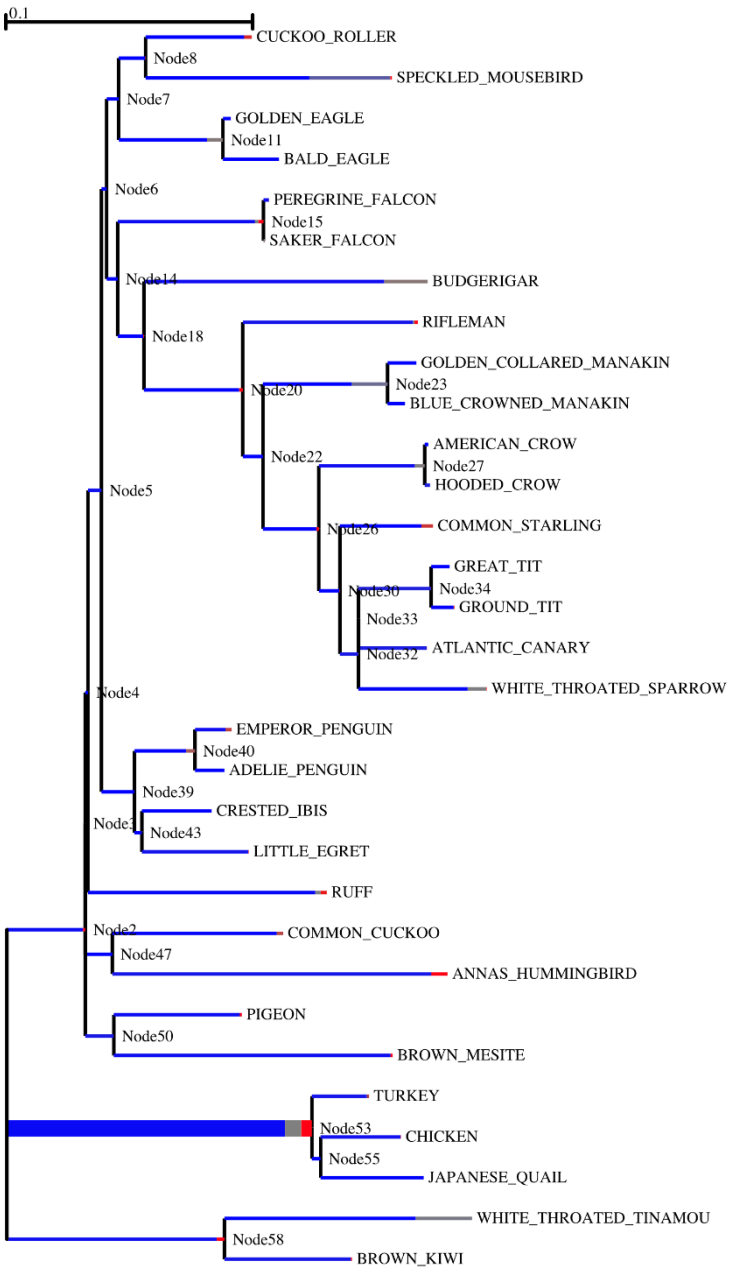


Figure 2.S4. BSR analysis of avian LGP2. The tree topology shows the phylogenetic relationship of the birds. Strength of selection are represented by colors, with red corresponding to dN/dS > 5, blue to dN/dS

= 0 and grey to dN/dS = 1. The width of each color component represents the proportion of sites in the corresponding class. Thicker branches have been classified as undergoing episodic positive selection by the sequential likelihood ratio test at corrected $p \leq 0.05$.

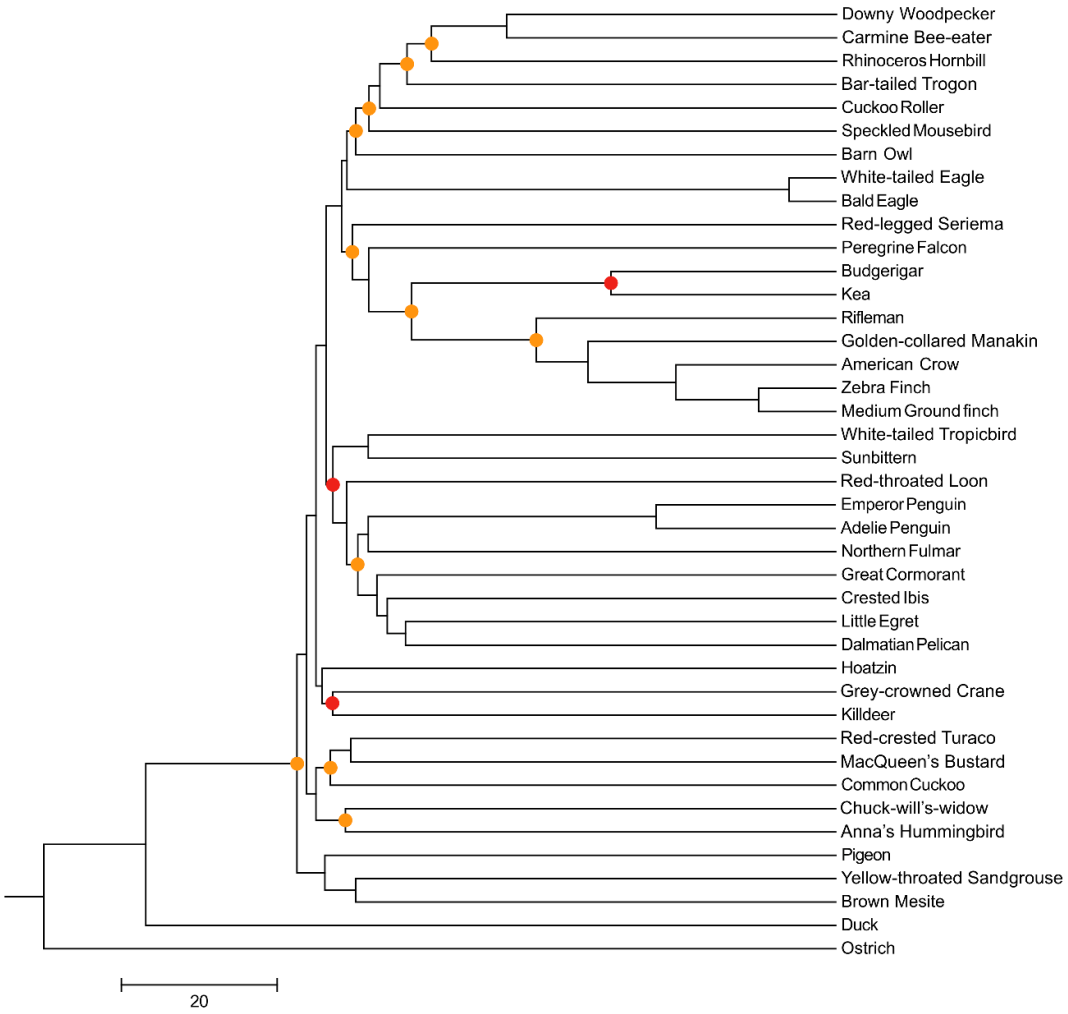


Figure 2.S5. Nodes showing large discordance between their ranks of the PICs of species-to- ancestor dN/dS and ERV load. Dots in red indicate the nodes that have a difference of > 30 between the two ranks; in orange indicate the nodes that have a difference of > 20 but ≤ 30 between the two ranks. Branch lengths are scaled to divergence time (Mya).

Chapter 3

Genome-Wide Evolutionary Interaction with Endogenous Retrovirus Load in Mammals and Birds

Abstract

Mammals and birds differed largely in the average load of endogenous retrovirus (ERV) (a metric of proportion of ERV in the host genome). It has been unknown whether the host-ERV relationship including conflicts and co-option is related to this difference. Through phylogenetic gene-phenotype association tests on about 5000 genes in mammals and birds separately, I detected genes that evolved in association with ERV loads in each group. Genes showing high degree of association between the gene evolutionary rate and the ERV load occur more frequently in birds than in mammals. Compared with positive association, negative association is more pervasive in both mammals and birds. Gene set enrichment analysis based on the association results reveals evolutionary interaction of biological processes with the ERV load. The evolutionary interaction (1) is remarkable for gene silencing in both mammals and birds; (2) is stronger for negative immune regulation than positive immune regulation in mammals, while in birds it is the opposite way; (3) shows higher weight for DNA recombination in mammals than in birds.

