TITLE: Life finds a way: reconstructing the evolutionary history of a functionally diverse gene family reveals complexity at the genetic origins of novelty

Authors:

Ivan Koludarov[1*], Timothy NW Jackson[2*], Vivek Suranse[3], Andrea Pozzi[1], Kartik Sunagar[3], Alexander S Mikheyev[1,4]

1. Ecology and Evolution Unit, Okinawa Institute of Science and Technology, Okinawa, Japan
2. Australian Venom Research Unit, Department of Pharmacology and Therapeutics, University of Melbourne, Melbourne, Australia
3. Evolutionary Venomics Lab, Centre for Ecological Sciences, Indian Institute of Science, Bangalore, India
4. Evolutionary Genomics Research Group, Australian National University, Canberra, Australia

*These authors contributed equally.

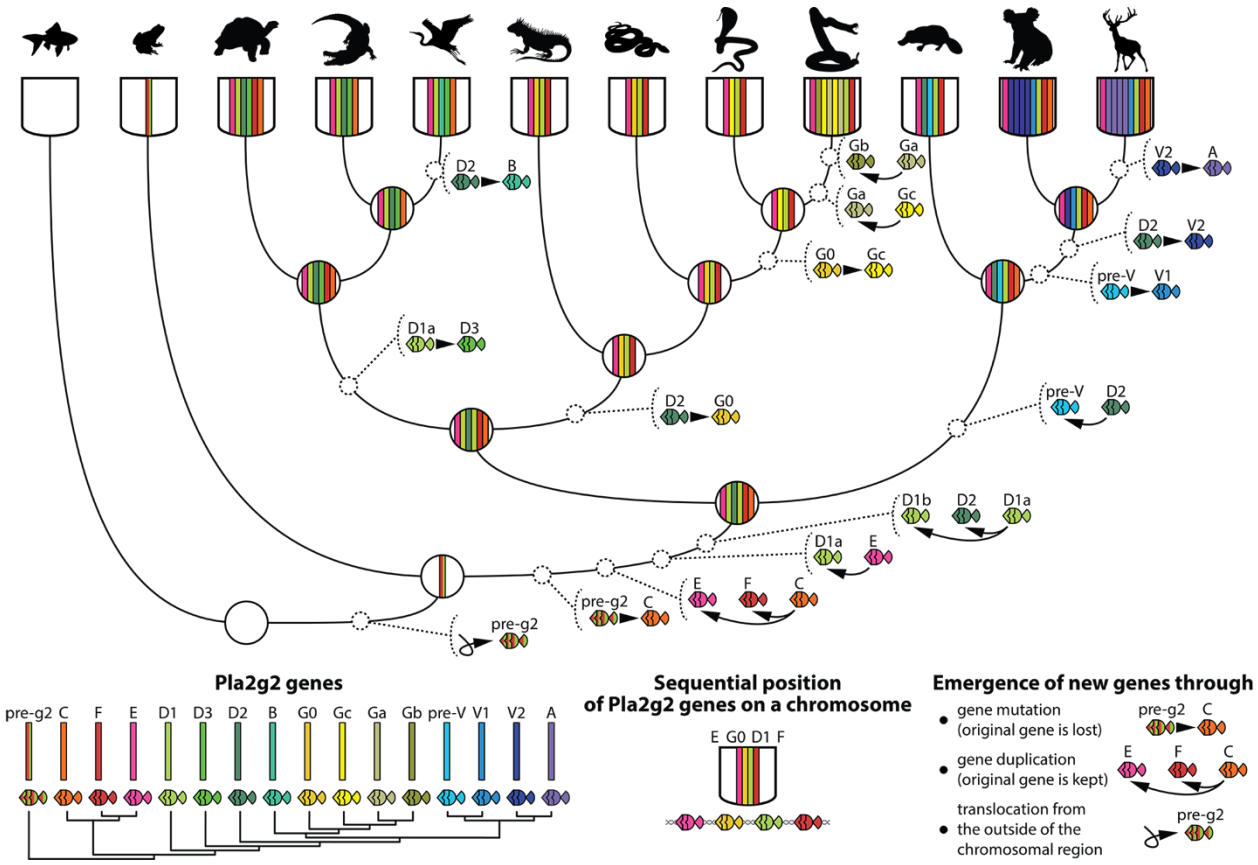
Corresponding authors:

Ivan Koludarov (jcoludar@gmail.com), Timothy NW Jackson (timothy.jackson@unimelb.edu.au), Alexander S Mikheyev (sasha@homologo.us)

Abstract

Gene duplication is associated with the evolution of many novel biological functions at the molecular level. The dominant view, often referred to as “neofunctionalization”, is that duplications precede many novel gene functions by creating functionally redundant copies which are less constrained than singletons. However, numerous alternative models have been formulated, including some in which novel functions emerge prior to duplication. Unfortunately, few studies have reconstructed the evolutionary history of a functionally diverse gene family sufficiently well to differentiate between these models. In order to understand how gene families evolve and to what extent they fit particular evolutionary models, here we examined the evolution of the g2 family of phospholipase A2 (EC 3.1.1.4) in the genomes of 93 species from all major lineages of Vertebrata. This family is evolutionarily important and has been co-opted for a diverse range of functions, including innate immunity and venom. The genomic region in which this family is located is remarkably syntenic. This allowed us to reconstruct all duplication events over hundreds of millions of years of evolutionary history using manual annotation of gene clusters, which enabled the discovery of a large number of previously un-annotated genes. Surprisingly, we found that even at this level of resolution our data could not be unambiguously fit to existing models of gene family evolution. This suggests that each model may describe a part-truth that doesn’t capture the full complexity of gene family evolution.

Graphical abstract



Introduction

Perhaps the most important goal in evolutionary biology remains the explanation of the origins of novelty – how do new functions, traits, and ultimately organisms arise? Gene duplication is widely considered to be one of the most important mechanisms facilitating the evolution of novel functions (Ohno 1970; Innan and Kondrashov 2010). However, duplication itself is often treated as a “black box” – a form of “random” mutation – and numerous apparently contradictory models have been articulated to explain the fates of duplicate genes (Conant and Wolfe 2008; Innan and Kondrashov 2010). Although discussion of duplication and redundancy arguably goes back to Darwin (who had no knowledge of genes and spoke of redundant “organs”) and has a rich history in the 20th Century (Taylor and Raes 2004), the “neofunctionalization” model of Susumu Ohno has loomed large in the field of molecular evolution since the publication of the seminal text “Evolution by Gene Duplication” in 1970 (Ohno 1970). Briefly, this model describes gene duplication (a neutral process) facilitating the genesis of novelty by creating functionally redundant gene copies which, no longer constrained by the functional role of the molecule encoded by the “parent” gene, enjoy a period of relaxed selection in which neutral mutations may accumulate. Any potentially beneficial mutations acquired during this period of neutral change may then be fixed by positive selection.

Despite the continued influence of Ohno's model, a vibrant literature on gene duplication has subsequently produced many other models, a number of which are reviewed in (Innan and Kondrashov 2010), which either expand upon or contradict the basic neofunctionalization framework. A number of these attempt to account for what has been dubbed "Ohno's Dilemma" (Bergthorsson, Andersson, and Roth 2007) – how do duplicate genes survive long enough under neutral conditions to acquire the necessary changes of sequence or expression regulation that result in functional divergence? Several possible fates for duplicates are frequently discussed. One likely outcome is that duplicates are simply deleted, either by further random events or as a direct result of selection stabilizing gene dosage (Bergthorsson, Andersson, and Roth 2007; Birchler and Veitia 2012).

Two primary models describe the fate of duplicates that survive and go on to fulfil functional roles – "subfunctionalization" and "neofunctionalization" (Force et al. 1999; Conant and Wolfe 2008). In the former the parent gene performed multiple functions, which are subsequently distributed between the duplicates, each of which may acquire function-impairing "degenerate" mutations resulting in the necessary maintenance of both copies to fulfil the functional role of the parent gene. In the latter, a novel function is discovered during the period of relaxed constraint immediately following duplication. These models have been further nuanced by the recognition of distinct forms of subfunctionalization such as "escape from adaptive conflict", in which a novel specialized function emerges following the partitioning of the ancestral function; and of neofunctionalization, including models in which duplication is intrinsically advantageous and thus may be positively selected. In these latter models, the initial benefit of duplication may be the consequence of increased gene dosage, increased robustness (protection against deleterious point mutations), or the spontaneous origin of a novel function (Innan and Kondrashov 2010). In all such cases, the consequent accumulation of redundant copies in gene family networks may generate a hotspot for functional novelty.

All of the previously discussed theoretical models give pride-of-place to gene duplication as a facilitator of functional change and thus it may appear as though duplication must precede the origin of novel functions. However, it should also be noted that novel functions may emerge as the result of changes of tissue-specific expression patterns in the absence of duplication, a process known as "gene sharing" (Wistow and Piatigorsky 1987) or "moonlighting" (Copley 2014). Following this period of functional sharing, duplication may facilitate the emergence of distinct proteins capable of subdividing the shared function between them (Force et al. 1999) and specializing for one of the ancestral functions (Hughes 1994). It should be noted that even in subfunctionalization, which does not directly describe the origins of a "novel function" (since the gene in question is already pleiotropic during the period described by the model), the secondary function is "novel" (i.e. originates later in the evolutionary history of the gene) relative to the "original" function. Thus, in these scenarios acquisition of a novel function occurs prior to duplication (Piatigorsky and Wistow 1991).

As recognized by Ohno (Ohno 1970) and supported by much recent research, gene duplication often results in an increase in the dosage of the product encoded by the multiplied genes (e.g. Conant, Birchler, and Pires 2014; Margres et al. 2017). In light of

this, much of the recent literature on gene duplication centers on the importance of gene dosage in determining the fate of duplicates. A key observation in this regard is the divergent fates of duplicates that originate in whole genome duplication (WGD) events and those that are locally (segmentally) duplicated (LD) (Birchler and Veitia 2012; Conant, Birchler, and Pires 2014). In the case of WGD, preserved duplicates are typically those with numerous interaction partners with which they must maintain precise stoichiometric balance – if one half of a pair is lost, a dosage imbalance may occur. Conversely, duplicates preserved after LD tend to be genes with few interaction partners – they can persist in the genome because their origin does not cause a dosage imbalance.

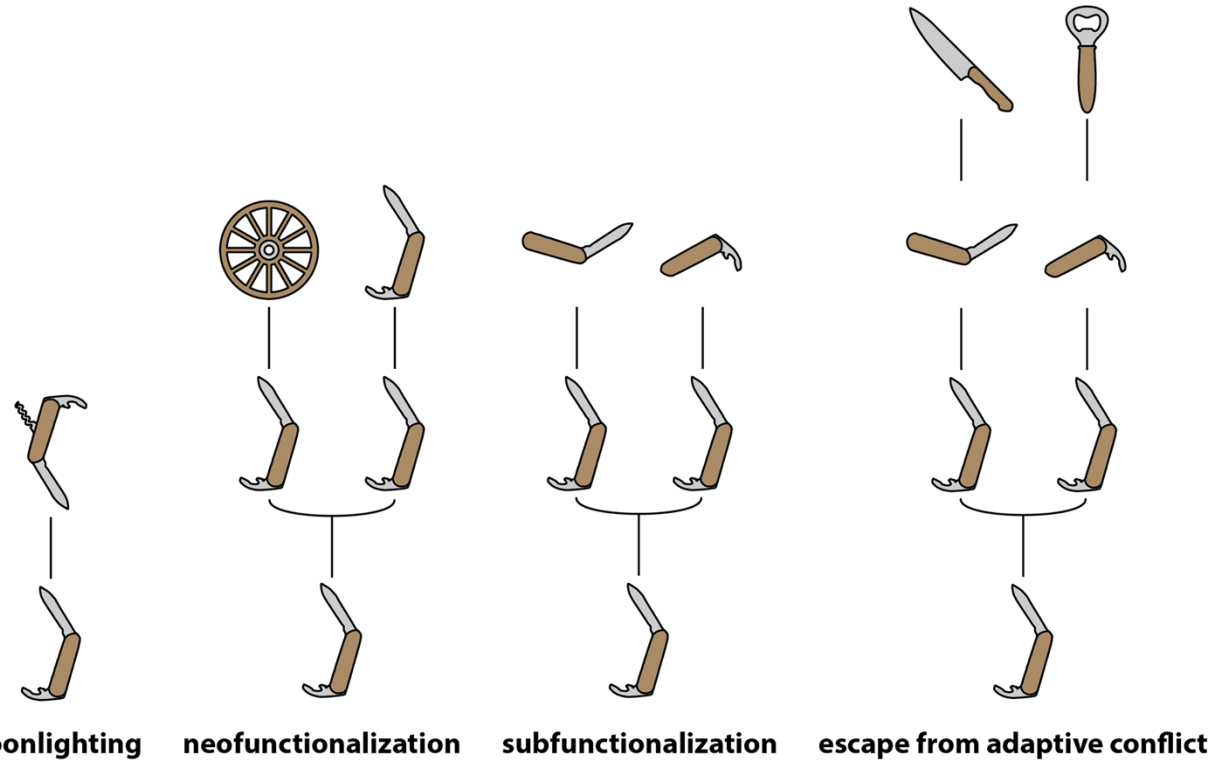


Fig. 1. Models of gene evolution. Schematic representation of the most prominent models of gene evolution, with genes and their functions being represented as human tools. Note that moonlighting occurs in the absence of any duplication, with the product of a single gene fulfilling multiple functional roles.

