Reduction in major transcription factor binding domain families during the evolution of the avian lineage

Allie M. Graham[1,2*] and Jason S. Presnell[1]

[1] Oregon State University, Department of Integrative Biology, Corvallis, OR
[2] University of Miami, Coral Gables, FL

* Corresponding author: grahaall@oregonstate.edu

ABSTRACT

Transcription factors (TF) are characterized by certain DNA binding-domains (DBD) which regulate their binding specificity, and thus their ability to effect a change on gene expression of their downstream targets. TFs are central to organismal development, and morphology; therefore, they potentially are instrumental in producing phenotypic diversity. We measured TF abundance of 49 major TF DBD families in 48 bird genomes, which we then compared with 5 reptile genomes, in an effort to assess the degree to which TF DBD are potentially connected to increased phenotypic diversity in the avian lineage. We hypothesized that there would be increased TF DBD abundance in multiple TF families correlated with the increased phenotypic diversity found in birds; instead ultimately, we see a general loss of major TF DBD families, reflecting general genome reductions seen between reptiles and birds, with largest losses in TF DBD families associated with multiple developmental (feather, sex-determination, body-plan, immune, blood) and metabolic processes.

INTRODUCTION

Transcription factors (TFs) are proteins that bind to DNA in a sequence-specific manner and enhance or repress gene expression. In response to a broad range of stimuli, TFs coordinate the regulation of gene expression of essential genes for defining morphology, functional capacity, and developmental fate at the cellular level. Although transcription factor binding domains are very well conserved, the other associated domains, largely responsible for protein-protein interactions, readily diverge among homologs. Therefore, the structure and function of transcription factors are inherently modular. This attribute is thought to allow gene-regulatory networks to evolve via changes in TF domain architecture (Wray 2007; Jarvela & Hinman 2015), and could account for the seemingly large phenotypic difference between closely related groups (Liu *et al*. 2014; Nadimpalli *et al*. 2015).

A long-standing question has been whether changes in gene regulation or protein sequences have made a larger contribution to phenotypic diversity seen between species (Britten & Davidson 1969; King & Wilson 1975). Now, it is understood that changes in *cis*-regulatory systems more often underlie the evolution of morphological diversity than gene duplication/loss or protein function (Levine & Tjian 2003; Carroll 2008; Wittkopp & Kalay 2011). These *cis*-regulatory elements (CREs) typically regulate gene transcription by functioning as binding sites for transcription factors. However, another avenue would be through whole-scale changes in transcription factor function through changes in domain modularity, either through their DNA-binding domains (DBD) and/or through other domains usually involved in protein-protein interactions (Wagner & Lynch 2008, 2010; Schmitz *et al*. 2016), otherwise known as *trans*-regulatory elements.

Ultimately, observed phenotypic variation ranging from individual organisms to broad groups has been attributed to a combination of changes associated with *cis*- and *trans*-regulatory elements (Schmitz *et al*. 2016).

Although transcription factor diversity has been correlated with increased "complexity" across the eukaryotic lineage (Charoensawan *et al*. 2010; de Mendoza *et al*. 2013; Lehti-Shiu *et al*. 2016), no study has measured such transcription factor diversity within a specious, but highly related, animal clade. However, recently, forty-eight avian genomes representing all the major families of birds (Aves) have recently been published (OBrien *et al*. 2014; Zhang *et al*. 2014a; Eöry *et al*. 2015), providing a unique opportunity to do just that. Birds represent one of the most diverse vertebrate lineages, and has the distinction of being the tetrapod class with the most living species, with half of them being Passerines (over 10,000 species) (Gill & Wright 2006). Not only do birds live worldwide and range in size (5 cm - 2.75 m), but also vary widely in morphology, physiology and behavior, and have unique features (e.g., feathers). From a genomic standpoint, these organisms have relatively low rates of gene gain/loss in gene families and have similarly sized genomes (0.91-1.3 Gb) (Zhang *et al*. 2014b).