3.1. Introduction

The endogenous retrovirus (ERV) is a relic sequence of the retrovirus in the host genome. Retrovirus infection in vertebrates can be traced back to over 450 million years ago with phylogeny of ERVs (Aiewsakun and Katzourakis 2017). ERVs constitute the majority of endogenous viral elements (EVE) in the host genome and form part of a larger assemblage of long terminal repeats (LTR)-retrotransposons (Gifford et al. 2018). After entering a host cell, the retrovirus needs to integrate into the host genome to complete its replication cycle (Hayward and Katzourakis 2015). The retroviral sequences integrated in germline cells are potential to transmit to offspring of the host. If such a sequence reach fixation (or spread to a certain extent) in the host population, it is considered as an ERV (Feschotte and Gilbert 2012; Hayward and Katzourakis 2015; Kapusta and Suh 2017). ERVs can increase their copy numbers in a genome through retrotransposition or reinfection in the germline (Belshaw et al. 2004; Bannert and Kurth 2006; Magiorkinis et al. 2012). As a result, ERVs constitute 8% of the human genome (Lander et al. 2001) and 11% of the mouse genome (McCarthy and McDonald 2004). The proportion of ERVs ranges from 2.39% to 11.41% of the genome in mammals, but is much lower in birds, ranging from 0.16% to 3.57% (Cui et al. 2014). Cui *et al* (Cui et al. 2014) discussed three possible reasons for the difference in the EVE proportion between mammals and birds. The three possible reasons can be applied to ERVs: (1) birds have been exposed to fewer retroviral infections than mammals; (2) birds are more resistant to retrovirus integration following infection; (3) DNA deletion rate is higher in birds (Kapusta et al. 2017) and therefore ERVs are more frequently purged from the genome. The third

hypothesis was supported by the observation of higher density of TEs, including ERVs in the W chromosome of birds, which shows low recombination rate (Kapusta and Suh 2017). However, DNA deletion may not fully explain the difference of ERV proportion between mammals and birds. Higher DNA deletion rate will also lead to smaller genome size, but a species with small genome size can have high ERV proportion. For example, the mouse-eared bat (*Myotis davidii*) indeed has a small size of genome (2.089 Gb), which is similar to many of birds, but has an ERV proportion of 4.751%, which is higher than all the birds investigated and is also higher than several mammals (Cui et al. 2014). To fully explain, I need to consider the unique relationship between host and ERVs/retroviruses. High or low ERV load in the genome may be involved in adaptation and phenotype evolution of the host, specifically speaking, ERV load may affect or be affected by the relationship between host and ERVs/retroviruses. For example, it was reported that RIG-I, the innate immune receptor for intracellular viral RNA, evolved in correlation with ERV load in birds, showing the possibility of association of host immunity evolution with ERV load change (Zheng and Satta 2018).

The relationship between host and ERVs/retroviruses contains two sides: conflict and co-option. Some ERVs can still generate transcripts, proteins and reverse-transcripts, although most ERVs no longer encode intact viruses (Stoye 2001). The host has to retain defense against both exogenous retrovirus and pro-virus/ERV re-insertion, since the insertions will threaten the integrity of the host genome. However, host defense in this conflict can be mitigated. Expression of ERVs may require a certain degree of tolerance of the host immune system to avoid chronic inflammation (Kassiotis and Stoye 2016) and. ERV expression were found pathogenic in many

autoimmune diseases such as systemic lupus erythematosus (SLE) (Ogasawara et al. 2000; Ogasawara et al. 2010), multiple sclerosis (MS) (Perron et al. 1997; Antony et al. 2004; Tselis 2011). On the other hand, some ERVs are co-opted by the host and involve various physiological functions. These ERVs can promote host evolution of gene-regulatory network and phenotype. The most prevalent case in vertebrates is that the proteins encoded by ERV genes *env* or *gag* can restrict later viral entry or integration (Frank and Feschotte 2017). Another notable case is the important role of ERVs in the origin and evolution of placenta in mammals. ERV-derived proteins gained a variety of novel functions in placenta such as syncytins mediate cell fusion to form the barrier from maternal immune cells and exogenous viruses (Sha et al. 2000; Blaise et al. 2003; Dupressoir et al. 2005); and ERVs also constitute a substantial fraction of regulatory elements in placental cells (Chuong et al. 2013). ERVs also provide ligands and regulatory elements in many other biological functions and the cases of co-option were extensively discovered in mammals (Frank and Feschotte 2017). The extensive co-option and higher ERV loads together suggest that mammals might establish a specific relationship with ERVs during evolution. To retain this relationship, mammals might evolve to be more tolerant than birds with ERVs/retroviruses in general.

To better understand the host-ERV relationship during long-term evolution in different vertebrate groups, I conducted a comparative study about the evolutionary interaction between host genes and ERV loads in mammals and birds, with particular attention to biological processes that involves the most potential host restrictions to ERV load, including immune responses, gene silencing and DNA recombination (the main source of DNA deletion).

3.2. Materials and Methods

3.2.1. Selected Mammal and Bird Species

Association analyses for evolutionary rate and traits require reliable endogenous retrovirus (ERV) copy number and genome size data, as well as species representative of large ranges of birds or mammals. Since this study uses published data, the quality of ERV load estimation depends on the reliability of published data of ERV copy number and genome size. To ensure that the estimated ERV load is reliable for association analyses, I selected the species of which the ERV copy numbers were obtained from whole-genome data of good quality. Our threshold for good quality is set as scaffold N50 $\geq$ 0.5Mb (5×10^5 bp) for a mammal, and scaffold N50 $\geq$ 1Mb (10^6 bp) for a bird. Under this condition, 12 placental mammals and 21 birds (see Table 3.1) were selected from whole-genome-sequenced species for association analyses. All data treatments and analyses were performed separately for birds and mammals.

Table 3.1. ERV Load Categories of the 12 Mammals and 21 Birds.

Species Name	Common Name	ERV Load (copies/Gb)	ERV Load Category
Mammals			
Canis familiaris	Domestic dog	214	0
Ailuropoda melanoleuca	Giant panda	244	0
Sus scrofa	Domestic pig	283	0
Equus caballus	Horse	461	1
Bos taurus	Domestic cow	473	1
Callithrix jacchus	Common marmoset	745	1
Pan troglodytes	Chimpanzee	910	2
Macaca mulatta	Rhesus macaque	969	2
Rattus norvegicus	Brown rat	1028	2
Homo sapiens	Human	1238	3
Cavia porcellus	Guinea pig	1921	3
Mus musculus	House mouse	2126	3
Birds			
Struthio camelus	Ostrich	107	0
Cuculus canorus	Common Cuckoo	165	0
Aptenodytes forsteri	Emperor Penguin	184	0
Pygoscelis adeliae	Adelie Penguin	199	0
Columba livia	Pigeon	226	1
Egretta garzetta	Little Egret	238	1
Nipponia nippon	Crested Ibis	244	1
Anas platyrhynchos	Duck	254	1
Manacus vitellinus	Golden-collared Manakin	279	2
Meleagris gallopavo	Turkey	285	2
Falco peregrinus	Peregrine Falcon	286	2
Chaetura pelagica	Chimney Swift	340	2
Opisthocomus hoazin	Hoatzin	352	2
Calypte anna	Anna's Hummingbird	380	3
Charadrius vociferus	Killdeer	381	3
Picoides pubescens	Downy Woodpecker	427	3
Melopsittacus undulatus	Budgerigar	434	3
Gallus gallus	Chicken	517	3
Taeniopygia guttata	Zebra Finch	587	3
Geospiza fortis	Medium Ground-finch	732	3
Corvus brachyrhynchos	American Crow	942	3

3.2.2. Coding Gene Alignments

I retrieved aligned coding sequences of orthologous genes of 48 birds and outgroup species from Jarvis *et al* (Jarvis et al. 2015) and those of 39 placental mammals an outgroup species from Douzery *et al* (Douzery et al. 2014). I first removed all the sequences containing premature terminal codons (PTCs). A gene would be used in this study if its alignment data meets the following criteria: it contains sequences of all the selected species; it contains at least one species of the outgroup of the selected species. The outgroup for birds include the American alligator (*Alligator mississippiensis*), the green sea turtle (*Chelonia mydas*), the green anole lizard (*Anolis carolinensis*) and the human (*Homo sapiens*). The outgroup for placental mammals include the tammar wallaby (*Macropus eugenii*), the Tasmanian devil (*Sarcophilus harrisii)*, The gray short-tailed opossum (*Monodelphis domestica*) and the platypus (*Ornithorhynchus anatinus*). In this way, I got alignments of 5178 genes for mammals and 4890 genes for birds.

3.2.3 Endogenous Retrovirus Load

Endogenous retrovirus load is defined as the copy number divided by genome size in Gb for each species. Copy numbers of ERVs in bird species were retrieved from Cui *et al* (Cui et al. 2014) and those in mammal species from Katzourakis *et al* (Katzourakis et al. 2014). In both of those two literatures the ERVs were screened using tBLASTn and a library of representative viral protein sequences built by each research group. Genome sizes corresponding to the genome versions used in the

above literatures were retrieved from NCBI Genome (NCBI 2018) and Archive Ensembl (Ensembl 2018).

3.2.4. Phylogenetic Gene–Phenotype Association Tests

I applied two methods of phylogenetic association analysis to detect the association between the evolutionary rate of coding genes and ERV load (rate-load association, for simplicity in this paper). The first method is phylogenetic general least squares (PGLS) models (Pagel 1997). This method estimates the correlation of the variations of two continuous traits and the effect of phylogenetic relationships on the variations are controlled. Applied to this study, the root-to-tip dN/dS is the first trait and the ERV load is the second trait. Evidence of association is evaluated with the Bayes factor (BF) or its logarithm form (log BF), which is the ratio of marginal likelihood of correlated model over that of non-correlated model estimated from Markov chain Monte Carlo (MCMC). Positive or negative association with a detected gene can be reflected in the positive or negative sign of the average correlation coefficient R of the correlated model. This method is implemented in BayesTraits v3.0.1 (Meade and Pagel 2017). Root-to-tip dN/dS were estimated under a branch model with PAML v4.9 (Yang 2007).