Virtually all studies of gene families have taken a comparative approach focusing on statistical patterns such as copy number variation (CNV) over time. This powerful approach is general and can facilitate the development of evolutionary models to describe the observed patterns, as well as test alternate hypotheses. However, few ancient gene families have been reconstructed in sufficient detail to validate these proposed models. The few exceptions that have been studied, such as the Hox family, are unusual cases, and hard to generalize. As a result, it is not clear whether the models of gene family evolution derived from these studies fit global patterns. Therefore there is a pressing need to reconstruct gene families in detail, a task that is difficult due to (a) breaks in genomic synteny over large timescales and (b) challenges in assigning orthology within a family after multiple rounds of duplication. Furthermore, understanding the evolution of gene families that have undergone neofunctionalization

and positive selection is particularly important as they underlie the origins of phenotypic novelty.

One such family is Phospholipase A2 group 2 (Pla2g2) – a family of enzymes with multiple interaction partners that exhibits CNV in vertebrate genomes. Pla2g2 are additionally interesting in that they exhibit structural variation and possess multiple functional roles resulting from independent neofunctionalization events in divergent vertebrate lineages – mammalian Pla2g2A is an important component of innate immunity (Nevalainen 2007; Nevalainen 2008; Birts, Barton, and Wilton 2010), whilst Pla2g2G are a major component of the venom of viperid snakes (Kini 2003) (Fig1). Pla2g2 is also an excellent candidate for comparative genomics research because the cluster is located in a region known to be syntenic across vertebrate genomes – in all these genomes, the region of interest is flanked by OTUD3 and UBXN10 genes (Yamaguchi et al. 2014; Dowell et al. 2016). Although Pla2g2 exists in only 5-6 copies in the genomes of many species, in others the family has undergone considerable expansion associated with the acquisition of novel functions. Notably, the functions associated with gene family expansion are extracellular and “exochemical” – directed towards interaction partners originating outside the body of the producing organism. These functions are likely facilitated (“exapted” – Gould and Vrba 1982) by the ancestral activity of the gene family, hydrolyzing membrane phospholipids, which makes them suitable for deployment as a “weapon” against the cells of other organisms, be they microbes (use in innate immunity) or potential prey/predators (use in venom) (Burke and Dennis 2009).

In recent years the availability of genomic data has increased dramatically. However, our ability to process and interpret this information is yet to catch up with our ability to generate it. Genomic annotation pipelines are still at their best when used to annotate genomes of model organisms or species closely related to them (Wang et al. 2017; Yandell and Ence 2012). Since most organisms don’t fall into this category, much *ab initio* annotation remains challenging.

A potential solution to this problem of *ab initio* annotation is to utilize transcriptomic and proteomic data as a template upon which genomic annotations can be based. A weakness of these homology-based annotation pipelines is that they may attempt to align an entire protein or mRNA sequence with a genomic sequence, a practice that is likely to miss alternative splicing variants or pseudogenes, information concerning which is crucial in evolutionary studies (Danchin et al. 2018; Zhang et al. 2018). In addition, these pipelines have trouble annotating tandem-array duplications, especially when there’s high similarity between the copies (Zallot et al. 2016; Nobre et al. 2016). Assembly of transcriptome libraries is itself challenging and erroneous assemblies may introduce errors that are then perpetuated in genomic annotations based upon them (Venturini et al., 2018).

To address the aforementioned concerns we implemented simple and robust method that aligns genomic sequences with mature exons instead of mRNA, creating an exonic map that then can be further translated into coding sequences (see Materials and Methods and Supplementary Materials for detailed explanation). This method was used to successfully locate and annotate several gene families albeit at far less ambitious scale: vertebrate NAD glycohydrolases (Koludarov and Aird, 2019), mammalian

kallikreins (Casewell et al. 2019) and Indian cobra three finger toxins (Suryamohan et al. 2020).

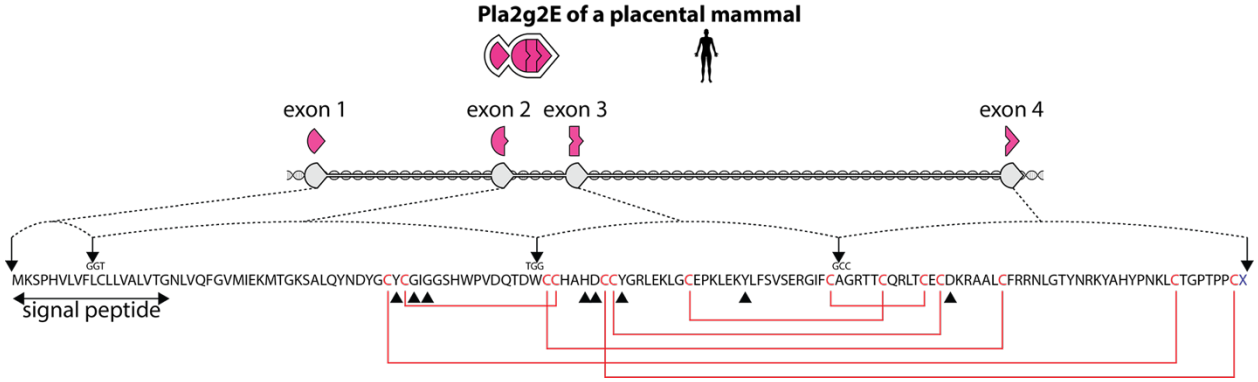
Comparative genomics is increasingly being heralded as one of the most promising frameworks for understanding the origins of novel functional traits at the molecular level (Drukewitz and von Reumont 2019) as well as organismal evolution and diversification more broadly (Zhang et al. 2014; Dunn and Munro 2016). In the present study, we utilized a comparative approach combining manual genomic annotation, phylogenetics, selection rate estimates, and analysis of synteny, to reconstruct the evolutionary history of the Pla2g2 gene family. In the process many new genetic lineages were discovered, necessitating an appraisal and modification of the nomenclature of subclades within this family (see Materials and Methods for the detail). By examining 110 genomic sequences from 93 species across the animal kingdom, we were able to track each duplication event that has occurred since the most recent common ancestor (MRCA) of amniotes, nearly 324 Million Years Ago (Hedges and Kumar 2009). Our analyses identified a number of multiplication events in the gene family's history, including perhaps the most consequential one, which occurred after the split of Amphibia from Amniota and created the g2 cluster. All extant Pla2g2 genes result from this event. In addition, we demonstrate that a single locus, the same in all lineages, was independently involved in all subsequent acquisitions of novel functionality within the family – birds, snakes and mammals derive proteins with novel functions from the same ancestral gene. Our results indicate that no single model amongst those popularly invoked to explain the evolutionary role of gene duplication effectively describes the history of this gene family. We therefore include a discussion of these hypotheses and concur with prior assertions (e.g. Conant, Birchler, and Pires 2014) that a "pluralistic framework" incorporating multiple models may be the most appropriate way to understand the fate of duplicate genes and the origins of functional novelty in molecular evolution.

Results and Discussion

Structure of Pla2g2 genes and proteins

A typical Pla2g2 gene has 4 exons, the first of which encodes half of the signal peptide, while the three other exons encode what becomes the mature protein. In all figures, we have graphically represented this structure by styling the first exon (signal peptide) as the "tail" and the following three exons (mature protein) as the "body" of each gene (Fig. 2). Our analyses were unable to recover the first exon of some genes and other genes have multiple exons encoding the signal peptide or transmembrane domain (e.g. mammalian g2F – see Fig. 3 for a brief on clades of Pla2g2 and their functions). However, in all the cases the three exons that encode the mature peptide retain the same organization. The number of cysteines (and therefore disulfide bonds) varies from 12 to 15 among the genes, but the most common (and likely plesiotypic) condition is 14

214 (see SM1 Fig. 13 for consensus primary structures and SM6 for individual sequences).



215

216 **Fig. 2.** Structure of a typical Pla2g2 gene and the way it corresponds to protein sequence, using *Homo*
217 *sapiens* g2E as an example (intron lengths are up to scale, cysteines and disulfide bonds are in red,
218 triangles mark catalytic sites, arrows mark the position of the splice sites with respect to codons).

219 *Pla2g2 gene cluster synteny is conserved in amniotes*

220 In the present study, we used previously published genomes (see SM2 for the full list)
221 and manual re-annotations to examine the genomic region in which the Pla2g2 gene
222 cluster is located. We discovered remarkable synteny in this region: upstream and
223 downstream regions flanking the Pla2g2 cluster share more than ten genes in almost
224 exactly similar positions across the entire Tetrapoda clade (Fig. 4 and SM1 Fig. 2 and
225 SM1 Fig. 3). This allowed us to reconstruct duplication events spanning 300 million
226 years of the family's evolutionary history. Interestingly, against this background of
227 conservation, several unrelated species (e.g. *Gecko japonicus, Pelodiscus sinensis*)
228 exhibit substantial rearrangements in this region. In addition to species-specific
229 rearrangements, the most prominent long-term rearrangement is shared by all
230 squamate reptiles (lizards and snakes), making them more divergent from crocodylians
231 in this region than crocodylians are from humans. Thus, phylogenetic distance is not
232 necessarily an accurate predictor of syntenic conservation in this genomic region.

233 *Pla2g2 gene clades*

234 All analyzed genes, except those of amphibians, fall into two major clades that diverged
235 from one another some time following the split of Amniota from the amphibians and
236 prior to the evolution of the inferred most recent common ancestor (MRCA) of extant
237 amniotes. These two clades contain Pla2g2s E, F & C; and Pla2g2 D, respectively.
238 Members of the "EFC clade" occupy the flanks of the cluster with g2E genes positioned
239 next to OTUD3 and g2F or g2C positioned next to UBXN10 (Fig. 4). Members of the
240 EFC clade are single-copy genes – E is ubiquitously present, whereas only mammals
241 have functional forms of both F & C (with a few exceptions, e.g. human, in which C is
242 pseudogenized). Birds and crocodylians typically retain C, whereas squamate reptiles
243 retain F. Turtles retain a functional copy of F, but C is typically pseudogenized (see SM1
244 Fig. 24 for gene maps). The mammalian arrangement, in which both genes are typically
245 functional, likely results from the gene conversion of F from a secretory protein into a
246 transmembrane protein, a novel function unique to this lineage (Thul et al. 2017;
247 Petryszak et al. 2016). Selection analyses indicate that all members of the EFC clade

are evolving under the constraint of negative selection (see SM1 Table 1 and SM1 Table 2).

On the other hand, members of the “D clade” occupy the center of the cluster and are involved in all subsequent expansion and neofunctionalization events within the Pla2g2 family (Fig. 4). Both viperid snake g2G toxins and mammalian antimicrobial g2A derive from within this clade.

