Brief Communication: The predictable network topology of evolutionary genomic constraint

Authors: Katharina C. Wollenberg Valero[1*]

Affiliations:

[1] School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland.

*Corresponding author. Email: katharina.wollenbergvalero@ucd.ie

Abstract

Large-scale comparative genomics studies offer valuable resources for understanding both functional and evolutionary rate constraints. It is suggested that constraint aligns with the topology of genomic networks, increasing towards the center, with intermediate nodes combining relaxed constraint with higher contributions to the phenotype due to pleiotropy. However, this pattern has yet to be demonstrated in vertebrates. This study shows that constraint intensifies towards the network's center in placental mammals. Genes with rate changes associated with emergence of hibernation cluster mostly towards intermediate positions, with higher constraint in faster-evolving genes, which is indicative of a "sweet spot" for adaptation. If this trend holds universally, network node metrics could predict high-constraint regions even in clades lacking empirical constraint data.

Key words: Constraint, networks, evolution, pleiotropy, cost of complexity

Introduction

In a genome, near-infinite numbers of gene modifications and resulting allelic combinations can emerge through mutation and natural selection. Phenotypic convergence and the reuse or re-evolution of genes for similar adaptations raise the question of what genomic characteristics enable such seemingly improbable evolutionary outcomes (Futuyma 2010). Several studies have provided transcript-level, base-level, or genome-wide scores of constraint for various organisms, including *Anolis* lizards (Tollis et al. 2018), lacertid lizards (as tree root-to-tip-length, (Garcia-Porta et al. 2019), whales (Tollis et al. 2019), birds (Zhang et al. 2014; Feng et al. 2020), and humans (Chen et al. 2023). Some genomic regions appear to retain mutations less frequently, leading to the formation of, for example, ultra conserved elements (Katzman et al. 2007; Faircloth et al. 2012), which were conserved to more than 98% across 240 pla- cental mammal genomes (Christmas et al. 2023). Other genomic regions showed accelerated evolution crucial for lineage-specific adaptations across placental mammals (Christmas et al. 2023). In addition, both highly-

and lowly- constrained regions across genomes appear to be homologous between mammals and birds, with approximately 13% found within protein-coding sequences (Zhang et al. 2014). What remains unexplored in these studies is to which extent such elements of constraint arise through functional interactions among gene products, which have been proposed to limit the degrees of freedom for selection (Payne and Wagner 2019). Empirical classification of evolutionary constraint at the level of protein-protein interactions (PPI) networks is limited, particularly due to the scarcity of multi-clade genomic alignments, especially in multicellular organisms (but see studies on yeast, Frost et al. 2012; Schoenrock et al. 2017; Wollenberg Valero 2020). The 240-mammal alignment within the Zoonomia genomic resource now enables a first test for the influence of network topology in shaping levels of evolutionary rate constraint and the location of nodes relative for mammal-specific adaptations (Christmas et al. 2023).

The recent publication by the Zoonomia consortium (Christmas et al. 2023) presented a comprehensive assessment of mammalian evolutionary rate constraint at single-base resolution using the "Phylop" metric. This metric (Pollard et al. 2010; Hubisz et al. 2011) describes the deviation of nucleotide substitution rates from neutrality in a clade-specific manner, and allows discriminating between rate acceleration and deceleration. Significant Phylop scores (-log10(p-value)) different from 0 (neutral evolution) indicate higher levels of sequence constraint (positive) or acceleration (negative), with constraint reflecting functional importance maintained by purifying selection. Remarkably, the study found that around 10% of mammalian genomes exhibited strong constraint, often within crucial developmental pathways. Notably, while approximately 80% of highly constrained regions occur outside of protein-coding exons, species-specific adaptations were observed in genes with relaxed constraint, such as a correlation between olfactory gene number and olfactory turbinals, enhancing environmental sensing capabilities (Christmas et al. 2023). Additionally, the study revealed that single-base constraint coincided with higher-level functional elements, including CTCF transcription factor binding sites and functionally important peptide regions like start/stop codons and splice sites (Christmas et al. 2023). Such constraints collectively ensure the functionality of cellular and organismal processes, extending from peptide functionality to genomic network constraint (Wollenberg Valero 2020).