The second method is phylogenetic substitution models with mixed rate matrices between genotype and phenotype (O'Connor and Mundy 2009). Evidence of association is evaluated with a likelihood ratio test (LRT) of the model that nucleotide substitution rate is weighted by shifts among up-to-four semi-quantitative phenotype categories over the model without weighted. Applied to this study, the phenotype

categories are categories of the ERV load. I made an ordered list of the ERV loads of the mammal or bird species and divided the list in equal of number of species into four sections. The ordered list for mammals consisted of the 12 selected species. The ordered list for birds consisted of 48 species including the 21 selected species. Since 21 is not the multiples of four, I used 48 species of which information for calculating ERV load are available. The ERV load range in real numbers from the highest to the lowest value in each of the four sections defines an ERV load category. Analyses for positive and negative association were performed, separately. I excluded genes that showed critical LRT results of p-value< 0.05 in both positive and negative association analyses from the set of genes of high degrees of association. In addition, a corrected p-value for multiple tests, the q-value, was calculated following Storey *et al* (Storey and Tibshirani 2003) for all the genes. The second method is implemented in GetGPA (O'Connor and Mundy 2009).

3.2.5. Gene Ontology Annotation and Gene Set Enrichment Analysis

Gene ontology (GO) annotation and gene set enrichment analysis was performed in Blast2Go v5 (Götz et al. 2008). Gene set enrichment analyses (GSEA) (Subramanian et al. 2005) were performed on pre-ranked lists of rescaled log BF. The classic method of gene set enrichment analysis takes the rank alone into account while the weighted method additionally takes the expression differentiation values into account. In this study, expression differentiation is represented with a rescaled value of log BF. The main purpose of rescaling was to reduce the magnitude of differences in log BF values that may bias the weighted GSEA caused by the impact of

occasional extreme values. Even though, classic GSEA is recommended for pre-ranked data (Subramanian et al. 2005) thus the result should be viewed with caution. The rescaled log BF is

$$Sgn(\log BF)\ \frac{|\log BF| - B_n}{B_{n+1} - B_n}\ ,$$

where B_n is the nth of five boundaries of four sections for log BF absolute values and the log BF belongs to the section between B_n and B_{n+1}. B_1 to B_5 take 0, 2, 5, 10 and 100.

3.2.6. Evaluation of Convergent or Divergent Rate-Load Association

Convergent rate-load association with ERV load of a gene is defined as synchronous strong evidence of rate-load association in both mammals and birds. Divergent rate-load association is defined as showing strong evidences of rate-load association in one of the two groups but rate-load independence in the other group. If a gene shows rescaled log BF >2 (original log BF >5) in both mammals and birds, it will be identified of convergent rate-load association. If a gene shows rescaled log BF >2 in one group but rescaled log BF <-2 in the other group, it will be identified of divergent rate-load association.

3.3. Results

3.3.1. Genes Showing Rate-Load Association with ERV Load Are Detected

5178 genes of 12 placental mammals and 4890 genes of 21 birds are used for studying the association between the genome-wide coding gene evolution and the endogenous retrovirus (ERV) load (rate-load association, for simplicity in this paper) in each of the two vertebrate groups. The species are shown in Table 3.1. I used two different methods, a phylogenetic general least square (PGLS) (Pagel 1997; Pagel 1999) work frame and a mixed matrices method (O'Connor and Mundy 2009), to detect the gene-load association. Since the two methods were implemented in BayesTraits and GetGPA, respectively, I refer to them as B method and G method in this text. Degree of association is indicated with a Bayes factor for method B. A logarithm of Bayes factor (log BF) greater than two indicates evidence, greater than five indicates strong evidence and greater than 10 indicates very strong evidence (Raffety 1996). For method G, the degree of association is indicated by p-values of likelihood ratio tests (LRT).

For mammals, analyses of 5173 genes were successful with B method. Among them, 1262 genes show log BF>2; 153 genes show log BF>5; 9 genes show log BF>10, of which 7 show positive association and 2 show negative association. With G method, analyses of 5095 genes were successful and 266 of them show p-value<0.05. Among those, 61 show q-value<0.05, of which 17 show positive association and 44 show negative association. For birds, I got successful results of

4869 genes with B method. Among them, 1735 genes show log BF>2; 581 genes show log BF>5; 133 genes show log BF>10, of which 98 show positive association and 35 show negative association. With G method, analyses of 4594 genes were successful and 1791 of them show p-value<0.05. Among those, 1275 show q-value<0.05, of which 353 show positive association and 922 show negative association.

B method and G method detected quite different set of genes but showed consistency to some extent. I classified the degree of association of each gene from the results of the two methods and investigated the co-occurring of all combinations of the degree classes (Figure 3.1). The lower left and upper right zones represent consistent discoveries between the two methods while the upper left and lower right blocks represent conflict discoveries between the two methods, regarding the sign of association. For mammals, 5090 genes have results from both methods, with 31 of them located in the consistent discovery zones (Figure 3.1), from in total 1240 genes that show log BF>2 in B method and 266 genes that show p<0.05 in G method. For birds, 4576 genes have results from both methods, with 318 of them located in the consistent discovery zones (Figure 3.1), from in total 1620 genes that show log BF>2 in B method and 1791 genes that show p<0.05 in G method. Genes that located in the blocks with the highest level of degree of association in both methods are found only in bird and the number is 10 (Table 3.2).

Figure 3.1. Comparison of the results between B method and G method by co-occurring of discovery classes of rate-load association. X and Y-axes are classes of degree of association named with method initial B or G, the symbol ‘neg’ or ‘pos’ indicating negative or positive association and the degree rates 1 to 3. The degree rates 1 to 3 for B method indicate log BFs in-between 2 and 5, in-between 5 and 10 and greater than 10, respectively, while for G method indicate p-values of the likelihood ratio test in-between 0.05 and 0.01, in-between 0.01 and 0.005 and less than 0.005. A black cross separates the co-occurring plane into four zones. The lower left and upper right zones represent consistent discoveries between the two methods while the upper left and lower right blocks represent conflict discoveries between the two methods, regarding the sign of association. The block colors and values present the count of genes.

Table 3.2. Bird Genes of the Highest Degree of Association in Both Methods

Gene	Protein Encoded	B method		G method		
		Log BF	R	p-value	q-value	Association Type
CLSTN2	Calsyntenin-2	11.862580	-0.727648	3.59^{-05}	3.2^{-04}	Negative
NFATC2	Nuclear factor of activated T-cells, cytoplasmic 2	11.098068	-0.664603	3.10^{-13}	0	Negative
SUGP2	SURP and G-patch domain-containing protein 2	10.570130	-0.640964	1.22^{-17}	0	Negative
RAPH1	Ras-associated and pleckstrin homology domains-containing	16.325324	0.768969	8.58^{-04}	9.72^{-03}	Positive
TRHR	Thyrotropin-releasing hormone receptor	14.165068	0.737685	4.62^{-04}	5.58^{-03}	Positive
CAB39L	Calcium-binding protein 39-like	13.791844	0.693361	3.23^{-15}	0	Positive
FAM8A1	Protein FAM8A1	12.585644	0.656408	4.33^{-169}	0	Positive
ST6GAL2	Beta-galactoside alpha-2,6-sialyltransferase 2	11.217192	0.673644	0	0	Positive
YWHAG	14-3-3 protein gamma	10.936956	0.454083	2.98^{-150}	0	Positive
TMNT	Thymocyte nuclear protein 1	10.737454	0.669516	5.21^{-200}	0	Positive

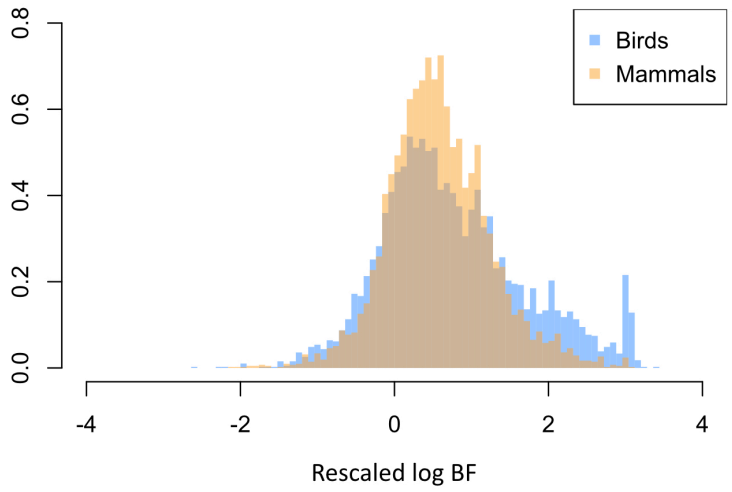
From the above I see that much more genes that evolve in association with ERV load are detected in birds than in mammals with both methods as well as in the consistent discovery zones. Considering the inconsistency between the results of B method and G method, I choose to focus on the result of B method in this study because B method is more convincing since it models two continuous traits as our data exactly are. In contrast to B method, G method only models nucleotide sequences and categorized phenotypes. I had to convert the continuous ERV load into four categories for applying G method and this may cause loss of information and subjective categorizing may introduce bias to the association analysis. Another advantage of B method is that the protein functional evolution metric dN/dS (non-synonymous substitution rate over synonymous substitution rate) could be used as a trait. G method doesn't discriminate between synonymous and non-synonymous substitutions, which may introduce disturbance to the association analysis. The top 10 genes for mammals and birds from method B are shown in Table 3.3.

The distribution of rescaled log BF of mammals and birds show close locations of peak but the distribution of log BF for mammals is nearly bell-shaped while that for birds has a bolder and longer tail on the right side (Figure 3.2A). I also found that positive association is more common than negative association for both mammals and birds (Figure 3.2B).