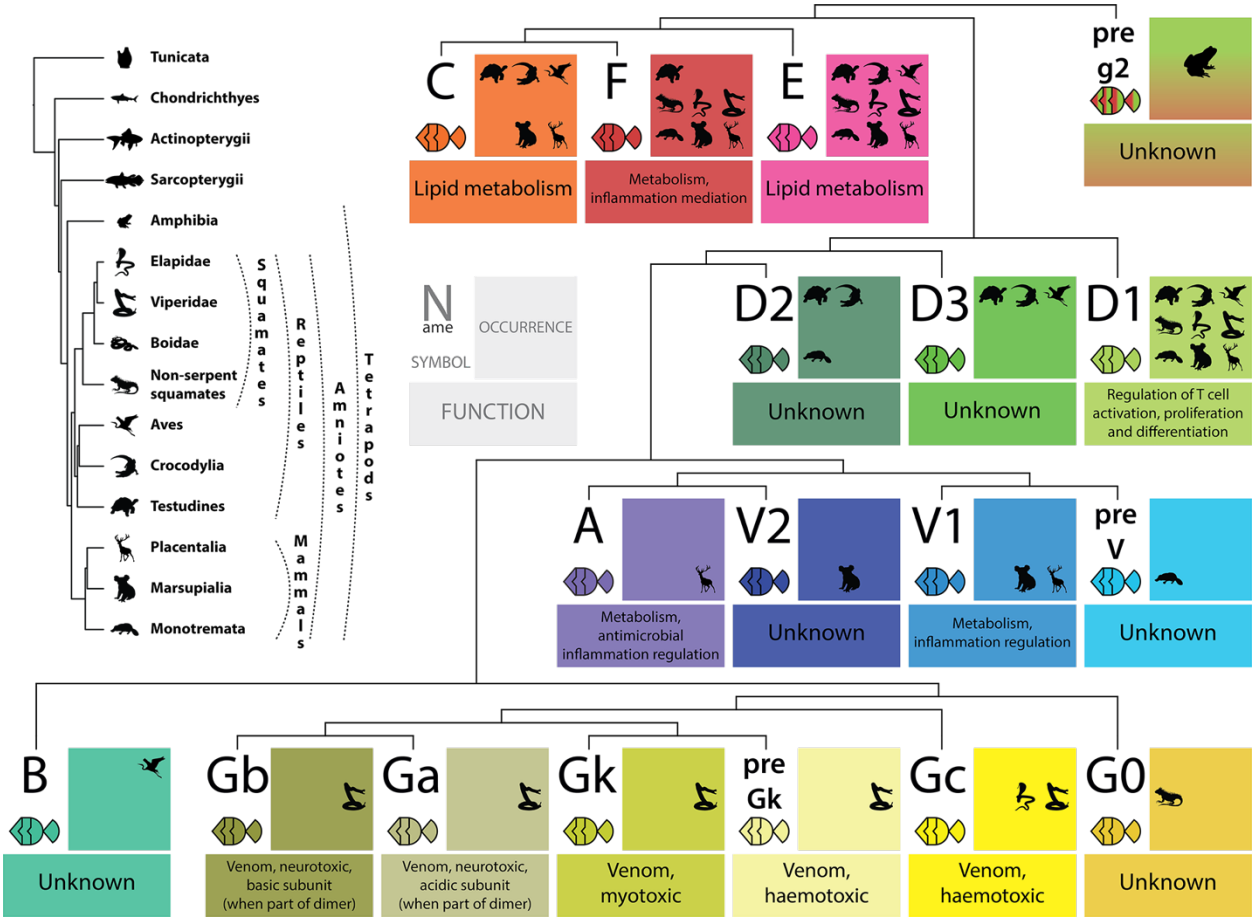


Fig. 3. Phylogenetic tree of taxa surveyed in this study and key to color-coding of Pla2g2 clades, along with their phylogenetic relationships, functions and taxonomic presence. Functions as in (Petryszak et al. 2016; Thul et al. 2017; Six and Dennis 2000) and (Fry 2015).

In our grouping of genes we have tried to preserve previously published Pla2g2 nomenclature created for individual taxa (mostly mammals and vipers, e.g. Six and Dennis 2000), while at the same time reinterpreting this nomenclature in light of the evolutionary history revealed by our analyses (see below for detailed review of Pla2 nomenclature). We have expanded the definition of each Pla2g2 group to make it as concordant with the gene phylogeny as possible and changed some of the names to reflect this phylogeny where it was deemed necessary (Fig. 3).

Evolution of the Pla2g2 cluster

Selection Analyses. To investigate the role of selection in generating structural diversity within the Pla2g2 family, various Bayesian and maximum-likelihood based site-

and branch-site specific models were employed. When grouped into organismal lineages, site-specific analyses indicated that all Pla2g2 (as well as otoconin-22-like genes in all taxa) are evolving under the influence of strong negative selection. The Bayes Empirical Bayes (BEB) analyses failed to identify positively selected sites in all lineages, except one site identified in the avian lineage, while Mixed Effects Model Evolution (MEME) detected 2-7 episodically diversifying sites in all lineages (see SM1 Table 1 and SM1 Table 2). Although the Fast Unconstrained Bayesian AppRoximation (FUBAR) model detected signatures of pervasive positive selection on certain sites in mammals and squamates, it identified a predominant effect of pervasive negative selection on all lineages (SM1 Table 1). These findings may be a consequence of the crucial role played by Pla2g2 in physiological processes, which constrains the accumulation of non-synonymous substitutions that may perturb function. In branch-site specific analyses, we estimated the ω parameter using the two-ratio model test, which indicated the occurrence of positively selected sites in certain clades. Mammalian g2A and g2V exhibited an ω of 2.42, with 1 positively selected site. When all subclasses of squamate reptile g2G were lumped together, the group exhibited an ω of 2.34, with 5 positively selected sites. When the non-toxic form from non-colubroid squamates (g2G0) was removed from this set, analyses returned a higher ω of 2.75 (with the same 5 positively selected sites identified), indicating that the derived toxin forms experience an increased influence of positive selection in comparison to the plesiomorphic gene (SM1 Table 2). The two-ratio model estimated an ω of 2.17 (with 0 positively selected sites) for the incipient toxin g2Gc, however, this analysis failed the likelihood ratio test (LRT). This suggests that this lineage may have experienced positive selection, although the signature was too weak to be confidently identified. Similarly, this test also failed to detect the signatures of positive selection on avian g2B, and the ancient amniote forms g2C, g2E and g2F (SM1 Table 2). Together with the results of the site-selection models and synteny analyses, this highlights the strong evolutionary constraints experienced by the plesiotypic Pla2 gene clusters. Importantly, the combined results highlight the relationship between duplication and positive selection within this gene family, as it is only clades (e.g. g2A and g2G) in which lineage-specific duplications have occurred that exhibit evidence of positive selection (see further discussion below).

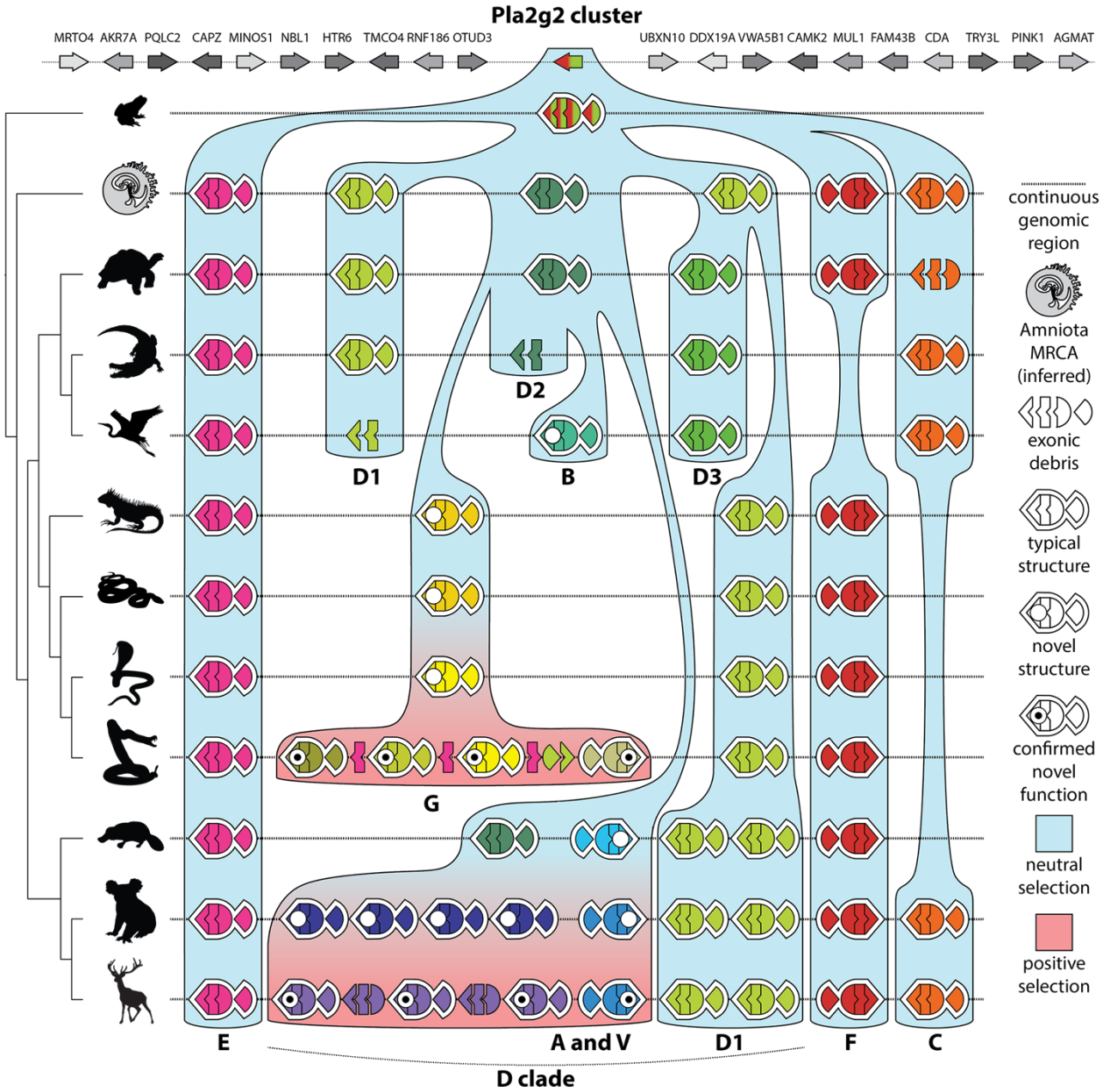


Fig. 4. Location of Pla2g2 cluster, representative arrangement of the cluster for each organismal clade; schematic of the phylogenetic relationships among groups; color-coded estimation of selection regimes (blue = negative/neutral; red = positive). Pla2g2 genes are color-coded by group (see Fig. 3 for details).

Reconstructing the ancestral state. Our results allow us to reconstruct the evolutionary history of the Pla2g2 gene family from its origins in an ancient lobe-finned fish (Sarcopterygii) between 400 and 300 million years ago to its diversification in more recent vertebrate lineages. The deeper we go into the evolutionary past, the more we must rely on inference to guide our reconstruction and thus the less credence our conclusions should be given. Nonetheless, the following scenario is suggested by the evidence we have uncovered: after the split of Teleostei and Sarcopterygii, reshuffling introduced a Pla2 gene into the OTUD3-UBXN10 genomic region. This region underwent duplication, either as a large segment or during the whole genome

313 duplications that, according to the 2R hypothesis, occurred early in vertebrate evolution
314 (Van de Peer, Maere, and Meyer 2010). Subsequent to that, a genomic rearrangement
315 resulted in two regions with one Pla2 each – the ancestral g2 gene and Pla2 "otoconin-
316 22-like" gene, the latter being the closest relative to Pla2g2 occurring outside the
317 OTUD3-UBXN10 region. Each of these genes is present in a single copy in *Xenopus*
318 however we were unable to locate any g2 genes *Rana catesbeiana* genome and found
319 only a single g2-like Pla2 gene in *Ambystoma mexicanum.* In contrast to Pla2 "otoconin-
320 22-like" which is always a single-copy gene (if present at all), g2 persisted and
321 presumably gained functional significance in the Amniota clade – by the time of the
322 inferred amniote MRCA it had undergone ancestral expansion to form a cluster of 5-6
323 genes (Figs 2 & 3).

324 Based on our analysis, the ancestral g2 gene possessed a structure similar to that of
325 modern g2C. This gene may have triplicated via tandem inversion (TID, SM1 Fig. 29).
326 The evidence for that is the close relationship between the sequences of E, F and C
327 genes that flank the cluster; their direct-reversed-direct position characteristic of the
328 results of TID event; and the presence of a short palindromic sequence upstream of the
329 g2 gene in *Xenopus*, which is necessary to facilitate such an event (Reams and Roth
330 2015). Alternatively, the gene may have multiplied via two separate tandem
331 duplications.