Here, we aimed to characterize the major metazoan TF families by their major DNA-binding domains (DBD) in the 48 avian and 5 reptile genomes. This is the first study to analyze the evolutionary history and phylogenetic distribution of transcription factors in the diverse genomes available for the avian group, and the closely related reptile lineage. We would expect an increase in TF DBD abundance in multiple TF families correlate with the increased phenotypic diversity found in birds, however we

observed no such evidence, and instead our results show a wholescale loss within major TF DBD families, reflecting general genome reductions seen between reptiles and birds.

MATERIALS AND METHODS

TF DBD Identification

We obtained data on complete genomes from publicly available databases for birds (http://avian.genomics.cn/en/jsp/database.shtml) and reptiles (http://crocgenomes.org/). A PfamScan was performed on the protein models, using a custom database containing the major DBD families, and selecting the gathering threshold option as a conservative approach, which can underestimate total counts for some domains but minimizes false positives (Eddy 2011). We looked for the presence of major DBDs, which were TFs selected based on previous studies (de Mendoza *et al*. 2013). For the major DBDs, in all cases, we defined a one-to-one relationship between TF class and DBD class (i.e., non-duplicates). In cases in which two or more DBDs were found in the same gene, those were relegated to a separate list, compared to those which showed only one DBD. DBDs that appeared just in combination with other DBDs (i.e., duplicates) were analyzed separately, to avoid an overestimation of TF numbers, due to problems detecting repeated domains (de Mendoza *et al*. 2013). We counted the number of genes/proteins containing a given DBD each species.

Transcription Factor DBD enrichment and characterization

We tested for enrichment of TF numbers using a Mann-Whitney U test, with a significance threshold of $P < 0.01$, as performed by de Mendoza *et al*. (2013). The

protein sequences were then filtered to include sequences of the DBD that were significantly different between reptiles and birds; these sequences were then subjected to a BLAST search, mapped, annotated, and analyzed using Blast2GO basic version (Conesa *et al*. 2005). Gene Ontology (GO) term maps were created for biological processes, cellular, and molecular functions, with a threshold of 10% (i.e., 1,500).

RESULTS

Transcription Factor DBD Identification

Of the 49 different DNA binding domains, a total of 34,318 non-duplicate and 19,668 duplicate proteins were identified (SUPP Tables). To ascertain which families may have experienced family expansion or contraction, comparisons were made against the 5 reptile genomes (crocodile, alligator, gharial, green sea turtle, and soft-shelled turtle). In the non-duplicate group, of the 49 DBD, 21 were significantly different (MWU, $P > 0.01$; Table 1, SUPP Table 1). A majority of these families experienced a reduction in DBD presence, with an average decrease of 51.14%, while only 2 families showed an increase (Homeobox_KN, and HTH_psq), though this increase was substantial (~393%). In the duplicate group, only 11 showed a significant difference (Table 2, SUPP Table 1), also showing a general reduction in DBD presence. In this category, all families decreased in number (~60%), with evidence of complete loss of members with certain DBD combinations (Forkhead, HMG box, MH1, and Tub). Most of these DBD families are not significantly different in proteins with only one representative DBD (Table 1).

Transcription Factor DBD Characterization

We then assessed any additional commonalities of transcription factor families (beyond DNA binding, and general gene expression) that underwent a change between reptiles and birds, utilizing gene-ontology (GO) for the proteins that contained the 21 non- duplicate DBDs that were significantly different (Table 1). This resulted in additional characterization of 21,463 proteins across the 47 bird and 5 reptile genomes. The GO families associated with Cellular and Molecular functions were, not surprisingly, variations on DNA-binding, and cell parts/nucleus, while Biological Processes were those associated with various metabolic (nitrogen, macromolecule, nucleic acid, protein) and developmental processes (at Level 3). Of those GO term classes pertaining to development, the most common terms were those associated with skeletal development, immune system development, circulatory system development, hematopoietic/blood development, embryonic morphogenesis, nervous system development, and animal organ development (Figure 1).