Table 3.3. Top 10 Rate-Load Associated Genes for Mammals and Birds

Gene	Protein Encoded	log BF	R
Mammals			
INTU	Protein inturned	15.062398	-0.900868
CYB5R3	NADH-cytochrome b5 reductase 3	12.516580	0.765672
TFAP2C	Transcription factor AP-2 gamma	12.469400	0.849238
SYT13	Synaptotagmin-13	12.250774	-0.196901
SUPT16H	FACT complex subunit SPT16	11.932724	0.807048
PICK1	PRKCA-binding protein	11.325090	0.838844
TRAPPC1	Trafficking protein particle complex subunit 1	11.239916	0.880502
MBTPS1	Membrane-bound transcription factor site-1 protease	10.750780	0.788573
KXD1	KxDL motif-containing protein 1	10.655768	0.860930
TFF2	Trefoil factor 2	9.911860	-0.464864
Birds			
LRRC23	Leucine-rich repeat-containing protein 23	43.766168	0.947823
RCOR3	REST corepressor 3	32.453716	0.901074
ENY2	Transcription and mRNA export factor ENY2	25.877020	0.854776
C9ORF152	(Uncharacterized protein)	25.667366	0.867732
TPD52L2	Tumor protein D54	25.595244	0.844478
UBE2J2	Ubiquitin-conjugating enzyme E2 J2	24.215536	0.835109
SLC46A3	Solute carrier family 46 member 3	21.929244	0.834818
COPB1	Coatomer subunit beta	21.813896	0.781284
C1H11ORF70	(Uncharacterized protein)	21.027612	0.803636
PDCD5	Programmed cell death protein 5	20.551992	0.806962

A



B

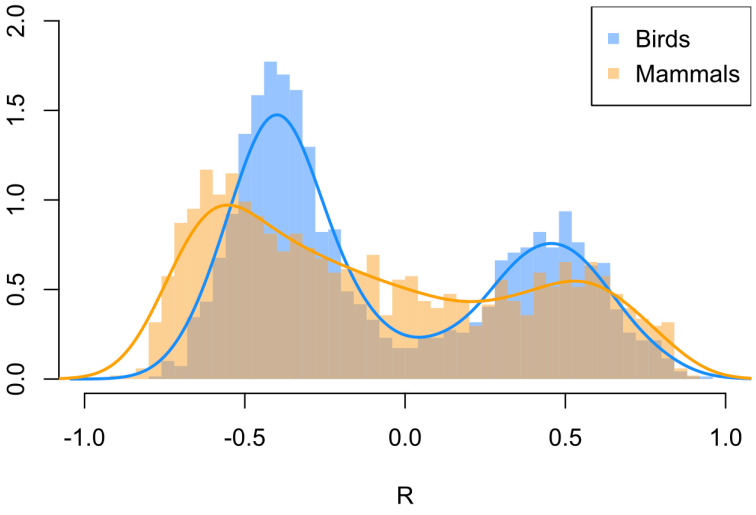


Figure 3.2. Overview of rate-load association of genes in mammals and birds. (A) Distribution of rescaled log BF in mammals and birds. (B) Distribution of the rate-load correlation coefficient of genes in mammals and birds. Genes with log BF > 5 are used.

3.3.2. Genes of Convergent or Divergent Rate-Load Association with ERV Load between Mammals and Birds

Some genes may evolve in association with ERV load in both mammals and birds while others in only one of the two groups. I call the former case convergent rate-load association and the latter divergent rate-load association. I looked for signals of convergent or divergent rate-load association with ERV load between mammals and birds in the 2114 homologous genes of mammals and birds determined by BLAST searches of bird sequences against mammal sequences. Convergent or divergent rate-load association of a gene can be identified with congruous or incongruous high levels of the log BF between mammals and birds (see Materials and Methods and Figure 3.3). Divergent rate-load association (the red-shaded blocks at the upper left and lower right of the plane in Figure 3.3) is found in only one gene (frequency: 0.04%), *GDPD3*, which encodes a lysophospholipase that crosses endoplasmic reticulum membrane (Ohshima et al. 2015). *GDPD3* shows strong evidence of rate-load association with ERV load in birds but no evidence in mammals. On the other hand, seven genes (frequency: 0.33%) show convergent rate-load association (the blue-shaded block at the upper right of the plane in Figure 3.3) with ERV load and distribute in various biological processes or molecular functions such as transcription regulation, cell division, Ca^{2+}-binding (sensing), cell junction and protein transport (The Uniprot Consortium 2019). The rescaled log BF values of the genes identified with convergent or divergent rate-load association are shown in Table 3.4.

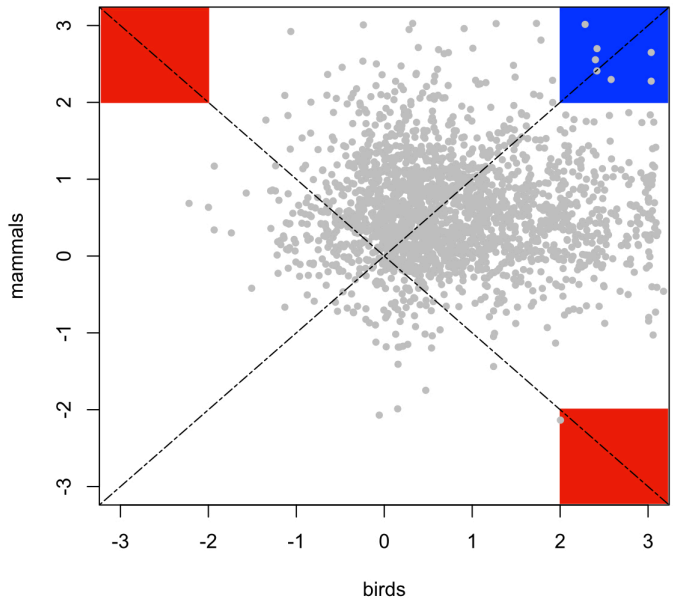


Figure 3.3. Mammals VS. Birds in the rescaled log BF values of 2114 genes. The blue-shaded block indicates the region of convergent rate-load association and seven genes are located in this region. The red-shaded blocks indicate the regions of divergent rate-load association and one gene is located in these regions.

Table 3.4. Genes of Convergent or Divergent Rate-Load Association

Gene	Protein Encoded	Mammal Rescaled log BF	Bird Rescaled log BF
Convergent			
ETV6	Transcription factor ETV6	2.649313	3.035952
TNKS	Poly [ADP-ribose] polymerase tankyrase-1	2.275826	3.035835
PICK1	PRKCA-binding protein	3.014723	2.284043
C2orf76	UPF0538 protein C2orf76 (Uncharacterized)	2.699245	2.419549
C5orf51	UPF0600 protein C5orf51 (Uncharacterized)	2.554476	2.401253
CGN	Cingulin	2.297543	2.579953
RP2	Protein XRP2	2.409116	2.418267
Divergent			
GDPD3	Lysophospholipase D GDPD3	-2.135310	2.003435

3.3.3. GO Terms Enriched in the Rate-Load Association with ERV Load

To understand what biological processes are involved in the rate-load association with ERV load in mammals and birds, I performed gene set enrichment analyses (GSEA) based on the results of association analyses and I ranked the gene sets of gene ontology (GO) terms, by the normalized enrichment score (NES). Based on the B method result of 5173 genes of mammals and 4869 genes of birds, respectively, the top ranked biological process (BP) GO term in the classic GSEAs (see Materials and Methods) for mammals is negative regulation of lymphocyte activation, with representative genes *TNFSF18* (log BF = 6.212182) and *INHBA* (log BF = 5.370952), and for birds is cellular response to lipopolysaccharide, with representative genes *CD36* (log BF = 10.963968), *FN1* (log BF = 9.647306) and *MAP2K3* (log BF = 9.199852) *etc.*. Both of the top ranked BP GO terms are related to immune response.

Critical enrichment of GO terms under the criterion q-value <0.25 were not reported from the classic GSEAs based on results of B method, in mammals or birds, but were reported from the weighted GSEA using rescaled log BF values. In birds, it is the term glutamate secretion, while in mammals it is the term multi-organism transport. The genes such as *NUP133*, *NUP153*, *THOC6* and *NUPL2* in mammals, which contribute the most to the enrichment of GO multi-organism transport, are involved in the export of mRNAs from the nucleus into the cytoplasm (The Uniprot Consortium 2019) and *NUPL2* may also be involved in the docking of viral protein R (Vpr) of human immunodeficiency virus (HIV) at the nuclear envelop (Le Rouzic et

al. 2002). On the other hand, relation between retrovirus and glutamate secretion has not been studied in birds but it was reported that excess glutamate secretion of infected monocyte-derived macrophages is related to HIV associated dementia (Erdmann et al. 2007). The result of weighted GEAS should be viewed with much caution (see Materials and Methods).