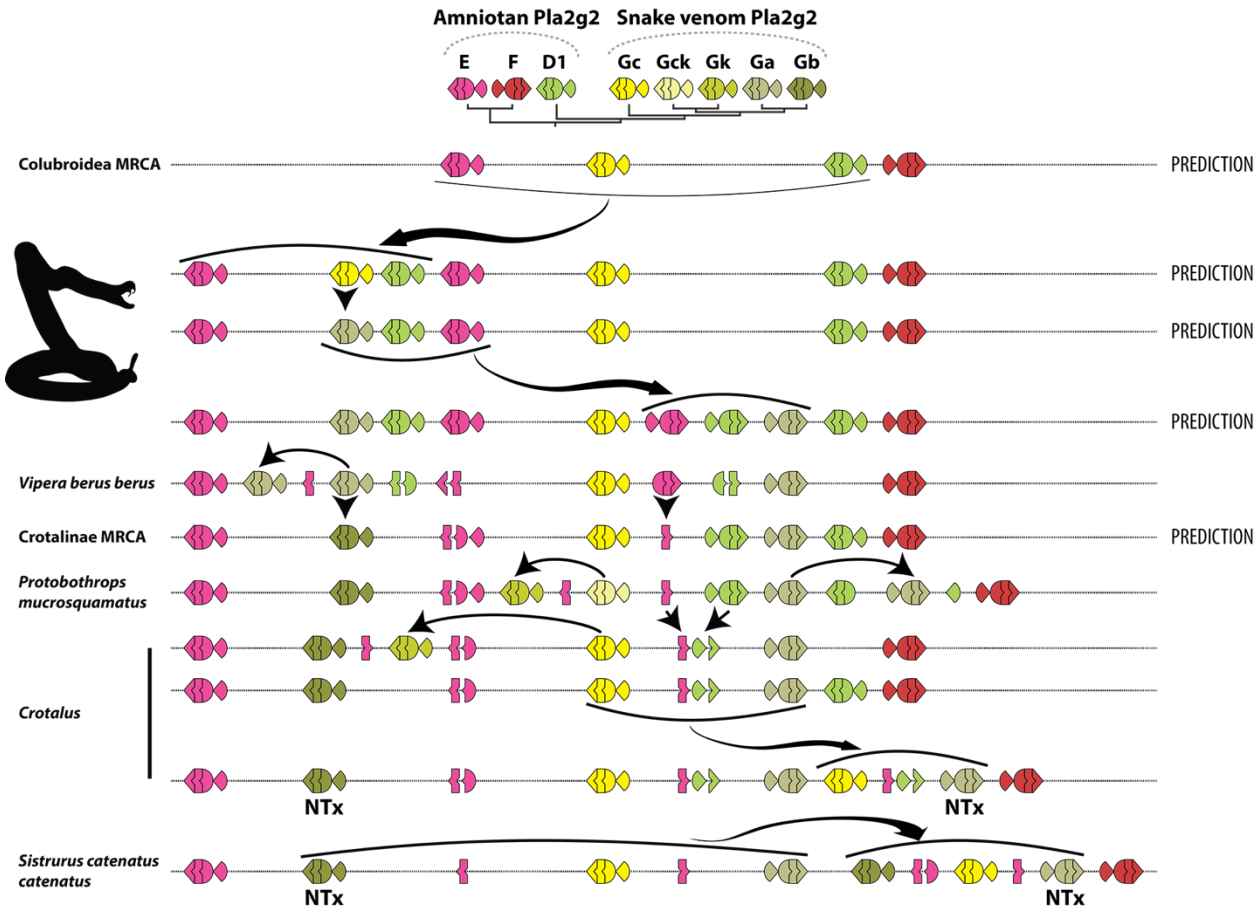
332 As mentioned above, while the g2E gene is present in all species studied, there's a
333 clear taxonomic bias concerning the preservation of the g2F or g2C gene, both of which
334 are present only in the mammals and turtles (with C typically pseudogenized in the
335 latter). In other lineages, sequence similarity between C and F genes may have
336 conferred functional redundancy leading to the elimination of either one or the other.

337 Early in amniote evolution one of the ancestral EFC genes (possibly the g2E of the
338 Amniota MRCA based on its genomic position) duplicated to create the g2D gene,
339 which in turn spawned at least three additional copies (Fig. 4). This is indicated by the
340 fact that all extant lineages have at least one g2D gene and one or two differentiated D-
341 clade genes. Based on our analysis of synteny, it appears that the amniote MRCA had
342 at least three g2D genes sandwiched between g2E on one side and g2F and g2C on
343 the other. We say "at least", because there may have been additional copies. The
344 syntenic position and phylogenetic/structural relationships of extant genes indicate that
345 two of the three were members of the plesiotypic g2D1 clade (see Fig. 4).

346 **Expansion of D-clade genes is associated with lineage-specific**
347 **neofunctionalization.** The major differences in Pla2g2 clusters between different
348 lineages of Amniota concern the evolution of new D-clade genes unique to each
349 taxonomic lineage. All of these genes appear to be descendants of the same ancestral
350 g2D2 gene, which is still present in a plesiomorphic form in crocodiles and turtles (Fig.
351 4, SM1 Fig.24). The ancestral g2D2 appears to have undergone mutation independently
352 in mammals, birds, and squamate reptiles. These mutated derivations of g2D2 are the
353 ancestors of the g2V (mammalian), g2B (avian), and g2G (squamate reptilian) clades,
354 each of which possesses a unique genomic structure. The avian g2B protein has an N-
355 terminal region unique to this group that sets it apart from almost all other g2. It is
356 always present in only a single copy and its function is unknown. The mammalian and
357 squamate forms are virtually the only ones to have undergone expansion since the

358 ancestral amniote duplication. According to selection analyses, both of these are
359 evolving under the influence of positive selection (SM1 Table 1), however, both of them
360 evolved their unique sequences prior to duplication (Fig. 4).

361 *Evolution of snake venom Pla2g2 genes*



362

363

364 **Fig. 5. Evolution of Pla2g2 cluster in Viperidae.** Note the presence of multiple fragments of "exonic
365 debris" (mainly from g2E and g2D) that make possible the reconstruction of the evolutionary history,
366 including all duplication events, of this genomic region in viperid snakes. Arrows indicate birth and death
367 events. A number of the duplication events (bold arrows) do not involve single genes (unlike g2A of
368 mammals) but rather small groups which are duplicated as units ("cassettes"), typically composed of a
369 g2G gene flanked by parts of g2E and g2D. See SM1 Fig. 24 for precise exonic maps of *Crotalus* and
370 other viperid and colubroid snakes omitted here.

371 The squamate reptile g2G gene is present in lizards and *Boa constrictor* in a single
372 copy. This gene is the basal member of the clade that contains genes variously labelled
373 as "Gc", "Gk", "Ga", and "Gb" (Dowell et al. 2016). Since it is the plesiotypic (non-toxic)
374 form, we have labelled it g2G0 (Fig. 3). In the ancestor of "advanced snakes"
375 (Colubroidea – i.e. the shared ancestor of cobras and vipers) it underwent structural
376 change whilst remaining a single copy (becoming g2Gc) and this, in addition to the gene
377 family's ancestral membrane-degrading activity, may have further exapted it for its
378 subsequent functional recruitment into the venom arsenal of vipers. This form, unique to

379 advanced snakes, was recovered from the genomes of the natricid snake *Thermophis*
380 *baileyi,* the elapid snakes *Notechis scutatus*, *Ophiophagus hannah*, and *Pseudonaja*
381 *textilis*, and all viperid species (i.e. all colubroid snakes with genomes of sufficient
382 quality), indicating that it is likely a synapomorphy of colubroid snakes. Readers familiar
383 with snake toxinology should bear in mind that this gene is not the one recruited to the
384 venom arsenal of elapid snakes, which is of a member Pla2 group 1.

385 The colubroid ancestor may have been venomous (Jackson et al. 2017) and thus the
386 potential exists at that early stage for positive selection acting upon genes encoding
387 orally secreted toxins. It is unclear, however, whether g2Gc was in fact utilized as a
388 venom toxin by early colubroid snakes and whether this function may have provided the
389 selection pressure leading to the fixation of this form in the colubroid ancestor.
390 However, the ancestral membrane-degrading activity of Pla2g2 gene products (Six and
391 Dennis 2000) exapts them for utilization as toxins or as antimicrobial components of
392 innate immunity – note that these are not mutually exclusive as immune components
393 are frequently co-opted for use as toxins (e.g. Georgieva et al. 2011; Wong and Belov
394 2012). Not all genes within the Gc group have been functionally characterized at this
395 stage and data concerning their expression in various tissues is limited (Fig. 3); these
396 data are important in resolving the evolutionary pathways leading to the deployment of
397 this gene family in the venom of viperid snakes. Regardless, Gck is selectively
398 expressed in the venom gland of extant crotaline viperid snakes (Aird et al. 2017;
399 Dowell et al. 2018). The homologous gene (Gc) is not expressed in the venom gland or
400 accessory gland of the elapid snake *O. hannah* (despite being 94% similar in sequence
401 to the viperid form SM1 Fig. 32), indicating that it is not utilized as a toxin by this
402 species; it is also expressed at extremely low levels in pooled tissues, which may be
403 indicative of its incipient toxicity (Vonk et al. 2013). Whilst elapid snakes do utilize
404 phospholipases as toxins, all known elapid Pla2 venom toxins are members of group 1,
405 which is unrelated to group 2 (Fry 2015). Group 1 Pla2s exhibit quite a different
406 evolutionary pattern from that of group 2, and are the subject of a follow-up study.

407 Pla2g2Gc is the founder member of the colubroid snake-specific Pla2g2G clade, which
408 all viperid PLa2g2 toxins are members of. Given its presence as a single copy in the
409 genomes of the other colubroid snakes (a clade which is ancestrally venomous), we
410 have uncovered no evidence of duplication associated with the acquisition of a toxic
411 function in venom for this gene. Rather, in viperid snakes Gc, as well as transitional and
412 derived forms are expressed in the venom gland (Aird et al. 2017; Dowell et al. 2018).
413 Thus we infer that this novel function arose *prior* to duplication, possibly via co-option
414 facilitated by a shift in tissue-specific expression patterns which resulted in its
415 expression in the venom gland. Stochastic gene expression of this kind has been linked
416 to the phenotypic diversity from which the origins of novel adaptations may arise (True
417 & Carroll 2002; Kaern et al. 2005; Woods 2014). At present, our ability to pinpoint the
418 origin of the toxin function in Viperidae is limited as we lack transitional forms within that
419 family. Duplication of the gene occurred sometime between the split of viperid snakes
420 from the main stem of Colubroidea and the origin of the MRCA of extant Viperidae,
421 which possessed additional copies (Fig. 5).