DISCUSSION

The results of this analysis suggest there was an overall major reduction in transcription factor families across the avian lineage, and did not show any substantial expansions associated with any specific DBD family. The only groups which saw any increase in number were the Homeobox KN and HTH psq families, however that increase was through a gain of ~1-2 additional gene in each group, on average (Table 1).

Although avian and reptile genomes are generally similarly sized currently, a reduction in genome size between reptiles and birds did occur in the saurischian dinosaur

lineage between 230 and 250 million years ago (Organ *et al*. 2007). This coincided with a major reduction in repetitive elements, intron size, and whole-scale loss of syntenic protein coding regions, typically attributed to the general metabolic requirements for flight (Organ & Shedlock 2009; Zhang & Edwards 2012; Lovell *et al*. 2014; Wright *et al*. 2014; Zhang *et al*. 2014b). Thus, the reduction of TF DBD families seen in our results, mirror the general genome reduction seen in other studies. Despite this fact, it is somewhat surprising to see a lack of any instances of TF DBD family expansions, since increases in the number of regulatory proteins, including TFs, have frequently been connected to phenotypic innovations (Miyata & Suga 2001; Levine & Tjian 2003; Kusserow *et al*. 2005; Schmitz *et al*. 2016).

Nonetheless, such a pattern suggests that TF DBD families are not correlated with avian diversification, and can potentially be ascribed to other factors, such as non-TF gene family duplications, or *cis*-regulatory changes (Zhang *et al*. 2014b; Seki *et al*. 2017). Duplications of non-TF gene families can play a major role in species evolution: redundancy provides a medium for novelty while maintaining initial function (Lynch & Conery 2000; Zhang 2003). However, another possibility is that the absence of an increase TF DBD between reptiles and birds may instead suggest changes in TF modularity, appearing through increased interconnectivity/occurrence of DBD and domains involved in protein-protein interactions. Thus, surveying domains associated with protein-protein interactions could potentially reveal an additional facet to how TF families may have evolved, in conjunction with a reduction in genome size, during the reptile-bird transition. An increase or an enrichment in these domains would hint at a scenario of TF families being re-organized, or their functions becoming specialized, and

thus would alter their expression profiles or binding properties causing an effect on expression of many target genes, often with a major functional impact (Lespinet *et al*. 2002).

It is interesting to note that the TF DBD families who experienced the sharpest declines were ones associated with heart development (T-box), vocal learning (Forkhead), feather formation (Ets), wing development (HMG box), sex determination (HMG box, DM), immune function (HMG box, IRF, STAT bind, RHD), and aspects of blood (Runt, GATA) (Table 3). These genes have been drivers of major physiological/developmental changes between reptiles and birds (Brusatte *et al*. 2015; Chatterjee 2015), especially the development of feathers from scales (Chuong *et al*. 2000), sex-determination through chromosomal differences rather than temperature (Sarre *et al*. 2004), and even changes in immune system functionality (Zimmerman *et al*. 2010). Even so, it is difficult to speculate how reductions in these major families may be associated with such changes seen between avian and reptile lineages.

CONCLUSTIONS

Ultimately, these results represent the first foray into TF DBD characterization between the avian and reptile genomes. Furthermore, the results of these analyses strengthen the notion that *cis*-regulatory regions and non-TF protein-coding gene families are behind much of the extant avian diversification. Still, whole-scale reductions in TF DBD families in the genome likely posed a significant issue, unless these families were comprised of multiple members that were functionally redundant. Overall, the results of this analysis represent a broad characterization of TF DBD family composition. The

specific composition of TF families should be probed further, especially those with the largest reductions seen in this study. In addition, whether non-major TF DBD families have also seen a general reduction requires future analysis, as does the composition of domains associated with protein-protein interactions.

Acknowledgements

We would like to acknowledge our PhD advisors (Kevin McCracken and William Browne) who allowed us to explore other topics and work on side-projects, at the same time as our thesis work.