3.3.4. Gene Set Enrichment on Rate-Load Association with ERV Load Differs Between Mammals and Birds in Immunity and Gene/Chromatin Silencing

As is mentioned in the introduction, immune response, regulation of gene expression and recombination are the three biological processes most possible to be the casual factors of ERV variation. If any of these were true, the genes involved in the three processes should show different patterns of rate-load association between mammals and birds since they show large difference in ERV load distribution. As is shown in the last result section that GO terms related to immunity rank top in the result of classic GSEA for both mammals and birds, I looked further into the rankings of GO terms relevant to the above three biological processes and compared between mammals and birds. I ranked the GO terms in bird and mammal datasets, respectively, by the NES of classic GSEAs and normalized the rankings with feature scaling so that the rankings are presented with the range in (0,1]. Then I selected GO terms from those containing keywords “immune response”, “silencing”, or “recombination” in the 3261 shared GO terms between the bird and mammal datasets. The normalized rankings (NRs) of the selected GO terms show characteristics as follows. (1) GO gene

silencing ranks very high at the 5th percentile in both mammals and birds and its subcategory GO terms also rank high with nine of the 10 within the 20th percentile in both mammals and birds; (2) GO adaptive immune response ranks high at or close to the 10th percentile in mammals and birds; (3) GO innate immune response ranks medium in both mammals and birds (NR in mammals: 0.4039; NR in birds: 0.2929) but one of its subcategory GO term positive regulation of innate immune response show higher ranking in birds (NR in mammals: 0.7887; NR in birds: 0.0534); (4) when further looking into regulation of immune response, GO positive regulation of immune response rank higher in birds (NR in mammals: 0.4400; NR in birds: 0.0803) and seven out of the 10 subcategory GO terms including the above-mentioned GO positive regulation of innate immune response also rank higher in birds; (5) on the contrary, GO negative regulation of immune response rank higher in mammals (NR in mammals: 0.2395; NR in birds: 0.8267) and all the three subcategory GO terms also rank higher in mammals; (6) GO DNA recombination ranks higher in mammals (NR in mammals: 0.1898; NR in birds: 0.5633) and two out of the three subcategory GO terms also rank higher in mammals (see Figure 3.4). The characteristics (4), (5) and (6) demonstrate the differences between mammals and birds in the biological processes of immune response regulation and recombination.

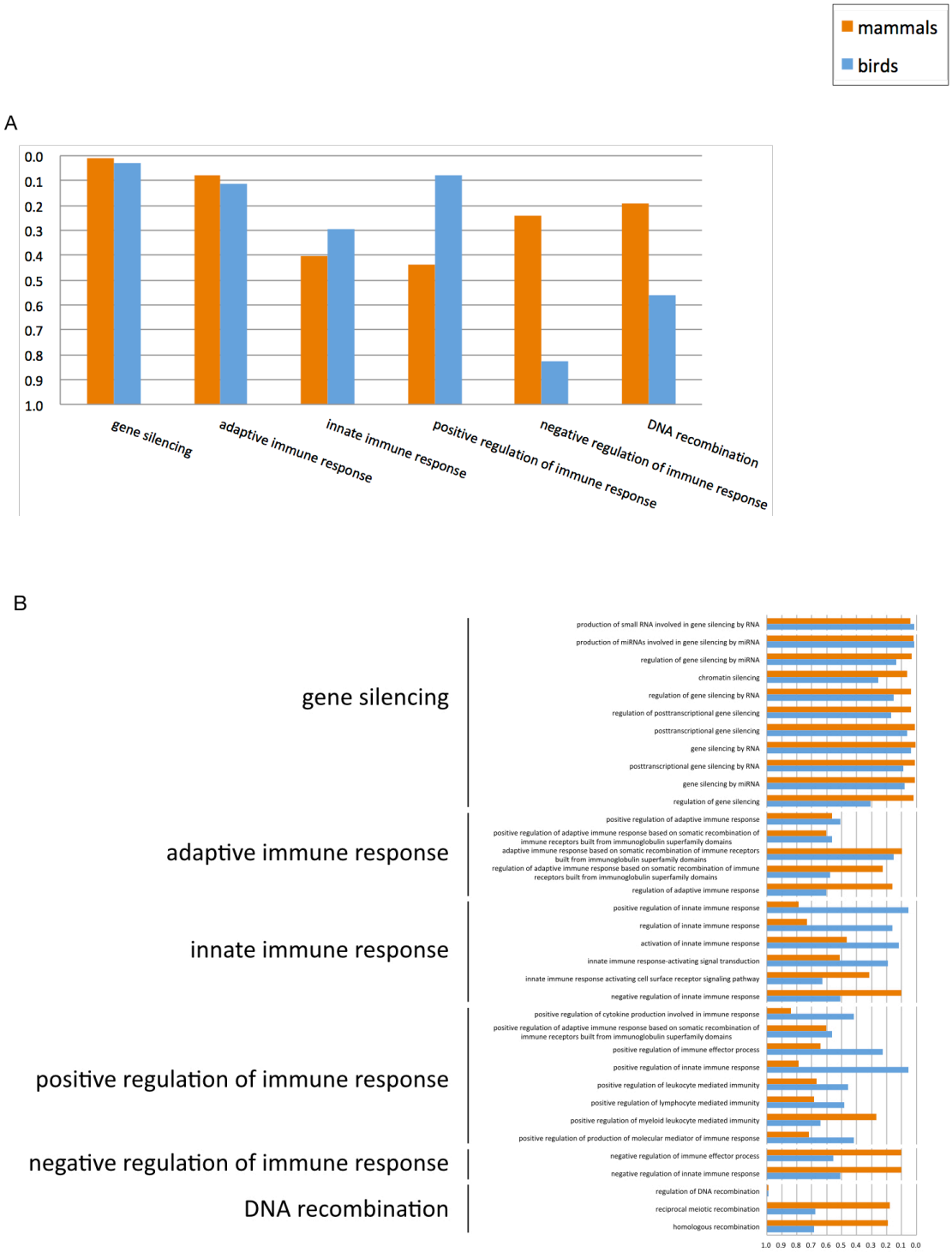


Figure 3.4. Rankings of normalized enrichment score (NES) of potential ERV-load-related GO terms. The charts are based on the 3261 biological process (BP) GO terms shared between the mammal and bird datasets

from the gene set enrichment analysis (GSEA) using B method result. (A) Columns indicate normalized rankings of the GO terms. (B) Bars indicate the normalized rankings of the GO terms. These GO terms shown are subcategories to the five GO terms as sown on the left corresponding to those in (A).

Some genes involved in immune response, gene silencing and DNA recombination show strong evidence of rate-load association, especially in birds. In birds, five genes involved in the above processes show very strong evidence (log BF > 10) of association. *CD36* (log BF = 10.963968) and AP1G1 (log BF = 15.178164) participate in immune responses with the former involved in natural killer cell activation and the latter involved in cellular response to lipopolysaccharide; *FKBP6* (log BF = 15.832288) participates in gene silencing as a co-chaperone required for repressing transposable elements in spermatogenesis (Xiol et al. 2012); *RUVBL1* (log BF = 14.763746) and *THOC1* (log BF = 11.315910) are involved in DNA recombination. In mammals, there are no genes involved in the above processes show very strong evidence (log BF > 10) of association. The gene of the largest log BF is *SETDB2* (log BF = 9.801188) involved in gene silencing, which is a histone methyltransferase participating in chromosome condensation (Falandry et al. 2010).

I also looked into other GO terms about gene expression in a broader sense than gene silencing (Figure 3.s1). For both mammals and birds, all those GO terms rank at top 30% and the GO "negative regulation of gene expression, epigenetic" ranks at top 10%. Genes of very strong evidence (log BF > 10) of association annotated with this GO term are also annotated with the GO gene silencing, except one gene in mammals, *TFAP2C* (log BF = 12.469400), which is a transcription factor binding to enhancer elements (Kang et al. 2014). In addition, *SEH1L* (log BF = 12.94417) in birds is annotated with both the GO terms "gene silencing" and "negative regulation of gene expression, epigenetic" because it encodes a component of the nuclear pore complex (NPC) (Loiodice 2004) and the NPC can import the

miRNA (Reactome 2019). This involvement is quite indirect compared to the other genes mentioned.

3.4. Discussion

Phylogenetic association analyses reveal evolutionary association between genes and ERV load, but it doesn't tell whether gene evolution is the cause or consequence of ERV load changes or both ways. Therefore, the detected association alone indicates possibility of certain kinds of interaction, including cause (driving force), consequence (response) or both (co-evolution). ERV load change can be driven by precedent gene evolution, for example, some microbial environment such as infectious disease breakout may drive evolution of immune system and as a consequence, the evolved immune system show changed behaviors toward ERVs and retroviral infections thus change the integration rate of retroviruses. The microbial environment includes retroviruses and other microbial. ERV load change can also be the driving force of gene evolution, if ERVs express and cause disorders (Kassiotis and Stoye 2016), or if ERVs are welcomed for utilization by the host (Frank and Feschotte 2017). Even though, the genome-wide maps of host rate-load association have shown a big difference between mammals and birds and must be resulted from the difference in their evolutionary interactions with ERVs and/or exogenous retroviruses. Despite the explanatory limitation of association analysis, the distribution of rate-load association and ranking of biological processes in the enrichment of rate-load association, as our result presents, can reveal the commons and differences between mammals and birds in the host-ERV interactions during

evolution and provide information for hypothesizing host-ERV interactions. I discuss them as follows.

A big difference in the distribution of rate-load association between mammals and birds is found. Genes of high degree of evolutionary association with ERV load occur with higher frequency in birds than in mammals and this indicates a higher level of genome-wide evolutionary interactions between the host and ERVs in birds, in both breadth and intensity. This implies that a wide range of physiological processes in birds may take parts in the host-ERV/retrovirus interaction. This difference in the distribution of rate-load association may be the key to an explanation of lower ERV load in birds than in mammals. If the hypothesis proposed in the introduction is true that mammals are tolerant while birds are alert to retrovirus/ERVs in their evolutionary histories, the higher level of rate-load association in birds will represent a stronger restriction to ERV expansion. Meanwhile, negative association is more pervasive in both mammals and birds. Negative association suggests that the faster a gene evolves, the lower ERV load is in their genomes. In other words, relaxed selection or diversifying selection on a gene is related to ERV contraction, while purifying selection on a gene is related to ERV expansion.