422 The alternative possibility is that duplication occurred prior to “recruitment” to the venom
423 system, giving rise to the new gene g2Ga. The protein encoded by this new gene,

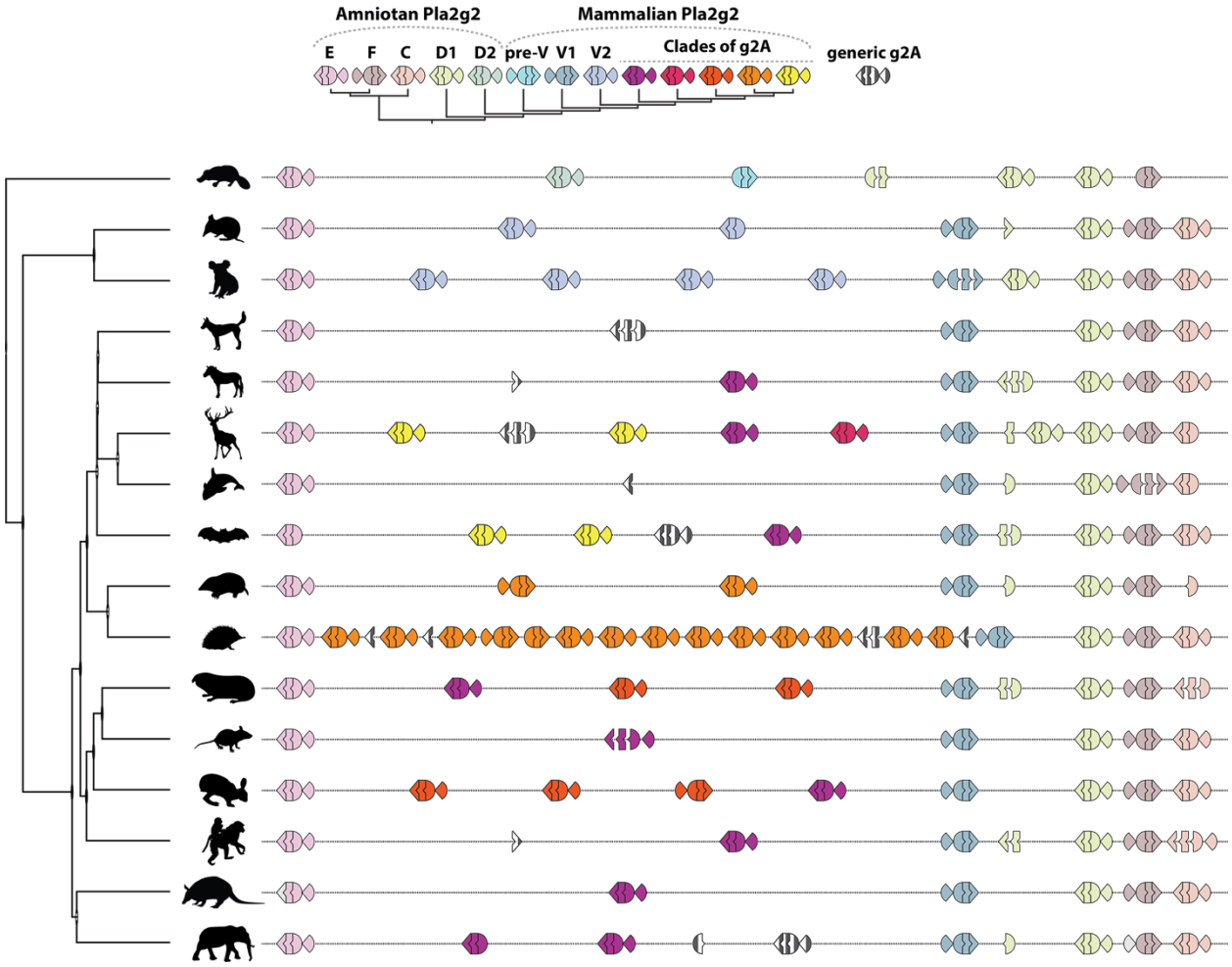
424 possessing by chance a greater toxicity than that of its parent gene (g2Gc), would have
425 been selected for venom gland-specific expression and the parent gene was co-
426 expressed due to the co-regulation of neighboring genes. This alternate scenario is
427 further complicated by the fact that g2Gc, initially a passively co-expressed (unselected)
428 gene in the venom system later evolves (in a Crotalinae specific derivation) into the
429 myotoxic g2Gk (a.k.a. "Lys49 Pla2s"). Thus instead of one, this alternate hypothesis
430 requires three "recruitment" events – one (of g2Ga) for the initial addition to the venom
431 arsenal, a second one associated with the mutation of g2Gc into g2Gk and a third one
432 with g2Gb becoming a basic subunit of neurotoxic dimeric Pla2g2.

433 In either case, changes in gene expression, which are untraceable at this level of
434 analysis (and possibly lost to the sands of time) are crucially important in the initial
435 acquisition of the novel, toxic function. Given the presence of additional "random"
436 (unselected) steps in the latter scenario (duplication precedes novel function), we prefer
437 the former (novel function precedes duplication – see below for a more detailed
438 discussion). However, additional research is required to definitively differentiate
439 between these hypotheses.

440 Viperid snakes diverged early from the main colubroid lineage (which includes elapid
441 snakes; the front-fanged lamprophiids *Atractaspis* and *Homoroselaps*; and many non-
442 front-fanged venomous species) and the most striking synapomorphy of the family is the
443 possession of large, hollow fangs which are the sole tooth located on a mobile maxillary
444 bone (Fry et al. 2012). These fangs, like those of other front-fanged snakes, are
445 connected to the venom gland by an enclosed duct, and the gland itself is surrounded
446 by compressor musculature which contracts during venom delivery. Thus viperids are in
447 possession of a "high-pressure" venom delivery system and, moreover, were the first
448 lineage of snakes in which such a system evolved. That members of the Gc group
449 apparently only became specialized for use as venom toxins after the divergence of
450 Viperidae from other advanced snakes suggests that the acquisition of this function may
451 have been associated with the evolution of a delivery system capable of inoculating the
452 toxin directly into the muscle tissue of potential prey organisms. This hypothesis is
453 consistent with the subsequent diversification of the subfamily in viperid snakes,
454 including the evolution of specialized myotoxic and presynaptically neurotoxic forms.
455 Myotoxic Pla2 are likely to be more effective as toxins if delivered intramuscularly – a
456 feat that non-front-fanged snakes, and even many front-fanged elapid snakes, are
457 unlikely to be capable of. It should be noted, however, that myotoxic phospholipases
458 have been independently recruited (from group 1 Pla2) as toxins in elapid snakes (Fry
459 2015). These toxins are particularly enriched in the venoms of large species capable of
460 exerting considerable bite force (e.g. *Pseudechis australis* – Georgieva et al. 2011) and
461 in species which feed on prey items that lack a layer of subcutaneous fat (e.g. some
462 hydrophiine sea snakes – Gopalakrishnakone 1997; Phillips 2002). It is plausible,
463 therefore, that the ability to inoculate venom intramuscularly has played a role in the
464 recruitment or enrichment of myotoxic Pla2g1 in elapid snakes – co-evolution of toxins
465 and associated delivery anatomy is a reasonable expectation, and has been reported
466 previously in toxicoferan reptiles (e.g. – Fry et al. 2012) and cnidarians (Surm et al.
467 2019). As far as the present case of Pla2g2 in viperid snake evolution is concerned,
468 additional investigation of snake bite force mechanics and feeding ecology is required to
469 test these conjectures.

Subsequent to the acquisition of the toxic function, a series of duplication events expanded this lineage in viperid snakes, the first of which gave rise to two new isoforms – the g2Ga (acidic) and g2Gb (basic) venom Pla2s (Fig. 5). In viper venoms these forms are more abundant than the plesiotypic g2Gc form (Aird et al. 2017; SM3, SM4). Pla2g2G venom genes were duplicated in multiple lineages to produce genes that became subunits of heterodimeric neurotoxins in several *Crotalus* and *Sistrurus* species (French et al. 2004; Mackessy 2010). As revealed by the arrangement and orientation of genes and exonic debris, these neurotoxins arose via independent duplications in each of these two genera (Fig. ; cf. Dowell et al. 2016), an example of convergent evolution made possible by the fact that a single point mutation is all that is required to "unlock cascading exaptations", leading to the derivation of this potent toxin (Whittington, Mason, and Rokyta 2018). In parallel, g2Gc (the plesiotypic form) mutated (again in the absence of duplication) into a g2Gck form in Crotalinae (pit vipers), and an additional duplication of this form became the non-catalytic myotoxin (g2Gk) (Fig. 5). In tandem with the evolution of these derived forms, the plesiotypical Gc and Gck appear to have had their expression suppressed, despite still being present in the genomes of many viperid snakes (Aird et al. 2017; Dowell et al. 2016; 2018).

Mammalian Pla2g2



489 **Fig. 6. Schematic depiction of expansion and differentiation of mammalian Pla2g2A.** Actual
490 complexity of grouping and duplication in placentals is greater, see SM1 Fig. 31 for details. Note that
491 "g2A" forms are confined to placental mammals, although marsupials exhibit distinct patterns of
492 duplication in the same region. Genes and fragments colored green and white belong to the g2A clade,
493 however their relationships within that clade cannot be unambiguously determined. Note that symbols of
494 non-g2A Plag2 forms are partially discolored in comparison to other figures in this study.

495 The g2V clade exhibits a similarly convoluted recent evolutionary history in mammals.
496 An early duplication of g2D produced "pre-g2V" (present in the platypus – Monotremata)
497 and the gene that later became g2V1 (a.k.a. Pla2g5), which is present in both placental
498 (Eutheria) and marsupial (Metatheria) mammals. While g2V1 is always present in a
499 single copy, it likely spawned a form similar to g2V2 of marsupials and then underwent
500 multiple independent expansions (Fig. 4, Fig. 6, Supplementary materials). In
501 marsupials, g2V2-like genes exist in several copies, all of which structurally resemble
502 an intermediate form between g2V1 and g2A of placental mammals, but do not form a
503 single clade. Group 2A – their counterpart in placental mammals – multiplied
504 independently in different lineages, while being retained as a single copy in some
505 placental mammals (Fig. 6). Most variation (in both structure and number of genes)
506 within placental mammals occurs at the center of the g2A sub-cluster, thus
507 recapitulating the general pattern of Pla2g2 evolution in which expansion takes place at
508 a single central locus (SM1 Fig. 31). It is hard to say how many copies of g2A the
509 eutherian MRCA possessed, given the CNV in both marsupials (Metatheria) and
510 placental mammals (Eutheria).

511 Given the aforementioned CNV and the myriad structural forms that have emerged in
512 parallel in distinct taxonomic lineages, group 2A phospholipases appear to be evolving
513 dynamically in placental mammals (Fig. 6). As an extreme example amongst the
514 surveyed taxa, the hedgehog genome contains 14 copies, four of which are successive
515 duplications of a form it shares with moles and shrews. At least one form is shared
516 (deep purple in Fig. 6) amongst (almost) all species that retain a functional g2A gene,
517 suggesting that this gene fulfils an important function. The sole exception to this pattern
518 is the hedgehog, which apparently compensates for the loss of the plesiotypic shared
519 form with its plethora of novel genes. Unfortunately, it is impossible to be precise about
520 the functional roles of products of these genes, given the absence of any functional data
521 or data on tissue-specific expression for any g2A forms from non-model organisms
522 (i.e. >95% of forms). Unlike their single copy g2V1 ancestors, g2A with characterized
523 functions are antimicrobial (Nevalainen, Graham, and Scott 2008). It must be noted that
524 "model" organisms appear anomalous in our data, in that the mouse, human, horse, dog
525 and cat genomes all contain only a single generic form of g2A or lack it entirely. Further
526 research may indicate that distinct forms of g2A should be grouped separately, given
527 that structurally g2A subclades differ as much if not more than venom g2G, and these
528 latter are known to possess strikingly different activities (see discussion above). It is
529 also peculiar that in our dataset only predominantly carnivorous mammals have
530 pseudogenized g2A: the entire Carnivora lineage and *Orcinus orca*. Whether this is
531 functionally significant is an interesting question for future research.

532 More broadly, based on current data (Fig. 3), it is reasonable to hypothesize that the
533 dynamic evolution of the g2A clade represents innovations built upon a plesiotypic
534 antimicrobial function, which is retained to a varying extent by all members of the group

535 functionally characterized to date. The sister clade of g2A in marsupials – g2V2 –
536 seems to evolve in a similar way, however the paucity of high-quality marsupial
537 genomes and a total lack of activity data for the products of these genes prevents us
538 from reaching any definite conclusion. However, it is clear that the general evolutionary
539 trend of g2V is stable throughout the Mammalia: varying numbers of lineage-specific
540 duplications resulting in novel forms that one can imagine are "fine-tuned" by selection
541 for the particular ecological needs of a particular species. This pattern would be
542 analogous to the evolution of odor receptors and toxin genes. This fascinating
543 evolutionary also story deserves further investigation.

544 **Thus, a single ancestral gene, g2D2, evolves into a new structural form**
545 **independently in squamates, birds and mammals.** In all the three cases (g2B, g2G
546 and g2A/V), these genes evolve new protein structures without prior duplication (Fig. 4).
547 Whilst the avian g2B remains in a single copy, in squamate reptiles and mammals an
548 expansion of the group took place. Interestingly, g2A/V and g2G genes are not just the
549 only Pla2g2 genes to multiply, but also the only ones that show evidence of evolving
550 under the influence of positive selection (SM1 Table 2). This pattern suggests that the
551 acquisition of a novel activity, associated with structural change, perhaps in concert with
552 an appropriate pattern of tissue-specific expression, was the change that facilitated the
553 accumulation of duplicates at this locus. Subsequent to the expansion of these gene
554 networks, the locus became a neofunctionalization hotspot, particularly within viperid
555 snakes and placental mammals.