Author contributions

AMG and JSP conceived, designed the study, and performed the data-collection. AMG wrote the manuscript. All authors discussed the results and implications, and commented on the manuscript.

References

Alonso CR (2002) Hox proteins: sculpting body parts by activating localized cell death. *Current Biology* **12**, R776-R778.

Anrather J, Racchumi G, Iadecola C (2005) cis-Acting element-specific transcriptional activity of differentially phosphorylated nuclear factor- κ B. *Journal of Biological Chemistry* **280**, 244-252.

Britten RJ, Davidson EH (1969) Gene regulation for higher cells: a theory. *Science* **165**, 349-357.

Brusatte SL, O'Connor JK, Jarvis ED (2015) The origin and diversification of birds. *Current Biology* **25**, R888-R898.

Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25-36.

Charoensawan V, Wilson D, Teichmann SA (2010) Lineage-specific expansion of DNA-binding transcription factor families. *Trends in Genetics* **26**, 388-393.

Chatterjee S (2015) *The rise of birds: 225 million years of evolution* JHU Press.

Chuong C-M, Chodankar R, Widelitz RB, Jiang T-X (2000) Evo-devo of feathers and scales: building complex epithelial appendages. *Current opinion in genetics & development* **10**, 449.

Conesa A, Götz S, García-Gómez JM*, et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676.

de Mendoza A, Sebé-Pedrós A, Šestak MS*, et al.* (2013) Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proceedings of the National Academy of Sciences* **110**, E4858-E4866.

Eddy SR (2011) Accelerated profile HMM searches. *PLoS computational biology* **7**, e1002195.

Eöry L, Gilbert MTP, Li C*, et al.* (2015) Avianbase: a community resource for bird genomics. *Genome biology* **16**, 21.

Escalante CR, Yie J, Thanos D, Aggarwal AK (1998) Structure of IRF-1 with bound DNA reveals determinants of interferon regulation. *Nature* **391**, 103-106.

Gill FB, Wright MT (2006) *Birds of the world: recommended English names* Princeton University Press Princeton, New Jersey.

Hannenhalli S, Kaestner KH (2009) The evolution of Fox genes and their role in development and disease. *Nature reviews. Genetics* **10**, 233.

Jarvela AMC, Hinman VF (2015) Evolution of transcription factor function as a mechanism for changing metazoan developmental gene regulatory networks. *EvoDevo* **6**, 3.

Kagoshima H, Shigesada K, Satake M*, et al.* (1993) The runt domain identifies a new family of heterometric transcriptional regulators. *Trends in Genetics* **9**, 338-341.

King M-C, Wilson AC (1975) *Evolution at two levels in humans and chimpanzees* na.

Kisseleva T, Bhattacharya S, Braunstein J, Schindler C (2002) Signaling through the JAK/STAT pathway, recent advances and future challenges. *Gene* **285**, 1-24.