With GSEA, commons and differences were shown between mammals and birds in the enrichment of biological processes in rate-load association and the enrichment ranking presents the weight of a biological process in the host-ERV evolutionary interaction. Different from usual GSEA with expression differentiation data, no biological processes show critical enrichment over the rest (q-value <0.25) since evolution of either placental mammals or birds has lasted for over 100 million years and many events involving various host biological processes might turn over

previous host-ERV balance. Therefore host interaction with ERV load should not be expected to be concentrated in limited biological processes. Nevertheless, I can learn the status of a biological process in the complicated host-ERV interaction during evolution from the NES ranking from the GSEA. GSEA is informative by integrating association degrees of genes. The frequencies of convergent and divergent rate-load associated genes between mammals and birds are 0.33% and 0.04%, which means both of them are rare and little informative. However, different genes may work in the same biological process, thus a biological process can be the target of natural selection.

The GSEA result suggests a role of mechanisms that suppress ERV replication in the host-ERV evolutionary interaction. The levels of enrichment in rate-load association of GO gene silencing and some GO terms about immune response are higher than that of DNA recombination in both mammals and birds and show differences between the two groups. Genes involved in the cellular response to lipopolysaccharide in birds show evidence of rate-load association with ERV load. Since lipopolysaccharide is a major component of the cell wall of gram-negative bacteria, those genes might evolve in response to bacteria environment and might coincidently have an influence on the immunity against retroviruses. As is also shown in this study, mammals have higher NES rankings of negative, while birds have higher NES rankings of positive, immune regulation in the rate-load association with ERV load. With the knowledge that positive regulation often occurs in early processes of infection and negative regulation occurs in late processes (Viganò et al. 2012), NES rankings suggest that the evolution of early processes of immune response is more associated with ERV load in mammals than late processes, and it is the opposite

way in birds. Since the early (other than the late) processes of immune response have vital effects on blocking pathogen proliferating processes such as cellular entry and integration of retroviruses, evolution of immune system can either be a precedent driving force of the ERV load changes or an acute responder to retrovirus infection in bird evolution. In contrast to this, late processes of immune responses are responsible for suppressing excessive reactions to avoid tissue damage and chronic inflammation, such as those that may be aroused by the expression of ERVs, which means that the immune system in mammals may evolve in response to ERV load change in the sense of suppressing excess inflammation that the expression of ERVs may lead to (Sharpe et al. 2007; Hamerman et al. 2016; Afonina et al. 2017). Based on our result and the fact that birds have much lower ERV load than mammals, I propose a hypothesis about how the defensive mechanisms control potential damage from ERVs to the host in mammals and birds: birds rely on early immune responses against retrovirus infections to prevent ERV load gaining while mammals use negative regulation to suppress excess inflammation caused by ERVs; however, gene silencing plays the most important role in both birds and mammals in restricting ERV load. A schematic diagram about this hypothesis is shown in Figure 3.5. This hypothesis can explain the difference in the average ERV load between mammals and birds. This hypothesis is also compatible with the another hypothesis proposed in the introduction that mammals could be more tolerant while birds could be more alert to retrovirus/ERVs in their evolutionary histories.

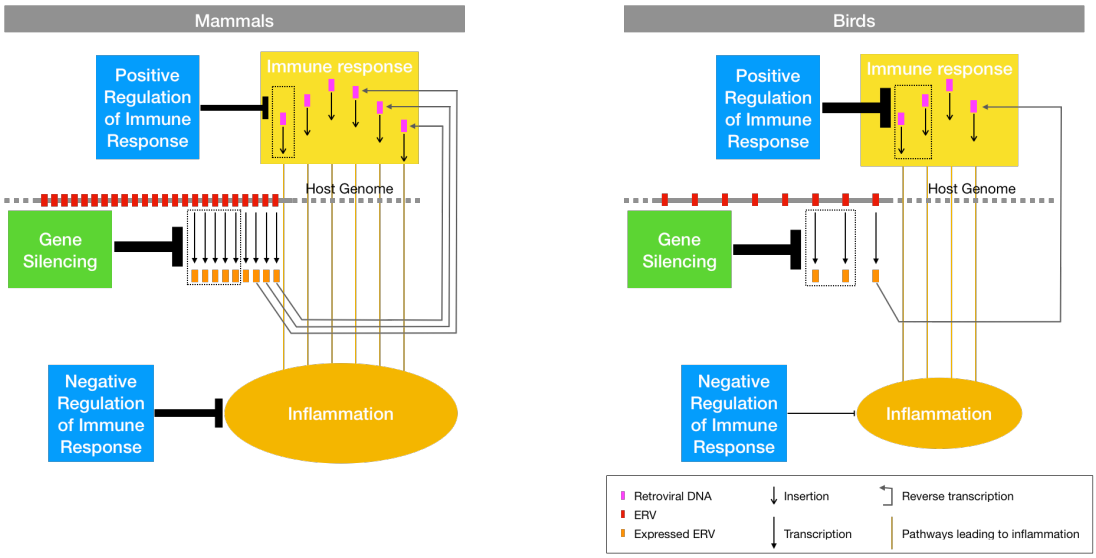


Figure 3.5. A hypothesis of how immunity and gene silencing interact with ERV/retrovirus in mammals and birds differently. Thickness of bar-headed lines indicates intensity of inhibition.

On the other hand, the GO DNA recombination in mammals shows a quite high ranking of enrichment in rate-load association (see Figure 3.4). This supports the possibility that recombination dependent deletion can influence the ERV load through an effect on deletion rate, as proposed in previous studies (Cui et al. 2014; Kapusta et al. 2017). In contrast to mammals, the GO DNA recombination shows a medium ranking of enrichment in rate-load association in birds (see Figure 3.4). This suggests two possible scenarios: (1) deletion rate (presumed to be determined by DNA recombination mechanism) is not a major factor of lower ERV loads during bird evolution; (2) the recombination process itself does not contribute much to the deletion rate in birds. The first scenario is possible in two ways. The first way is that some other strong forces drove the evolution of recombination process in birds and drowned out its evolutionary association with ERV loads; and the second way is that many other rival role players in the bird evolution can dilute the weight of deletion rate in the host-ERV evolutionary interaction. In the second way, deletion rate may still play a role, since two genes in birds involved in DNA recombination show very high degree of association (log BF > 10) with the ERV load while not one gene in mammals does so. On the other hand, regional deletion rate can change along with the chromatin organization (Makova and Hardison 2015), being an example of deletion rate changes without involvement in the mechanism of recombination process itself and therefore the second scenario is also possible. Some of the genes involved in regulation of gene expression (including gene silencing) can influence regional deletion rate because they conduct epigenetic modifications that are related to chromatin organization. It is still unknown if any other biological process may also influence deletion rate. I also suspect a possibility that the deletion rate may have a

smaller effect on the load of ERVs than on the load of other TEs resulting from the unique relationship between ERVs and some hosts during vertebrate evolution.

Besides the analysis about biological processes, I noticed that some genes (especially in birds) of high log BF values, namely of strong evidence of rate-load association, are implied to have high potential of interactions with ERVs by current knowledge gained mostly from studies in mouse or human. A gene involved in gene silencing, *FKBP6* (log BF = 15.832288) in birds is particularly notable for its possible role in repressing transposable elements because protein FKBP6 is found necessary for biogenesis of Miwi2-bound Piwi-interacting RNAs (piRNA) in the mouse germ cell (Xiol et al. 2012). Piwi proteins together with piRNAs mediate silencing of target sequence via DNA methylation during spermatogenesis (Kalmykova et al. 2005; Itou et al. 2015; Manakov et al. 2015) and piRNAs preferentially derive from LTR-retrotransposons (Kalmykova et al. 2005; Houwing et al. 2007), which includes ERVs. In the genes ranked top 10 by log BF values in birds, *RCOR3* (log BF = 21.929244) (Gaudet et al. 2011) and *ENY2* (log BF = 21.813896) (Lang et al. 2011) are involved in transcription process, implying that they may have effects on ERV expression; the two uncharacterized proteins, *C9ORF152* (log BF = 25.667366) of chickens shows the highest expression in the female gonad and *C1H11ORF70* (log BF = 21.027612) of chickens shows the highest expression in the testis (Bastian et al. 2008), implying that they may be involved in the germline repression of ERVs in birds. *LRRC23* (log BF = 43.766168) is the top ranked gene in birds and belongs to the leucine-rich repeats (LRR)-containing domain family of pattern recognition receptors initiating innate immune responses (Ng and Xavier 2011), implying that it may interact with ERV loads in immune responses against expressed ERVs or retroviruses. In the genes

ranked top 10 by log BF values in mammals, *TFAP2C* (log BF = 12.469400) encodes sequence-specific DNA-binding protein that interacts with enhancer elements to regulate transcription (Bamforth et al. 2001; Kang et al. 2014), implying a possible role in regulating ERV expression; *SUPT16H* (log BF = 11.932724) is involved in multiple processes that require DNA as a template such as mRNA elongation, DNA replication and DNA repair, implying its possible involvement in expression and/or deletion of ERVs (Keller et al. 2001; Belotserkovskaya et al. 2003; Pavri et al. 2006). In addition, the top ranked gene in mammals is *INTU* (log BF = 15.062398), which is essential for embryonic development in the aspects of cilia formation and normal orientation of elongating ciliary microtubules (Toriyama et al. 2016). However, its potential role in the interaction with ERVs is difficult to imply.