In a previous study using yeast as an example, I demonstrated a role of network topology in determining constraint levels. Three network statistical parameters—average shortest path length (ASPL), betweenness centrality (BC), and neighborhood connectivity (NC)—can be utilized to identify adaptable vs. resilient regions within the protein-protein interaction (PPI) network, contributing to our understanding of constraint dynamics (Wollenberg Valero 2020). The highest levels of constraint are expected in the central (hub) nodes of the network, characterized by the highest BC. Moving towards the network periphery, rate constraint gradually diminishes while passing through intermediate nodes with the highest NC, indicating a greater number of connections to other nodes. Finally, the peripheral nodes, marked by the highest ASPL and the fewest connections to other nodes, demonstrate the lowest levels of constraint (Wollenberg Valero 2020).

Here, I present a visualization and analysis of the PPI network topological architecture of mammalian genomic constraint, testing four hypotheses which are (i) that constraint is predicted by the structure of the network, (ii) that constraint differs between network node categories, (iii) that hibernation-associated genes are located in intermediate nodes in the network and differ in constraint values, and lastly (iv) that the positions of hibernation-relevant genes in the network are not just an outcome of chance.

Results and Discussion

In mammal genomes (modeled on a high-confidence STRING PPI network of *Homo sapiens,* Figure 1), ASPL, NC and BC emerge as significant predictors of base-level constraint summarized at gene level ($F = 449.6$, $p < 0.0001$, $df = 15,780$, explaining 8.55% of overall variance, Supplementary Table S1), which divide protein-coding genes (nodes) within the network into three regions (center, intermediate and periphery). Consistent with expectations, mammalian evolutionary constraint exhibits higher values in central hub nodes of the network, gradually declining towards the network periphery with significant pairwise differences in mean Phylop across three node categories—H (hub), I (intermediate), and P (peripheral; with a small but very strongly supported effect for the full model $F = 218.6$, $p < 2e\text{-}16$,

84 df=14,825, Supplementary Table S2, and large to huge pairwise differences Supplementary Table S3). This
85 pattern mirrors the findings previously observed in yeast evolutionary constraint (Wollenberg Valero 2020).
86 As predicted, only one of the genes whose evolutionary rate was significantly associated with hibernating
87 phenotype in mammals was situated in hub nodes; instead, most were associated with intermediate nodes
88 with only four out of 18 peripheral node positions (Figure 2). Surprisingly, faster-evolving genes exhibited
89 higher rate constraint compared to slower-evolving genes with large effect size and moderate statistical
90 support ($F = 6.229$, $p= 0.027$, Supplementary Table S5), and significant pairwise differences of small to
91 medium effect size. This at first glance counterintuitive observation can be explained by the "cost of
92 complexity" hypothesis (Wagner et al. 2008; Wang et al. 2010). Pleiotropic genes involved in multiple
93 biochemical pathways and having numerous interaction partners are subject to evolutionary rate constraint
94 due to purifying selection (Promislow 2004; He and Zhang 2006; Pavlicev and Wagner 2012), particularly
95 in highly constrained hub nodes. Hibernation genes located in intermediate nodes showed higher constraint
96 than those in peripheral nodes, although this trend did not reach statistical significance (Supplementary
97 Table S5). However, in general, genes with the highest number of connections are represented by
98 intermediate node positions within the network, not by hub node positions (Wollenberg Valero 2020, Figure
99 1). The combination of relatively lower constraint and higher complexity in nodes intermediate in the PPI
100 can result in higher rates of accumulation of mutations. These mutations, if beneficial, can in turn have a
101 greater phenotypic effect compared to nodes at the periphery, which accumulate adaptations with least
102 constraint but also with lower phenotypic effects due to a lower degree of pleiotropic interactions
103 (Wollenberg Valero 2020). This pattern is mirrored by genes decelerating in response to hibernation being
104 under lower constraint. Consequently, nodes with intermediate position and rate constraint are good
105 candidates for the emergence of novel and rapid phenotypic adaptations. Additional support for the
106 intermediate network of hibernation-relevant genes not being just an outcome of chance, comes from the
107 fact that their ASPL, NC, BC and Phylop values were significantly different from 1000 random draws
108 (Supplementary Tables S7-S10).