556 *Hypotheses concerning the role of gene duplication in the evolution of novel functions*

557 Our results indicate that duplication is not always necessary for the acquisition of a
558 novel function, a conclusion most clearly supported in the case of the g2Gc gene of
559 colubroid snakes. Whether or not this gene's product was first deployed as a venom
560 toxin in an ancestral colubroid or viperid snake, duplication does not appear to have
561 been a prerequisite for the acquisition of this novel "exochemical" (or "exophysiological",
562 being deployed to work outside of organism's inner "chemistry") function. This is in
563 contrast to the frequent assertion in the toxinology literature that toxin genes typically
564 acquire their function following the duplication of a non-toxin ancestor (e.g. Reza,
565 Swarup, and Kini 2005; Lynch 2007; B. G. Fry et al. 2012).

566 Despite duplication not being proximally involved in the acquisition of the toxic function,
567 it should be noted that the gene recruited was part of a gene cluster, which is likely an
568 important factor influencing its evolutionary trajectory. This is both because the locus
569 clearly possesses an ancestral propensity for duplication, and because the gene's
570 function may have been shared with the sister genes (in case of g2G ancestor,
571 presumably g2D or even g2E/F/C). This redundancy probably decreased the
572 evolutionary constraints on a single gene. Importantly, all of the novel clades of Pla2g2
573 in distinct animal lineages originate from this same locus. On the other hand, the
574 original expansion of the cluster is ancient and the general trend subsequent to this
575 expansion appears to be towards reduction through gene loss, rather than further
576 multiplication.

577 Several studies have downplayed the role of gene duplication in the acquisition of a
578 toxic function by certain gene families in certain venomous lineages. In the platypus, for

example, a majority of toxin-encoding genes do not exhibit lineage-specific expansions (Wong and Belov, 2012), rather toxins appear to have been recruited from gene families with pre-existing CNV, and no expansion has taken place subsequent to recruitment. In parasitoid wasps, duplication appears to have played even less of a role, with the predominant mode of venom diversification within and among lineages being shifts in *cis*-regulated gene expression in the absence of either gene duplication associated with acquisition of a toxin function, or gene deletion associated with its loss (Martinson et al. 2017).

Recent research has demonstrated that duplication of existing toxin genes can drive changes in gene expression, and provide raw material for future evolution (Margres et al. 2016). However, we suggest that novel functions first emerge when a gene product's *context* changes and it is exposed to a novel suite of interaction partners. This is unsurprising, given that a protein's function is fundamentally relational (Guttinger 2018), i.e., defined interdependently as the consequence of interaction between one protein and another. The change of context may occur in multiple ways: following a stochastic change in expression pattern that sees a gene being expressed in a novel tissue, i.e., a tissue in which the gene product in question is not typically expressed (Kaern et al. 2005; Woods 2014); following a structural change that modifies a protein's interactive propensity (i.e. exposes it to a novel context in terms of potential partners for interaction); or following the evolution of a "delivery system" (e.g. long hollow fangs) capable of delivering the gene product into a novel context (e.g. muscle tissue of prey animals).

Such changes of context may lead to the discovery of a "good trick" (Dennett 1995) by fortuitously facilitating an interaction with a positive impact on fitness. If both functions (ancestral and derived) persist for the same gene, this may create pressure for duplication, as the multiple functions (ancestral and derived) of the protein require segregation into discrete genes, a situation similar to that described in the "subfunctionalization" model (Force et al. 1999; Hargreaves et al. 2014).

Based on their expression patterns (Thul et al. 2017; Petryszak et al. 2016) and functions in extant species (Six and Dennis 2000; Fry 2015), Pla2g2 genes may have played an important role in the immune system of early terrestrial vertebrates. In any case, it seems plausible that the ancestral functional role of this group is associated with the independent gene mutation of an ancestral g2D2 gene into derived forms: g2B in birds, g2A in mammals and g2G in squamates. The role of antimicrobial mammalian A-clade genes in the innate immune response is well known (Nevalainen, Graham, and Scott 2008), though of course the majority of forms have not been studied. Amongst squamate G-clade genes, only viperid snake toxins forms are well-studied. And the function of avian g2B is unclear.

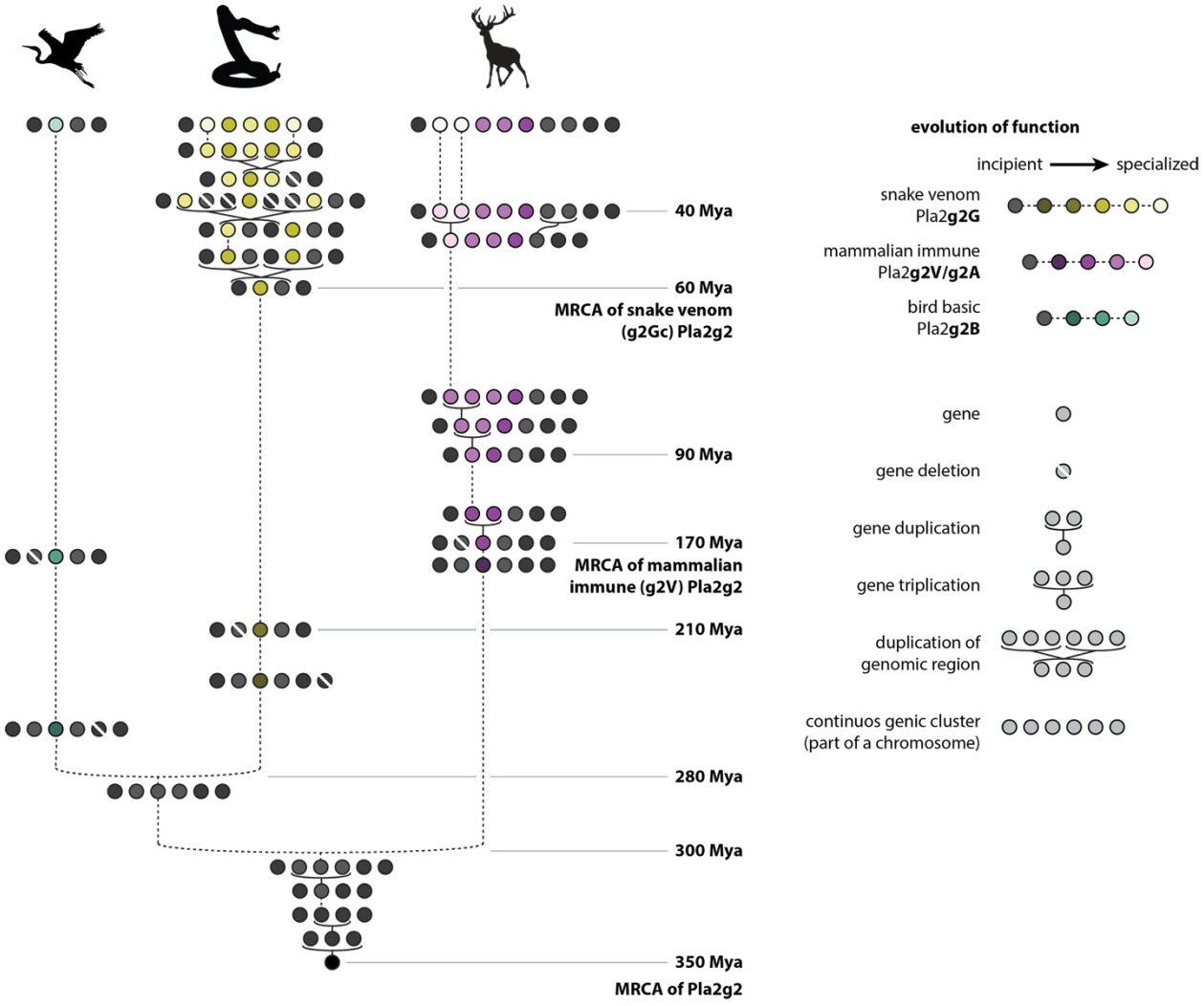


Fig. 7. Schematic depiction of the interplay between the emergence of novel functions and duplication in the evolution of Pla2g2. Time estimates are based on data accumulated in TimeTree.org, for details on functions see Fig. 3, for details of synteny see Fig. 4. The figure is based on *Pantholops hodgsonii* for mammals (the pseudogenized g2A is omitted), *Chlamydotis macqueenii* for birds and *Crotalus scutulatus* (form A) for snakes.

The pattern we have observed (Fig. 7), in which both the emergence of novel functions and subsequent gene family expansion take place at the same locus in distantly related taxa, suggests that such loci have a deep ancestral propensity for mutation and duplication. The propensity for duplication is likely conferred by genomic structure, as particular arrangements of genetic material (e.g. those described above for tandem inversion duplication) facilitate duplication (Reams and Roth 2015). This propensity, however, may typically be constrained. The alternative, still advocated by some biologists (e.g. Dunn and Munro 2016), is that duplications occur randomly throughout the genome and that regions only differ in CNV due to the differential preservation of

632 duplicates. This seems implausible for two reasons: 1) because exonic/intronic debris is
633 typically evident following deletion (unless the deletions are extremely ancient events) –
634 this debris was not detected throughout the genomes in the present study but only, *ex*
635 *hypothesi*, in isolated regions in particular genomes (i.e. those in which birth-and-death
636 is taking place); and 2) because down-regulation ("dosage sharing") or silencing with
637 methylation may facilitate the long-term preservation of segmental duplications in
638 genomes despite the predictions of the dosage balance hypothesis (Assis and Bachtrog
639 2015; Lan and Pritchard 2016; Guschanski, Warnefors, and Kaessmann 2017). Another
640 alternative is that individuals in which deleterious duplications occur are strongly
641 selected against and thus no evidence of these duplications persists in sequenced
642 genomes, but this also seems an unnecessarily extreme speculation as it requires that
643 such duplications be invariably lethal or render organisms sterile – i.e. individuals in
644 which duplication takes place must produce no offspring. We note that it is far from
645 novel to interpret patterns of duplication as non-random (e.g. Bailey et al. 2002), indeed
646 we feel that this should be considered the null hypothesis in the absence of the
647 evidence described above.

648 By casting the net widely, we have been able to detect a pattern that does not conform
649 to any one of the common theoretical models for duplicate preservation, but rather
650 subsumes several of them into a temporal series. We cannot definitively determine the
651 precise sequence of events based on our data, however, it may seem that the single
652 model our results most closely resemble is "subfunctionalization" (Force et al. 1999).
653 Additional processes (e.g. "moonlighting" and "neofunctionalization") not described by
654 that model also appear to have contributed to the origins of functional novelty within this
655 gene family, and periods of "degeneration" (a feature of the subfunctionalization
656 hypothesis) may or may not have occurred (see below for further discussion). The
657 "escape from adaptive conflict" model (Hughes et al. 1994) may describe our data even
658 more closely, since toxin genes have undoubtedly specialized for novel functions
659 following an initially pleiotropic period. However, as with subfunctionalization, this model
660 does not describe a subsequent period of classic neofunctionalization, which is
661 observable within the viperid-specific toxin forms on Pla2g2.

662 Conant et al. (Conant, Birchler, and Pires 2014) suggested that a "pluralistic framework"
663 incorporating multiple models may be the most appropriate way to understand the fate
664 of duplicate genes and our analysis corroborates this assertion. The following
665 paragraphs conjecturally describe events that may occur in episodes of
666 "neofunctionalization" (a term used here to describe the emergence of novel functions at
667 the molecular level, and not merely that emergence via Ohno's model). These should
668 not be thought of as an attempt to define a new formal model, but rather to show how
669 each of the previously proposed models may capture only part of the truth. Additional
670 processes not described here may occur in other cases – in evolution, it often seems to
671 be the case that whatever can happen, will happen.

672 The initial acquisition of a novel function may occur: i) when noisy expression patterns
673 (Kaern et al. 2005; Woods 2014) instigate a moonlighting scenario – a single copy gene
674 fulfilling multiple functions by virtue of expression in multiple locations (Copley 2014); or
675 ii) when structural change facilitates interaction with novel partners, whilst maintaining
676 the ancestral function. The novel function may then expose the gene to a distinct

677 selection regime, which may facilitate the accumulation and fixation of further mutations.
678 When a novel function is acquired by a single copy gene, this may create pressure for
679 the creation of duplicate copies such that the multiple functions can be segregated
680 between those copies, which may then specialize. Such a scenario, along with those in
681 which duplication is positively selected due to the benefits of increased gene dosage or
682 robustness (Innan and Kondrashov 2010), may result in selection driving an increased
683 duplication rate. An additional hypothesis describing positive selection on accumulation
684 of duplicate genes suggests that this may occur when duplication results in the
685 spontaneous origin of a novel "function", however, this might be better referred to as a
686 novel "propensity", as it is debatable whether a trait qualifies as functional *prior* to its
687 making a contribution to organismal fitness (i.e. prior to selection) (Jackson and Fry
688 2016; cf. Innan and Kondrashov 2010).