References

Afonina IS, Zhong Z, Karin M, Beyaert R. 2017. Limiting inflammation—The negative regulation of NF-B and the NLRP3 inflammasome. Nat. Immunol. 18(8):861.

Aiewsakun P, Katzourakis A. 2017. Marine origin of retroviruses in the early Palaeozoic Era. Nat. Commun. 8:13954.

Antony JM, Van Marle G, Opii W, Butterfield DA, Mallet F, Yong VW, Wallace JL, Deacon RM, Warren K, Power C. 2004. Human endogenous retrovirus glycoprotein-mediated induction of redox reactants causes oligodendrocyte death and demyelination. Nat. Neurosci. 7(10):1088–1095.

Bamforth SD, Bragança J, Eloranta JJ, Murdoch JN, Marques FIR, Kranc KR, Farza

H, Henderson DJ, Hurst HC, Bhattacharya S. 2001. Cardiac malformations, adrenal agenesis, neural crest defects and exencephaly in mice lacking Cited2, a new Tfap2 co-activator. Nat. Genet. 29(4):469–474.

Bannert N, Kurth R. 2006. The evolutionary dynamics of human endogenous retroviral families. Annu. Rev. Genomics Hum. Genet. 7:149–173.

Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, Robinson-Rechavi M. 2008. Bgee: Integrating and comparing heterogeneous transcriptome data among species. In: International Workshop on Data Integration in the Life Sciences. Berlin, Heidelberg: Springer. p. 124–131.

Belotserkovskaya R, Oh S, Bondarenko VA, Orphanides G, Studitsky VM, Reinberg D. 2003. FACT facilitates transcription-dependent nucleosome alteration. Science 301(5636):1090–1093.

Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M. 2004. Long-term reinfection of the human genome by endogenous retroviruses. Proc. Natl. Acad. Sci. 101(14):4894–4899.

Blaise S, de Parseval N, Benit L, Heidmann T. 2003. Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc. Natl. Acad. Sci. 100(22):13013–13018.

Chuong EB, Rumi MAK, Soares MJ, Baker JC. 2013. Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat. Genet. 45(3):325–329.

Cui J, Zhao W, Huang Z, Jarvis ED, Gilbert MTP, Walker PJ, Holmes EC, Zhang G. 2014. Low frequency of paleoviral infiltration across the avian phylogeny. Genome Biol. 15(12):539.

Douzery EJP, Scornavacca C, Romiguier J, Belkhir K, Galtier N, Delsuc F, Ranwez V. 2014. OrthoMaM v8: A database of orthologous exons and coding sequences for comparative genomics in mammals. Mol. Biol. Evol. 31(7):1923–1928.

Dupressoir A, Marceau G, Vernochet C, Benit L, Kanellopoulos C, Sapin V, Heidmann T. 2005. Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc. Natl. Acad. Sci. 102(3):725–730.

Ensembl. 2018. Archive Ensembl. Available from: http://www.ensembl.org/info/website/archives/index.html

Erdmann N, Zhao J, Lopez AL, Herek S, Curthoys N, Hexum TD, Tsukamoto T, Ferraris D, Zheng J. 2007. Glutamate production by HIV-1 infected human macrophage is blocked by the inhibition of glutaminase. J. Neurochem. 102(2):539–549.

Falandry C, Fourel G, Galy V, Ristriani T, Horard B, Bensimon E, Salles G, Gilson E, Magdinier F. 2010. CLLD8/KMT1F is a lysine methyltransferase that is important for chromosome segregation. J. Biol. Chem. 285(26):20234–20241.

Feschotte C, Gilbert C. 2012. Endogenous viruses: Insights into viral evolution and impact on host biology. Nat. Rev. Genet. 13(4):283.

Frank JA, Feschotte C. 2017. Co-option of endogenous viral sequences for host cell function. Curr. Opin. Virol. 25:81–89.

Gaudet P, Livstone MS, Lewis SE, Thomas PD. 2011. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief. Bioinform. 12(5):449–462.

Gifford RJ, Blomberg J, Coffin JM, Fan H, Heidmann T, Mayer J, Stoye J, Tristem M,

Johnson WE. 2018. Nomenclature for endogenous retrovirus (ERV) loci. Retrovirology 15(1):59.

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36(10):3420–3435.

Hamerman JA, Pottle J, Ni M, He Y, Zhang ZY, Buckner JH. 2016. Negative regulation of TLR signaling in myeloid cells—implications for autoimmune diseases. Immunol. Rev. 269(1):212–227.

Hayward A, Katzourakis A. 2015. Endogenous retroviruses. Curr. Biol. 25(15):R644–R646.

Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov D V., Blaser H, Raz E, Moens CB, et al. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell 129(1):69–82.

Itou D, Shiromoto Y, Shin-ya Y, Ishii C, Nishimura T, Ogonuki N, Ogura A, Hasuwa H, Fujihara Y, Kuramochi-Miyagawa S, et al. 2015. Induction of DNA methylation by artificial piRNA production in male germ cells. Curr. Biol. 25(7):901–906.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2015. Phylogenomic analyses data of the avian phylogenomics project. Gigascience 4(1):4.

Kalmykova AI, Klenov MS, Gvozdev VA. 2005. Argonaute protein PIWI controls mobilization of retrotransposons in the Drosophila male germline. Nucleic Acids Res. 33(6):2052–2059.

Kang HJ, Lee MH, Kang HL, Kim SH, Ahn JR, Na H, Na TY, Kim YN, Seong JK, Lee MO. 2014. Differential regulation of estrogen receptor α expression in breast cancer cells by metastasis-associated protein 1. Cancer Res. 74(5):1484–1494.

Kapusta A, Suh A. 2017. Evolution of bird genomes—a transposon's-eye view. Ann. N. Y. Acad. Sci. 1389(1):164–185.

Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. Proc. Natl. Acad. Sci. 114(8):E1460–E1469.

Kassiotis G, Stoye JP. 2016. Immune responses to endogenous retroelements: Taking the bad with the good. Nat. Rev. Immunol. 16(4):207–219.

Katzourakis A, Magiorkinis G, Lim AG, Gupta S, Belshaw R, Gifford R. 2014. Larger mammalian body size leads to lower retroviral activity. PLoS Pathog. 10(7):e1004214.

Keller DM, Zeng X, Wang Y, Zhang QH, Kapoor M, Shu H, Goodman R, Lozano G, Zhao Y, Lu H. 2001. A DNA damage-induced p53 serine 392 kinase complex contains CK2, hSpt16, and SSRP1. Mol. Cell 7(2):283–292.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860–921.

Lang G, Bonnet J, Umlauf D, Karmodiya K, Koffler J, Stierle M, Devys D, Tora L. 2011. The tightly controlled deubiquitination activity of the human SAGA complex differentially modifies distinct gene regulatory elements. Mol. Cell. Biol. 31(18):3734–3744.

Loiodice I. 2004. The entire Nup107-160 complex, including three new members, is targeted as one entity to kinetochores in mitosis. Mol. Biol. Cell 15(7):3333–

3344.

Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R. 2012. Env-less endogenous retroviruses are genomic superspreaders. Proc. Natl. Acad. Sci. 109(19):7385–7390.

Makova KD, Hardison RC. 2015. The effects of chromatin organization on variation in mutation rates in the genome. Nat. Rev. Genet. 16(4):213–223.

Manakov SA, Pezic D, Marinov GK, Pastor WA, Sachidanandam R, Aravin AA. 2015. MIWI2 and MILI have differential effects on piRNA biogenesis and DNA methylation. Cell Rep. 12(8):1234–1243.

McCarthy EM, McDonald JF. 2004. Long terminal repeat retrotransposons of Mus musculus. Genome Biol. 5(3):R14.

Meade A, Pagel M. 2017. BayesTraits V3.0.1. Available from: http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.1/BayesTraitsV3.0.1.html

NCBI. 2018. NCBI Genome. Available from: https://www.ncbi.nlm.nih.gov/genome

Ng A, Xavier RJ. 2011. Leucine-rich repeat (LRR) proteins: Integrators of pattern recognition and signaling in immunity. Autophagy 7(9):1082–1084.

O'Connor TD, Mundy NI. 2009. Genotype-phenotype associations: Substitution models to detect evolutionary associations between phenotypic variables and genotypic evolutionary rate. Bioinformatics 25(12):i94-i100.

Ogasawara H, Hishikawa T, Sekigawa I, Hashimoto H, Yamamoto N, Maruyama N. 2000. Sequence analysis of human endogenous retrovirus clone 4-1 in systemic lupus erythematosus. Autoimmunity 33(1):15–21.

Ogasawara H, Kageyama M, Yamaji K, Takasaki Y. 2010. The possibility that autoimmune disease can be induced by a molecular mimicry mechanism

between autoantigen and human endogenous retrovirus. Lupus 19(1):111–113.

Ohshima N, Kudo T, Yamashita Y, Mariggiò S, Araki M, Honda A, Nagano T, Isaji C, Kato N, Corda D, et al. 2015. New members of the mammalian glycerophosphodiester phosphodiesterase family: GDE4 and GDE7 produce lysophosphatidic acid by lysophospholipase D activity. J. Biol. Chem. 290(7):4260–4271.