Using tools like STRING V12.0, one can now easily obtain networks by uploading user-generated genomes or accessing over 10,000 existing organismal networks (Szklarczyk et al. 2023). Meanwhile, inferring base-level constraint still requires phylogenetic sampling and analysis of related genomes. The observed correlation between constraint and network position in model organisms such as yeast and humans indicates that network data might adequately predict constraint. Thus, networks could serve as proxies to identify genome regions with lower constraint and high complexity -- those both able to retain mutations accessible to selection, and capable of phenotypic change -- therefore allowing to test hypotheses about the evolvability of specific genes or gene clusters.

Conclusions

The predictability of node constraint based on network topology previously shown for yeast and demonstrated here for mammals could be a universal property of genomic networks, which opens up the possibility of using network properties as a predictive tool for identifying genomic regions with high constraint and potential for adaptation. This potentially has wide-ranging applications, including the prediction of disease heritability, phenotypic evolutionary constraint (Christmas et al. 2023), and even the identification of genomic regions where organisms may or may not respond to rapid climate change (Wollenberg Valero et al. 2021). Such predictive capabilities would be particularly valuable for animal groups or scenarios where empirical classification of evolutionary genomic constraint values is still lacking.

Methods

A human Protein-Protein Interactome was obtained via NDex from the STRING database and loaded into Cytoscape (V3.10). This “high confidence” network contained 17,185 nodes and 420,534 edges and was limited to edges with confidence scores >0.7 to exclude edges with spurious associations (V12.0, (Szklarczyk et al. 2017; Szklarczyk et al. 2023). The median confidence score of this network was 0.903, which is a summary measure for types of evidence supporting a protein-protein interaction, benchmarked

against common presence in KEGG pathways (Szklarczyk et al. 2023). Median number of types of evidence for each interaction was 19.82. Network statistics ASPL (Average Shortest Path Length), NC (Neighborhood Connectivity), and BC (Betweenness Centrality) were computed for each node of the connected network within Cytoscape as it was previously shown that they can partition a network by topology (Wollenberg Valero 2020). Subsequently, the Supplementary table from (Christmas et al. 2023) containing the constraint metric mean Phylop was merged to the node table to enable comparisons of human-derived network statistics with mammalian constraint measures per gene. Following simulation of additional training data with Synthetic Minority Oversampling TEchnique (SMOTE, Siriseriwan 2022), Support Vector Machine based classification (Meyer et al. 2023) was performed on nodes with highest values for these categories pre-classified (10% for ASPL and NC, and 1% for BC due to relatively lower amount of hub nodes in networks, Lawyer 2015). 1000 replicates of 9-fold cross-validation were performed on training sets and stratified training sets, to exclude the possibility of training set imbalance affecting classification. The full training set was then used to classify all remaining nodes into three categories (H - hub nodes, I - intermediate nodes, and P - peripheral nodes as described in Wollenberg Valero 2020). Classifications with low decision support values were subsequently declassified, denoted as “uncertain”, and excluded from further analysis. The response variable mean Phylop was normalized, and a linear model was run assuming to analyze the effect of ASPL, BC, and NC on mean Phylop with Cohen’s f for effect size estimation. The same analysis was repeated for node categories followed by pairwise *post hoc* tests and effect size estimation. Secondly, genes with significantly accelerated or decelerated evolutionary rates (rho, ρ), in conjunction with the emergence of the hibernation phenotype, were extracted from the dataset of Christmas et al. (Christmas et al. 2023). Ten faster evolving and eight slower evolving were matched in this network (Supplementary Table S3). For these genes, node position was calculated to explore the topology of hibernation relevant genes and their mean Phylop scores. A linear model was fitted to test for the effect of node category and direction of association with evolutionary rate on mean Phylop, followed by *post hoc* tests and effect size estimation. Graphs were generated in Cytoscape, as well as R. More details

on supplementary methods and supplementary results are available in Supplementary File 1, the R code can be accessed in Supplementary File 2, and the dataset in Supplementary File 3.