689 Certain novel functions lead to selection for increased expression of a gene product,
690 which also contributes to the fixation of duplicate copies (Margres et al. 2017). Notably,
691 in exochemical systems, since the interaction partners of gene products originate
692 outside the body of the producing organism and the products are secreted
693 extracellularly, the likelihood of a deleterious impact of mutations on fitness is
694 decreased (allowing for their accumulation) and there are no (internal) stoichiometric
695 constraints on dosage. Thus, products of duplicate genes in exochemical systems may
696 escape both negative selection and down-regulation or silencing, thereby having the
697 opportunity of diversifying and rapidly contributing to organismal fitness.

698 In contrast to the model proposed by Lan and Pritchard (Lan and Pritchard 2016) in
699 which coregulation of tandem duplications delays sub- and neo-functionalization, this
700 removal of constraint may facilitate rapid evolutionary divergence prior to genomic
701 separation of duplicate genes. This phenomenon may be termed "exochemical escape",
702 where "escape" refers to the evasion of dosage balance constraints and thereby the
703 solution to "Ohno's dilemma" (Bergthorsson, Andersson, and Roth 2007). A lack of
704 dosage constraint on exochemical/extracellular proteins may also explain the lack of
705 concordance between the evolution of these systems and the broader trend in
706 conservation or deletion of duplicates following whole-genome duplications versus
707 segmental duplications (Conant, Birchler, and Pires 2014) – in exochemical systems,
708 segmentally duplicated genes may persist even when they may have many interaction
709 partners and be involved in the formation of protein complexes.

710 Subsequent to initial duplication, specialization (a.k.a. "escape from adaptive conflict" –
711 Hughes 1994; Innan and Kondrashov 2010) may occur, in which one copy of the gene
712 maintains the original function and the other specializes for its exochemical role, e.g. a
713 role in venom in viperid snakes. This specialization may facilitate a tissue-specific
714 pattern of expression – although it has been suggested that the expression of tandem
715 duplicates is likely to be co-regulated until one copy undergoes chromosomal
716 displacement (Lan and Pritchard 2016) available expression data clearly indicate that
717 Pla2g2G are highly tissue-specific in their expression and that neighbouring genes
718 (Pla2g2E and Pla2g2D) are not expressed in the venom gland (supplementary material
719 SM4 and Aird et al. 2017; Dowell et al. 2016; Vonk et al. 2013).

720 This specialization may lead to increased selection on dosage, driving the accumulation
721 of duplicate genes now specifically expressed within the exochemical system. This is

722 particularly likely for systems in which more gene product is “better”, either leading to a
723 more toxic venom (Margres et al. 2016) or more effective response to infection (e.g.
724 mammalian g2A). At this point, classic Ohno-style redundancy occurs, as multiple gene
725 copies represent both a larger target for mutational change (and thus a network for
726 exploring phenotype space) and each becomes less constrained by purifying selection
727 (Aird et al. 2017). This in turn leads to neofunctionalization, in Ohno’s sense of the term,
728 in which specific gene copies evolve interactions with novel partners.

729 The aforementioned sequence describes a model (and a hypothesis in need of testing)
730 that loosely subsumes moonlighting, specialization/subfunctionalization and
731 neofunctionalization into a single temporal series. Models in which duplication is central
732 to the evolution of functional novelty have dominated discussion in recent years, but the
733 co-option of single copy genes is likely also widespread (True and Carroll 2002;
734 Martinson et al. 2017) and may be the first step on the pathway towards
735 “neofunctionalization”. Assertions that functional novelty may often precede duplication
736 are nothing new. Indeed, they date back at least to the work of Serebrovsky (1938,
737 referenced in Taylor and Raes 2004), who discussed the pleiotropic effects of a single
738 gene being distributed between daughter genes following duplication (see also
739 Piatigorsky and Wistow 1991). More recently, Hughes explicitly states that a period of
740 gene sharing precedes duplication-facilitated specialization (Hughes 1994) . Whether
741 these models, or that which we have outlined in the previous paragraph, should be
742 considered “subfunctionalization” (Force et al. 1999) is perhaps a moot point. The
743 formal subfunctionalization model includes “degeneration” (of regulatory elements or
744 functional structures) following duplication. Whilst this may occur, the significant
745 consequence of duplication, particularly in terms of venom toxins, appears to be
746 “escape from adaptive conflict” (Hughes 1994; Des Marais and Rausher 2008), which in
747 turn leads to neofunctionalization proper (Ohno 1970). This pattern conforms with the
748 analyses of Assis and Bachtrog (Assis and Bachtrog 2015), who demonstrated that
749 subfunctionalization was rare in comparison to conservation, specialization, or
750 neofunctionalization, and indicated that subfunctionalization may be merely a stage in
751 the evolutionary series leading towards neofunctionalization.

752 In any case, formal models are rarely more than schematics, and there is little reason to
753 expect real world sequences to conform to them precisely. Thus, whilst we do not
754 believe we have reconstructed a history that conforms to rigorously defined
755 “subfunctionalization”, clearly that history resembles this model, just as it resembles
756 elements of several others. Hargreaves et al. (Hargreaves et al. 2014) previously
757 argued that venom toxins likely acquire their toxic functions via subfunctionalization
758 rather than neofunctionalization. In this they were making a point of difference with
759 much of the molecular evolutionary work done in the field of toxinology (e.g. Reza et al.
760 2005; Lynch 2007; Fry et al. 2012), in which it had been previously well accepted that
761 Ohno-style neofunctionalization was the dominant process of protein “weaponization”.
762 Indeed, as more research is conducted on the genomes of venomous organisms, it is
763 becoming increasingly evident that even for toxin evolution there can be no one size fits
764 all explanation. For example, in the king cobra genome, considerable evidence of toxin-
765 specific gene family expansion was detected, which appeared to confirm the classic
766 neofunctionalization model’s applicability to toxin evolution (Vonk et al. 2013). A similar
767 pattern was observed in the genome of the anemone *Actinia tenebrosa* (Surm et al.

768 2019). The platypus genome, on the other hand, revealed a pattern in which toxin
769 genes are recruited from families with ancestral CNV, and no evidence of lineage-
770 specific (i.e. associated with the toxin function) expansion was uncovered for most of
771 these families (Wong and Belov 2012). In parasitoid wasps, yet another pattern was
772 observed in which duplication appears to play almost no role; rather, acquisition and
773 loss of toxic function was facilitated by changes in *cis*-regulated gene expression
774 (Martinson et al. 2017).

775 In our study, as described above, we have detected a pattern that suggests that both
776 co-option facilitated by changes in gene expression *and* lineage-specific gene family
777 expansion are important in toxin evolution. We therefore agree (with Hargreaves et al.
778 2014 and others) that Ohno's model does not account for all the details, but feel that it
779 describes an important part of the process characteristic of certain venom toxin families,
780 namely the expansion of these families via duplication and the attendant positively
781 selected evolution of multiple novel functions. We further recommend that the term
782 "neofunctionalization" not be too narrowly defined, as it, etymologically, merely refers to
783 the origin of novel functions. Ohno's initial coinage was a catchy one and we would like
784 the usage of this term to be legitimate, despite the fact that in its narrow definition is
785 does not capture all the details. Those that have read Ohno's monumental publication of
786 1970 (Ohno 1970), know that his thought was expansive and that he described
787 processes akin to subfunctionalization working alongside the neofunctionalization for
788 which he is remembered. In this sense he was like Darwin, whose thoughts on evolution
789 extended beyond Natural Selection and the conceptual tools of what became, in the
790 20th Century, Neo-Darwinism. Thus "Darwinism" is more expansive than "Neo-
791 Darwinism" and "neofunctionalization" may be legitimately considered more expansive
792 than its formal definition suggests.

793 This "highway to neofunctionalization" that we conjecture has shaped the evolution of
794 certain branches of the Pla2g2 family may be unique to rapidly evolving exochemical
795 systems, or may be more widespread. In other cases of multiplication within the Pla2g2
796 family, however, diversification takes place much more sedately. This is evidenced by
797 the fact that plesiotypical D-clade proteins in turtles and alligators are more similar to
798 each other and even to EFC-clade proteins than they are to the divergent forms of
799 mammals, birds or squamates (SM1 Fig. 13). Thus, sequence divergence and the
800 antiquity of the duplication event are not tightly correlated in this gene family – the
801 functional role of the gene in question dictates the dynamism of its evolution.

802 Conclusion

803 By utilizing a labour-intensive manual re-annotation method for genomic regions of
804 interest across 110 genomic sequences from 93 species, we have been able to
805 reconstruct the evolutionary history of the Pla2g2 gene family in unprecedented detail.
806 We have thus contributed qualitatively to our knowledge of the evolution of this gene
807 family and developed a method that can be applied to any gene family exhibiting copy
808 number variation and located in a genomic region with a moderate to high level of
809 synteny. We believe that this method, which we have described in detail in the
810 supplementary materials, is complementary with other annotation methods currently
811 being deployed and agree with the assertion made in a recent study of transcriptome

assembly pipelines that no single method of assembly and/or annotation is “complete”, or equally applicable to all systems (Venturini et al., 2018). In much the same way as scientific knowledge more generally advances as a consequence of the interaction of a variety of theoretical and experimentally supported arguments and perspectives put forth by a community of enquirers (Popper 1963; Hull 1988; Renn 2020), multiple complementary methods are required to extract the order from complex datasets such as genomes and transcriptomes.

The major theoretical contribution of the paper is the evidence it provides that novel gene functions emerge as the result of a change in a gene product’s context, which may occur with or without duplication. Indeed, in this case duplication appears as a likely *consequence* of “neofunctionalization”, not its antecedent. As a result, we have argued that whilst many published models of gene evolution tell part of the story, no single model captures the full range of possible pathways towards the evolution of functional novelty. We refer to this process as “neofunctionalization”, in deference to Ohno but with none of the theoretical commitments (particularly to duplication preceding the origins of novel function) that this term often implies.

Our analysis is by no means complete but indicates that further research is required in a wide range of model systems to differentiate between molecular evolutionary models and to fill in the gaps that may exist in all of them. Ultimately each system, once sufficiently well understood, may tell a slightly different story. We may have to acknowledge that there can be no one size fits all model for the evolution of functional novelty and that whatever can happen, will happen. Borrowing the words of two famous scientists (one fictitious), “life finds a way”[1], because “evolution is cleverer than you are”[2].

1. Ian Malcolm

2. Leslie Orgel

Materials and Methods

While this manuscript was in preparation, we used the approach developed for this study to successfully locate and annotate several gene families albeit at a far less ambitious scale: vertebrate NAD glycohydrolases (Koludarov and Aird, 2019), mammalian kallikreins (Casewell et al. 2019) and Indian cobra three finger toxins (Suryamohan et al. 2020). This shows that the approach presented here (for the excruciating detail see Research Notes in SM1) can be useful in studying the evolution of other gene families.