Pagel M. 1997. Inferring evolutionary processes from phylogenies. Zool. Scr. 26:331–348.

Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature 401(6756):877–884.

Pavri R, Zhu B, Li G, Trojer P, Mandal S, Shilatifard A, Reinberg D. 2006. Histone H2B Monoubiquitination Functions Cooperatively with FACT to Regulate Elongation by RNA Polymerase II. Cell 125(4):703–717.

Perron H, Garson JA, Bedin F, Beseme F, Paranhos-Baccala G, Komurian-Pradel F, Mallet F, Tuke PW, Voisset C, Blond JL, et al. 1997. Molecular identification of a novel retrovirus repeatedly isolated from patients with multiple sclerosis. Proc. Natl. Acad. Sci. U. S. A. 94(14):7583–7588.

Raffety AE. 1996. Hypothesis testing and model selection. In: Gilks WR, Richardson S, Spiegelhalter D, editors. Markov chain Monte Carlo in practice. Springer. p. 163–188.

Reactome. 2019. Reactome:R-HSA-5578749. Available from: https://reactome.org/PathwayBrowser/#/R-HSA-211000&SEL=R-HSA-5578744&PATH=R-HSA-74160&DTAB=MT

Le Rouzic E, Mousnier A, Rustum C, Stutz F, Hallberg E, Dargemont C, Benichou S.

2002. Docking of HIV-1 vpr to the nuclear envelope is mediated by the interaction with the nucleoporin hCG1. J. Biol. Chem. 277(47):45091–45098.

Sha M, Lee X, Li X ping, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, et al. 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403(6771):785.

Sharpe AH, Wherry EJ, Ahmed R, Freeman GJ. 2007. The function of programmed cell death 1 and its ligands in regulating autoimmunity and infection. Nat. Immunol. 8(3):239–245.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. 100(16):9440–9445.

Stoye JP. 2001. Endogenous retroviruses: Still active after all these years? Curr. Biol. 11(22):R914–R916.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102(43):15545–15550.

The Uniprot Consortium. 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47(d1):D506-515.

Toriyama M, Lee C, Taylor SP, Duran I, Cohn DH, Bruel AL, Tabler JM, Drew K, Kelly MR, Kim S, et al. 2016. The ciliopathy-associated CPLANE proteins direct basal body recruitment of intraflagellar transport machinery. Nat. Genet. 48(6):648.

Tselis A. 2011. Evidence for viral etiology of multiple sclerosis. Semin. Neurol. 31(3):307–316.

Viganò S, Perreau M, Pantaleo G, Harari A. 2012. Positive and negative regulation of cellular immune responses in physiologic conditions and diseases. Clin. Dev. Immunol. 2012:485781.

Xiol J, Cora E, Koglgruber R, Chuma S, Subramanian S, Hosokawa M, Reuter M, Yang Z, Berninger P, Palencia A, et al. 2012. A Role for Fkbp6 and the Chaperone Machinery in piRNA Amplification and Transposon Silencing. Mol. Cell 47(6):970–979.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24(8):1586–1591.

Zheng W, Satta Y. 2018. Functional Evolution of Avian RIG-I-Like Receptors. Genes (Basel). 9(9):456.

Supplementary

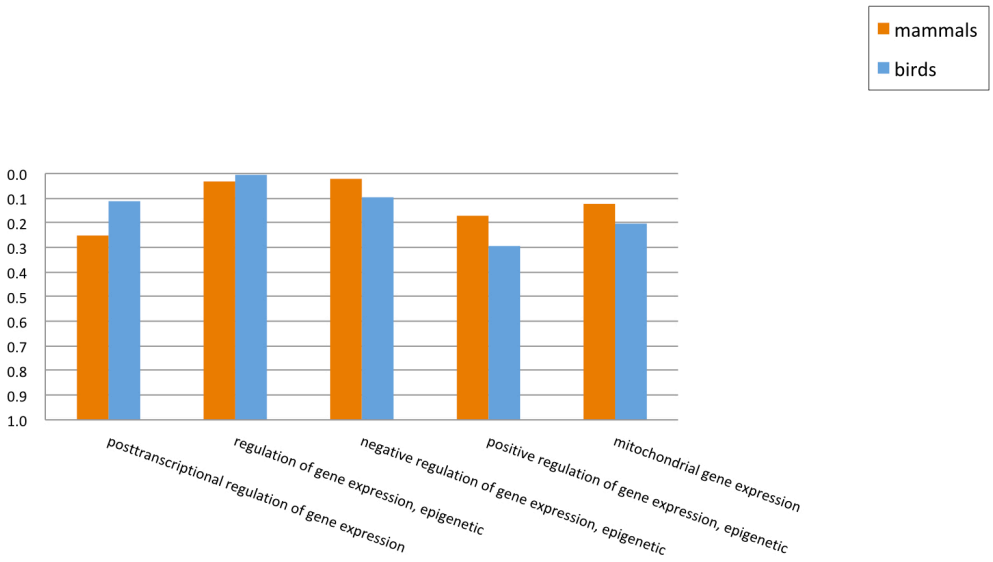


Figure 3.S1. Rankings of normalized enrichment score (NES) of potential ERV-load-related GO terms based on G method results.

Chapter 4

General Discussion

In conclusion, this thesis provided evidence of host-ERV evolutionary interaction in mammals and birds. I have and testified the long history of host-ERV relationships comprising of a balance between host-parasite conflict, tolerance and co-option. Besides, results suggest that the ERV load difference between mammals and birds can be related to gene silencing and immune response. Here, I'm going to discuss the significance of the findings in this thesis and the remained questions, as well as an extended outlook for future studies.

Chapters 2 and 3 have provided evidence that the immune system plays an important role in host-ERV evolutionary interaction. The relevant components of the immune system include innate sensors of retroviruses, including the innate RNA sensor RIG-I (Chapter 2), processes of regulation of immune responses, and even pathways of non-virus immunity (Chapter 3). This interaction reflects a long history, which lasts more than 100 million yeas, of host-parasite relationship between mammal/bird hosts and ERVs with the discrimination on ERVs as non-self. This is quite interesting since so many ERVs reside in mammalian genomes. The determinant for retaining the immune responses may be the necessity of resisting invasions of exogenous retroviruses. This suggests that immune pathways against retroviruses remain to be further explored and the recently discovered cGAS-STRING pathway (Gao et al. 2013) may not be the sole responsible pathway. Other pathways may exist,

especially in non-human vertebrates, and RIG-I and its related pathways could be revisited in future studies.

On the other hand, other biological processes also participated in the host-ERV evolutionary interaction of which gene silencing might play an even more important role than immunity (Chapter 3). Gene silencing may not only suppress ERV replication, but may also influence the regional deletion rate via its effect on chromatin organization (Makova and Hardison 2015) or other processes that are still unknown. Since epigenetic modifications are crucial for long-term gene silencing (Kim and Kim 2012; Mozzetta et al. 2015), further studies about the relation between epigenetic modifications and regional mutation rate, especially the deletion rate in vertebrate genomes are, are important for better understanding the role of gene silencing in genome size evolution and ERV load evolution.

Another notable finding from Chapter 3 is that rate-load associated genes are plenty (>24% showing log BF>2) in both mammals and birds, which may be associated with a distinguishable form of evolutionary interaction between the host and ERVs during long-term evolution. However, the specificity of this prevalence remains unknown. If a future study compares this result with host groups other than mammals and birds, we could elucidate if the prevalence of rate-load associated genes is specific to mammals and birds. Additionally, if future studies investigate the evolutionary interaction between host and other EVEs or TEs, we could determine whether this prevalence is specific to ERVs. The cause of this prevalence should also be investigated in future.

Overall, findings in this thesis added new evidences about the evolution of immune systems: the relationship between certain groups of parasites and the host is

possible to drive the evolution of host immune systems. My findings also suggest that the way in which eukaryotes demand for immune functions can largely vary. ERVs might have taken a larger part than expected in the diversification and phenotype evolution of eukaryotes, including biological processes in gene silencing and immunity.

Findings in this thesis can prompt new research directions. First, the difference in host-ERV evolutionary interaction between mammals and birds propose further studies into the diversity of evolutionary interaction with a certain clade of parasites across different hosts. Such studies may provide new findings about how viruses, or parasites in general, or symbiotic partners in the broader sense, could have taken part in shaping the diversity of eukaryotes. Second, different involvement of immune system between mammals and birds in this evolutionary interaction suggests that host-parasite evolutionary interaction may underlie the pathogenesis of infectious, inflammatory or autoimmune diseases, which is an very open subject to be studied. Furthermore, this study is a successful application of phylogenetic gene-phenotype association analyses and shows the potential of this method in addressing a broad range of biological questions in the light of evolution.

References

Gao D, Wu J, Wu YT, Du F, Aroh C, Yan N, Sun L, Chen ZJ. 2013. Cyclic GMP-AMP synthase is an innate immune sensor of HIV and other retroviruses. Science 341(6148):903–906.

Kim J, Kim H. 2012. Recruitment and Biological Consequences of Histone

Modification of H3K27me3 and H3K9me3. ILAR J. 53(3–4):232–239.

Makova KD, Hardison RC. 2015. The effects of chromatin organization on variation in mutation rates in the genome. Nat. Rev. Genet. 16(4):213–223.

Mozzetta C, Boyarchuk E, Pontis J, Ait-Si-Ali S. 2015. Sound of silence: The properties and functions of repressive Lys methyltransferases. Nat. Rev. Mol. Cell Biol. 16(8):499–513.