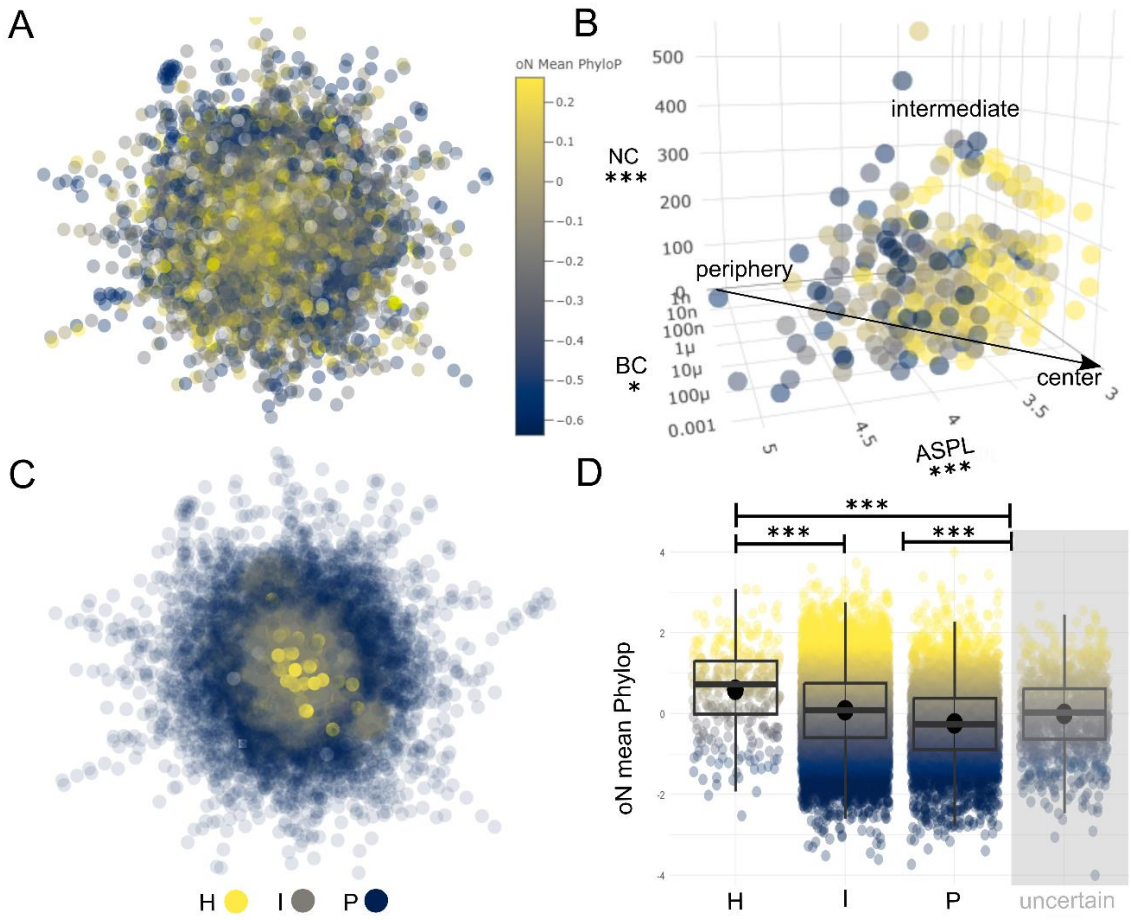

Fig. 1. Visualization of the network topology of mammalian evolutionary constraint. (A) The constraint metric mean Phylop computed by Christmas et al. (Christmas et al. 2023) is visualized on a high-confidence STRING human protein-protein interactome, showing constraint being higher in the center of the network. (B) Reduced dimensions of mean Phylop plotted for three network statistic metrics ASPL, BC, and NC, representing node positions within network topology. These dimensions characterize the center, intermediate and peripheral regions of the network. (C) Network with nodes classified into categories H (yellow), I (grey) and P (blue) using ASPL, BC and NC. (D) Mean Phylop significantly changes with network node category, with constraint decreasing from the center to the periphery of the

network. Significance levels for ANOVA (B) and Emmeans *post hoc* tests (D) are indicated with stars: p <0.001 = ***; p<0.01 = **, p<0.05 = *.