Kusserow A, Pang K, Sturm C, *et al.* (2005) Unexpected complexity of the Wnt gene family in a sea anemone. *Nature* **433**, 156-160.
Lehti-Shiu MD, Panchy N, Wang P, Uygun S, Shiu S-H (2016) Diversity, expansion, and evolutionary novelty of plant DNA-binding transcription factor families. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*.
Lespinet O, Wolf YI, Koonin EV, Aravind L (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome research* **12**, 1048-1059.
Levine M, Tjian R (2003) Transcription regulation and animal diversity. *Nature* **424**, 147-151.
Liu H, Chang L-H, Sun Y, Lu X, Stubbs L (2014) Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies. *Genome biology and evolution* **6**, 510-525.
Lotze MT, Tracey KJ (2005) High-mobility group box 1 protein (HMGB1): nuclear weapon in the immune arsenal. *Nature Reviews Immunology* **5**, 331-342.
Lovell PV, Wirthlin M, Wilhelm L, *et al.* (2014) Conserved syntenic clusters of protein coding genes are missing in birds. *Genome biology* **15**, 1.
Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151-1155.
Miyata T, Suga H (2001) Divergence pattern of animal gene families and relationship with the Cambrian explosion. *Bioessays* **23**, 1018-1027.
Morgan BA, Orkin RW, Noramly S, Perez A (1998) Stage-specific effects ofSonic HedgehogExpression in the epidermis. *Developmental biology* **201**, 1-12.
Nadimpalli S, Persikov AV, Singh M, Lieb JD (2015) Pervasive variation of transcription factor orthologs contributes to regulatory network evolution. *PLoS genetics* **11**, e1005011-e1005011.
OBrien SJ, Haussler D, Ryder O (2014) The birds of Genome10K. *GigaScience* **3**, 32.
Organ CL, Shedlock AM (2009) Palaeogenomics of pterosaurs and the evolution of small genome size in flying vertebrates. *Biology letters* **5**, 47-50.
Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV (2007) Origin of avian genome size and structure in non-avian dinosaurs. *Nature* **446**, 180-184.
Patient RK, McGhee JD (2002) The GATA family (vertebrates and invertebrates). *Current opinion in genetics & development* **12**, 416-422.
Plageman TF, Yutzey KE (2005) T‐box genes and heart development: Putting the "T" in heart. *Developmental Dynamics* **232**, 11-20.
Sarre SD, Georges A, Quinn A (2004) The ends of a continuum: genetic and temperature‐dependent sex determination in reptiles. *Bioessays* **26**, 639-645.
Scharff C, Haesler S (2005) An evolutionary perspective on FoxP2: strictly for the birds? *Current opinion in neurobiology* **15**, 694-703.
Schmitz JF, Zimmer F, Bornberg-Bauer E (2016) Mechanisms of transcription factor evolution in Metazoa. *Nucleic acids research*, gkw492.
Seki R, Li C, Fang Q, *et al.* (2017) Functional roles of Aves class-specific cis-regulatory elements on macroevolution of bird-specific features. *Nature Communications* **8**, 14229.

Smith CA, Roeszler KN, Ohnesorg T*, et al.* (2009) The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. *Nature* **461**, 267-271.
Wagner GP, Lynch VJ (2008) The gene regulatory logic of transcription factor evolution. *Trends in Ecology & Evolution* **23**, 377-385.
Wagner GP, Lynch VJ (2010) Evolutionary novelties. *Current Biology* **20**, R48-R52.
Wallis M, Waters P, Graves JA (2008) Sex determination in mammals—before and after the evolution of SRY. *Cellular and molecular life sciences* **65**, 3182-3195.
Welten M, Verbeek FJ, Meijer AH, Richardson MK (2005) Gene expression and digit homology in the chicken embryo wing. *Evolution & development* **7**, 18-28.
Wittkopp PJ, Kalay G (2011) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* **13**, 59-69.
Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* **8**, 206-216.
Wright NA, Gregory TR, Witt CC (2014) Metabolic 'engines' of flight drive genome size reduction in birds **281**, 20132780.
Zhang G, Li B, Li C*, et al.* (2014a) Comparative genomic data of the Avian Phylogenomics Project. *GigaScience* **3**, 26.
Zhang G, Li C, Li Q*, et al.* (2014b) Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311-1320.
Zhang J (2003) Evolution by gene duplication: an update. *Trends in Ecology & Evolution* **18**, 292-298.
Zhang Q, Edwards SV (2012) The evolution of intron size in amniotes: a role for powered flight? *Genome biology and evolution* **4**, 1033-1043.
Zimmerman L, Vogel L, Bowden R (2010) Understanding the vertebrate immune system: insights from the reptilian perspective. *Journal of Experimental Biology* **213**, 661-671.

Table 1: Non-duplicate DNA-Binding Domains (DBD) that are significantly different between reptiles, and birds (Mann-Whitney U test, $P>0.01$): "reptiles" represents the average number across the 5 lineages, and "birds" represents the average across the 48 lineages.