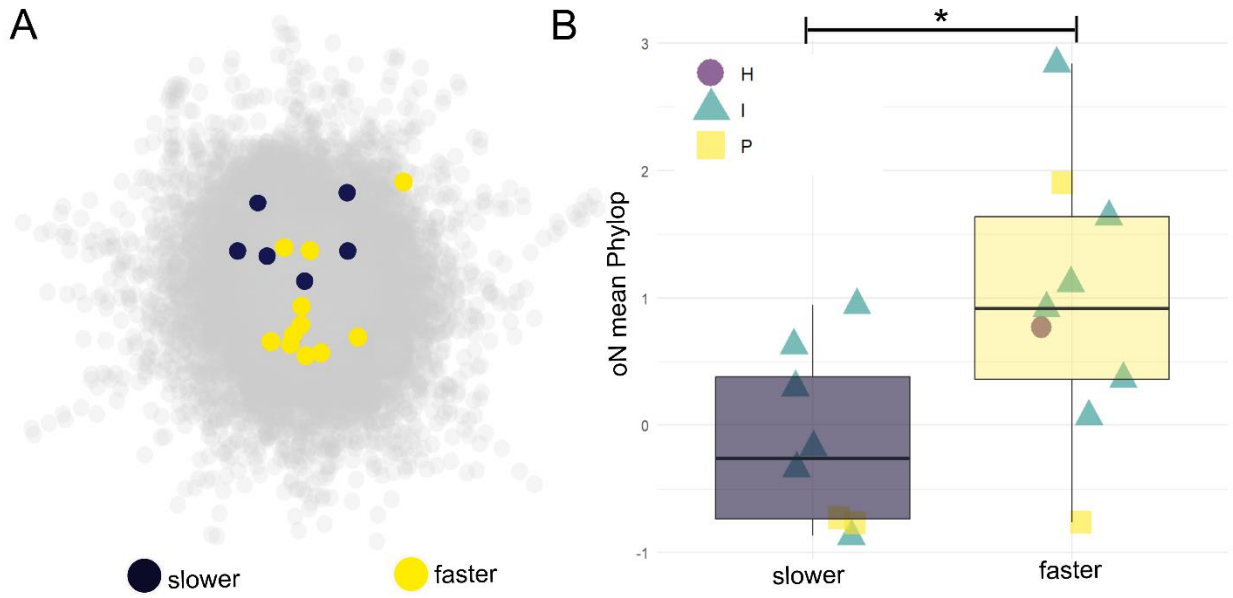

Figure 2. Association of constraint and network position in genes associated with hibernation phenotype. (A) Genes having significantly accelerated (yellow) or decelerated (blue) evolutionary rates in association with the evolution of the hibernating phenotype from Christmas et al. (Christmas et al. 2023), within the human high-confidence STRING PPI. (B) Association between the direction of evolution of hibernation-associated genes, their node class (purple circle- H, teal triangle- I, yellow square - P) , and rate constraint in mammals (orderNormalized mean Phylop). Boxes show medians and quartiles. Significance levels for ANOVA are indicated with stars p <0.001 = ***; p<0.01 = **, p<0.05 = *.

References

Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, Alföldi J, Watts NA, Vittal C, Gauthier LD, et al. 2023. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* 625:92–100.

Christmas MJ, Kaplow IM, Genereux DP, Dong MX, Hughes GM, Li X, Sullivan PF, Hindle AG, Andrews G, Armstrong JC, et al. 2023. Evolutionary constraint and innovation across hundreds of placental mammals. *Science* 380:eabn3943.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61:717–726.

Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH, Xie D, Chen G, Guo C, Faircloth BC, et al. 2020. Dense sampling of bird diversity increases power of comparative genomics. *Nature* 587:252–257.

Frost A, Elgort MG, Brandman O, Ives C, Collins SR, Miller-Vedam L, Weibezahn J, Hein MY, Poser I, Mann M, et al. 2012. Functional repurposing revealed by comparing S. pombe and S. cerevisiae genetic interactions. *Cell* 149:1339–1352.

Futuyma DJ. 2010. Evolutionary constraint and ecological consequences. *Evolution* 64:1865–1884.

Garcia-Porta J, Irisarri I, Kirchner M, Rodríguez A, Kirchhof S, Brown JL, MacLeod A, Turner AP, Ahmadzadeh F, Albaladejo G, et al. 2019. Environmental temperatures shape thermal physiology as well as diversification and genome-wide substitution rates in lizards. *Nat. Commun.* 10:4077.

He X, Zhang J. 2006. Toward a molecular understanding of pleiotropy. *Genetics* 173:1885–1891.

Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* 12:41–51.

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. 2007. Human genome ultraconserved elements are ultraselected. *Science* 317:915.

Lawyer G. 2015. Understanding the influence of all nodes in a network. *Sci. Rep.* 5:8665.

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C-C, Lin C-C. 2023. Package "e1071." Available from: https://cran.r-project.org/web/packages/e1071/e1071.pdf

Pavlicev M, Wagner GP. 2012. A model of developmental evolution: selection, pleiotropy and compensation. *Trends Ecol. Evol.* 27:316–322.

Payne JL, Wagner A. 2019. The causes of evolvability and their evolution. *Nat. Rev. Genet.* 20:24–38.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20:110–121.

Promislow DEL. 2004. Protein networks, pleiotropy and the evolution of senescence. *Proc. Biol. Sci.* 271:1225–1234.

Schoenrock A, Burnside D, Moteshareie H, Pitre S, Hooshyar M, Green JR, Golshani A, Dehne F, Wong A. 2017. Evolution of protein-protein interaction networks in yeast. *PLoS One* 12:e0171920.

Siriseriwan W. 2022. Package "smotefamily." Available from: https://cran.r-project.org/web/packages/smotefamily/smotefamily.pdf

Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, et al. 2023. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* 51:D638–D646.

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, et al. 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45:D362–D368.

Tollis M, Hutchins ED, Stapley J, Rupp SM, Eckalbar WL, Maayan I, Lasku E, Infante CR, Dennis SR, Robertson JA, et al. 2018. Comparative genomics reveals accelerated evolution in conserved pathways during the diversification of anole lizards. *Genome Biol. Evol.* 10:489–506.

Tollis M, Robbins J, Webb AE, Kuderna LFK, Caulin AF, Garcia JD, Bèrubè M, Pourmand N, Marques-Bonet T, O'Connell MJ, et al. 2019. Return to the Sea, Get Huge, Beat Cancer: An Analysis of Cetacean Genomes Including an Assembly for the Humpback Whale (Megaptera novaeangliae). *Mol. Biol. Evol.* 36:1746–1763.

Wagner GP, Kenney-Hunt JP, Pavlicev M, Peck JR, Waxman D, Cheverud JM. 2008. Pleiotropic scaling of gene effects and the "cost of complexity." *Nature* 452:470.

Wang Z, Liao B-Y, Zhang J. 2010. Genomic patterns of pleiotropy and the evolution of complexity. *Proc. Natl. Acad. Sci. U. S. A.* 107:18034–18039.

Wollenberg Valero KC. 2020. Aligning functional network constraint to evolutionary outcomes. *BMC Evol. Biol.* 20:58.

Wollenberg Valero KC, Garcia-Porta J, Irisarri I, Feugere L, Bates A, Kirchhof S, Jovanović Glavaš O, Pafilis P, Samuel SF, Müller J, et al. 2021. Functional genomics of abiotic environmental adaptation in lacertid lizards and other vertebrates. *J. Anim. Ecol.* 91:1049–1326.

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* 346:1311–1320.

Declarations

Data Availability Statement: Data are available publicly, and provided in the Supplementary Materials (S1 - methods and results, S2 - R code, S3 - analysis dataset)

Ethics approval and consent to participate Not applicable

Consent for publication Not applicable

Availability of data and materials All data are available in the main text, supplementary materials, or freely accessible databases and datasets.

Competing interests Authors declare that they have no competing interests

Funding European Research Council, MolStressH2O, 101044202 (KWV

Authors' contributions Conceptualization: KWV, Methodology: KWV, Writing KWV

Acknowledgements KWV acknowledges funding by the European Union (ERC, MolStressH2O, 101044202). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Authors' information (optional) Not applicable

274