DNA binding domain	**Reptiles**	**Birds**	**More/Less in Birds**	***P-value***
MADF_DNA_bdg	121.60	2.94	Less	7.70E-07
zf-BED	9.60	2.89	Less	7.70E-07
Homeobox	62.40	31.09	Less	5.39E-06
T-box	20.60	11.38	Less	1.77E-05
Fork_head	37.40	19.55	Less	3.23E-05
zf-C2H2	480.20	210.49	Less	3.23E-05
Ets	12.00	8.09	Less	1.37E-04
HMG_box	50.40	32.51	Less	1.42E-04
GATA	9.80	4.87	Less	1.62E-04
Homeobox_KN	0.20	1.57	*More*	5.82E-04
IRF	9.60	5.36	Less	8.06E-04
STAT_bind	6.40	3.57	Less	0.001040416
DM	3.60	0.85	Less	0.001419029
P53	3.80	1.83	Less	0.001921538
Runt	3.80	1.91	Less	0.002920399
CG-1	1.60	0.49	Less	0.003327485
ARID	15.20	10.77	Less	0.003717641
SRF-TF	6.20	2.91	Less	0.004266322
HTH_psq	1.20	2.43	*More*	0.0083064
Tub	5.40	3.15	Less	0.008685782
RHD	9.60	6.98	Less	0.009252932

Table 2: Duplicate DNA-Binding Domains (DBD) that are significantly different between reptiles, and birds (Mann-Whitney U test, $P>0.01$): "reptiles" represents the average number across the 5 lineages, and "birds" represents the average across the 48 lineages.

DNA binding domain	**Reptiles**	**Birds**	**More/Less in Birds**	***P-value***
YL1 nuclear protein	2.2	0.51	Less	1.62E-04
zf-MIZ	1	0.04	Less	3.59E-04
TEA	1.2	0.06	Less	4.29E-04
HLH	4.6	1.21	Less	5.93E-04
Fork_head	0.8	0	Less	9.05E-04
HMG_box	1	0	Less	9.05E-04
MH1	0.6	0	Less	9.05E-04
Tub	0.6	0	Less	9.05E-04
bZIP_Maf	35	25.13	Less	0.001431342
MADF_DNA_bdg	3.8	0.06	Less	0.005172839
bZIP_2	51.2	38.23	Less	0.007530705

Table 3: Significantly different DBD categories (see Table 1), and their respective biological function

Domain	Domain "Function"	Citation
zf-BED	general	PFAM: PF02892
Homeobox	general development	PFAM: PF00046
T-box	heart development	(Plageman & Yutzey 2005)
Fork_head	general	(Hannenhalli & Kaestner 2009)
	vocal learning	(Scharff & Haesler 2005)
zf-C2H2	general	PFAM: PF00096
Ets	feather formation	(Morgan *et al.* 1998)
HMG_box	wing development	(Welten *et al.* 2005)
	sex determination	(Wallis *et al.* 2008)
	immune function	(Lotze & Tracey 2005)
GATA	blood	(Patient & McGhee 2002)
Homeobox_KN	anterior-posterior axis formation	(Alonso 2002) InterPro: IPR001356
IRF	immune function	(Escalante *et al.* 1998) SMART: SM00348
STAT_bind	immune function	(Kisseleva *et al.* 2002) InterPro: IPR012345
DM	sex determination/differentiation	(Smith *et al.* 2009) InterPro: IPR001275
P53	cell cycle (general)	InterPro: IPR011615
Runt	blood	(Kagoshima *et al.* 1993) InterPro: IPR013524
CG-1	general	PFAM: PF03859
ARID	general	PFAM: PF01388
SRF-TF	cell cycle (general)	PFAM: PF00319
HTH_psq	cell cycle (general)	PFAM: [F05225
Tub	cell cycle (general)	PFAM: PF01167
RHD	immune/p53	(Anrather *et al.* 2005) InterPro: IPR011539

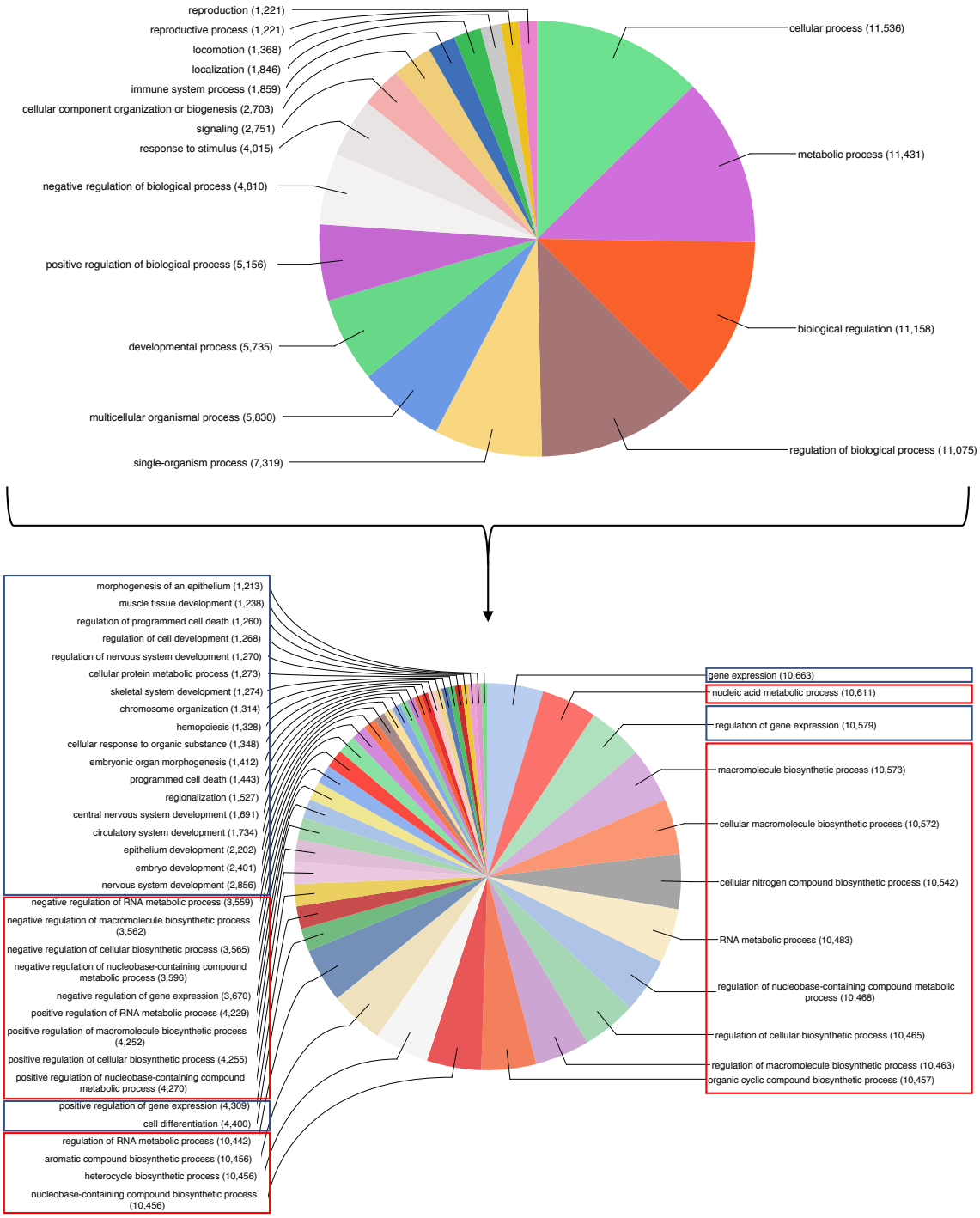

Figure 1: Gene Ontology (GO) terms associated with the significantly different counts of DNA-Binding Domains between reptiles and birds – blue, developmental GO terms; red, metabolism GO terms; top panel level 2 (more general), bottom panel level 5 (more specific).