Locating Pla2g2 cluster

We used published annotations to find genomic sequences that correspond to OTUD3-MUL1 region since those genes were previously shown to be flanking the Pla2g2 cluster in vipers (Chijiwa et al. 2003, Dowell et al. 2016). When no annotations were available, we used the blastn feature of NCBI-BLAST v.2.7.1+ suite to find them (blastn, e-value cutoff of 0.05, default restrictions on word count and gaps), using known sequences as queries. We used the well-annotated and supported by transcriptomic sequencing Protobothrops (Aird at al., 2017) and Crotalus (Dowell et al. 2016; Dowell et al. 2018) genomic scaffolds as starting points and traced single-copy genes contained in that region to their orthologs in non-snake reptiles (lizards,

855 turtles, alligators, birds) as well as mammals. By doing that we established the extreme synteny
856 of the region going as far back as the ancestor of Tetrapoda (see supplementary material SM1
857 Fig. 1 and SM1 Fig. 2 for genomic maps of the region and comparison between
858 representatives). In all the cases, Pla2g2 cluster was located downstream of OTUD3 gene with
859 MUL1 (squamates) and UBXN10 (all other lineages) genes flanking it on the other side.
860
861 *Creating of BLAST database of Pla2g2 exons*
862 All annotated previously annotated exons were extracted from the genomic sequences via
863 "extract" function in Geneious and saved as a fasta file from which initial BLAST database was
864 build from. All genomic sequences in the study were then queried against that database using
865 blastn (blastn, e-value cutoff of 0.05, default restrictions on word count and gaps). BLAST
866 output was formatted into an annotation gff file and imported into Geneious to visualize the
867 results. All exons that had hits against multiple other exons (that is, more than itself and a single
868 closely related species) were exported the same into a new database. All instances when blast
869 revealed previously unannotated exons, they were examined to make sure they complied with
870 the exon structure of Pla2g2 genes that were verified by the transcriptome of a good quality,
871 that is Human, Mouse or snake Pla2g2. Splice cites for each exon were assigned manually
872 based on conserved splice signals for eukaryots. All exons recovered via that method were also
873 put into the new database. Then the entire dataset was queried again using the same dataset
874 tblastx function of NCBI-BLAST suite with e-value cutoff of 0.01. And all the recovered BLAST
875 hits were carefully examined and extracted if they fit the criteria of being a Pla2g2 exon: no
876 frameshift mutations, existing splice sites that do not differ significantly from the known Pla2g2
877 exons. Then the process was repeated one final time. No new exons were discovered during
878 the last run.
879
880 *Creating a dataset for phylogenetic and selection analyses*
881 Since all previously described Pla2g2 genes have 3 exons that encode the mature protein, we
882 considered triplets of those exons (labelled as 2, 3 and 4 respectively) as a separate Pla2g2
883 gene if they were located in close proximity to each other. The only Pla2g2 group that didn't
884 conform to the 4 exon rule was mammalian g2F that had 5 exons, with the first two (labelled
885 exon F1 and exon F2) encoding a transmembrane domain, while the other three exons (labelled
886 2, 3 and 4 for the sake of consistency with other groups) were structurally identical to that of
887 other groups. Coding regions of all genes with intact exons (2,3,4) were extracted and
888 translated for the alignment using Geneious and then aligned with localpair function of MAFFT
889 software v7.305 (Katoh and Standley 2013) with 1000 iterations (--localpair --maxiterate 1000).
890 Alignments were refined by hand using AliView software (Larsson, 2014) to make sure that
891 obviously homologous parts of the molecule (like the cysteine backbone) were aligned properly.
892 The final dataset was trimmed to exclude sequences that might be pseudogenes and included
893 452 protein sequences
894
895 *Phylogenetic analyses of protein sequences*
896 Phylogenetic analysis was performed using ExaBayes v1.5 (Aberer, Kobert, and Stamatakis
897 2014) software with 10M generations of 4 runs and 4 chains running in parallel (total of 16
898 chains). The evolutionary model was not specified, which allows software to alternate between

different models, until the chains converge on the one that provides the best fit. Final consensus trees were generated with the consense command and the default 25% burn-in. To summarize the run statistics we used postProcParam function and assessed convergence between runs with plots of likelihood and parameter estimates, Effective Sampling Size values higher than 200 and the Potential Scale Reduction Factor lower than 1.1.
FigTree v1.4.3 was used to generate tree figures.
For the construction of species trees we used TimeTree online tool (Kumar et al. 2017), substituting species absent from the database for closely related species of the same genus.

Analysis of synteny

The overwhelming majority of genes were grouped together on the phylogenetic tree in the way that was totally consistent with their position on the chromosome. The few problematic sequences came from the groups for which only single genome was available – Monotremata and stem anquimorpha that have unique or transitional forms. In those cases we grouped their forms in accordance with the rules we developed for the nomenclature of this gene family (see SM1).

Selection analyses

The entire dataset that was used for the construction of the phylogenetic trees was used for the analysis of selection. To detect signatures of natural selection and gauge the regime of selection dictating the evolution of various Pla2g2 lineages, site- and branch-site specific maximum likelihood models implemented in CodeML of the PAML (Phylogenetic Analysis by Maximum Likelihood) package were used (Yang 2007), and the omega parameter (ω), or the ratio of non-synonymous to synonymous substitutions, was estimated. To determine the statistical significance of the results obtained from nested models M7 (null model) and M8 (alternate model), the likelihood scores were compared with a likelihood ratio test (LRT). Amino acid sites under the influence of positive diversifying selection were identified using the Bayes Empirical Bayes (BEB) approach in M8 (Yang, Wong, and Nielsen 2005). Data-monkey webserver was used to assess the influence of episodic selection using Mixed Effect Model of Evolution (MEME) and the pervasive effects of diversifying and purifying selection using Fast Unconstrained Bayesian AppRoximation (FUBAR) analysis (Murrell et al. 2012, 2013). For assessing the nature of selection underpinning various Pla2g2 lineages, branch-site specific two ratio model (Yang 1998; Yang and Nielsen 1998) was employed by marking the Pla2g2 lineage suspected to be evolving under positive selection as foreground (ω≥1, alternate model assuming positive selection), and by constraining others as background lineages (ω≤1, null model assuming negative selection or neutral evolution). The likelihood estimates of the null and alternate models were compared with an LRT for determining the significance

List of supplementary materials

File SM1.pdf contains a step by step detailed walkthrough of our methodology and the logic behind it, including (but not by any stretch of imagination limited to):

1. Pla2g2 genomic neighborhood of representative taxonomic clades

2. Convergent evolution of the cysteine arrangement formerly utilized as the defining characteristic of «Group 5» Pla2
3. Results of annotations of *Boa* genome assemblies utilizing different assembly algorithms.
4. Phylogenetic trees of Pla2g2
5. Phylogenetic and syntenic relationships within the placental Pla2g2A clade.
6. Full exonic map for all the species and genomic regions used in this study
7. Alignment of consensus sequences for each of the Pla2g2 group
8. Results of selection analysis

File SM2.xlsx contains curated compiled annotation of all genomic sequences used in this study with species names, genomic IDs, references, protein IDs, annotated genes as well as information on their status in previous annotations and key for the naming system used in the datasets.

File SM3_SM4.xlsx:

1. SM3: expression data for Pla2s based on previously published studies
2. SM4: expression data for Pm based on Aird 2017

File SM5_AnnotationsPla2g2.txt contains concatenated manual annotations made for this study

File SM6_Pla2g2_alignment.txt contains alignment of all the Pla2s used for constructing phylogenetic trees in fasta format

Acknowledgements

IK would like to thank all the attendees of Venom Evolution, Function and Biomedical Applications Gordon Research Conference 2018 for productive discussions in relation to this study, and Noah L. Dowell from S.B. Carroll's lab and Michael Broe from H.L. Gibbs' lab for sharing genomic sequences of *Crotalus* and information on Pla2g2 cluster in *Sistrurus* respectively. IK was partially supported by Japan Society for the Promotion of Science KAKENHI Grant Number 19K16206. KS was supported by funding from the Department for International Development (DFID: grant IAVI/BES/KASU/0002), the Department for Biotechnology, Government of India, through the DBT-IISc Partnership Program, DST-INSPIRE Faculty Award (DST/INSPIRE/04/2017/000071), and DST-FIST (SR/FST/LS-II/2018/233).

References

Aberer AJ, Kobert K, Stamatakis A. 2014. ExaBayes: massively parallel bayesian tree inference for the whole-genome era. Mol. Biol. Evol. 31:2553–2556.

Aird SD, Arora J, Barua A, Qiu L, Terada K, Mikheyev AS. 2017. Population Genomic Analysis of a Pitviper Reveals Microevolutionary Forces Underlying Venom Chemistry. Genome Biol. Evol. 9:2640–2649.

983 Assis R, Bachtrog D. 2015. Rapid divergence and diversification of mammalian duplicate gene
984 functions. BMC Evol. Biol. 15:138.

985 Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW,
986 Eichler EE. 2002. Recent segmental duplications in the human genome. Science
987 297:1003–1007.

988 Birchler JA, Veitia RA. 2012. Gene balance hypothesis: connecting issues of dosage sensitivity
989 across biological disciplines. Proc. Natl. Acad. Sci. U. S. A. 109:14746–14753.

990 Birts CN, Barton CH, Wilton DC. 2010. Catalytic and non-catalytic functions of human IIA
991 phospholipase A2. Trends Biochem. Sci. 35:28–35.

992 Burke JE, Dennis EA. 2009. Phospholipase A2 structure/function, mechanism, and signaling. J.
993 Lipid Res. 50 Suppl:S237–S242.

994 Casewell NR, Petras D, Card DC, Suranse V, Mychajliw AM, Richards D, Koludarov I, Albulescu
995 LO, Slagboom J, Hempel BF, Ngum NM, et al. 2019. Solenodon genome reveals
996 convergent evolution of venom in eulipotyphlan mammals. Proc. Natl. Acad. Sci. U. S. A.
997 116: 25745-25755.

998 Chijiwa T, Hamai S, Tsubouchi S, Ogawa T, Deshimaru M, Oda-Ueda N, Hattori S, Kihara H,
999 Tsunasawa S, Ohno M. 2003. Interisland mutation of a novel phospholipase A2 from
1000 *Trimeresurus flavoviridis* venom and evolution of Crotalinae group II phospholipases A2.
1001 J. Mol. Evol. 57:546–554.

1002 Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the
1003 interplay of multiple models for duplicate gene evolution over time. Curr. Opin. Plant Biol.
1004 19:91–98.

1005 Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new
1006 functions. Nat. Rev. Genet. 9:938–950.

1007 Copley SD. 2014. An evolutionary perspective on protein moonlighting. Biochem. Soc. Trans.
1008 42:1684–1691.

1009 Danchin A, Ouzounis C, Tokuyasu T, Zucker J-D. 2018. No wisdom in the crowd: genome
1010 annotation in the era of big data – current status and future prospects. Microb.
1011 Biotechnol. 11:588–605.

1012 Dennett DC. 1995. Darwin's dangerous idea. Sciences 35:34–40.

1013 Des Marais DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an
1014 anthocyanin pathway gene. Nature 454:762–765.

1015 Dowell NL, Giorgianni MW, Kassner VA, Selegue JE, Sanchez EE, Carroll SB. 2016. The Deep
1016 Origin and Recent Loss of Venom Toxin Genes in Rattlesnakes. Curr. Biol. 26:2434–
1017 2445.

1018 Dowell NL, Giorgianni MW, Griffin S, Kassner VA, Selegue JE, Sanchez EE, Carroll SB. 2018.
1019 Extremely divergent haplotypes in two toxin gene complexes encode alternative venom
1020 types within rattlesnake species. Curr. Biol. 28:1016-1026.

1021 Drukewitz SH, von Reumont BM. 2019. The Significance of Comparative Genomics in Modern
1022 Evolutionary Venomics. Frontiers in Ecology and Evolution. May 9;7.

1023 Dunn CW, Munro C. 2016. Comparative genomics and the diversity of life. Zoologica Scripta
1024 45:5–13.

1025 Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of
1026 duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.

1027 French WJ, Hayes WK, Bush SP, Cardwell MD, Bader JO, Rael ED. 2004. Mojave toxin in
1028 venom of *Crotalus helleri* (Southern Pacific Rattlesnake): molecular and geographic
1029 characterization. Toxicon 44:781–791.

1030 Fry BG. 2015. Venomous Reptiles and Their Toxins: Evolution, Pathophysiology and
1031 Biodiscovery. Oxford University Press

1032 Fry BG, Casewell NR, Wüster W, Vidal N, Young B, Jackson TNW. 2012. The structural and
1033 functional diversification of the Toxicofera reptile venom system. Toxicon 60:434–448.

1034 Georgieva D, Seifert J, Öhler M, von Bergen M, Spencer P, Arni RK, Genov N, Betzel C. 2011.
1035 *Pseudechis australis* venomics: adaptation for a defense against microbial pathogens
1036 and recruitment of body transferrin. J. Proteome Res. 10:2440–2464.

1037 Gopalakrishnakone P. 1997. Myotoxic phospholipases from snake venoms: general
1038 myoglobinuric and local myonecrotic toxins. Toxicon 4:474.

1039 Gould SJ, Vrba ES. 1982. Exaptation – a Missing Term in the Science of Form. Paleobiology
1040 8:4–15.

1041 Guschanski K, Warnefors M, Kaessmann H. 2017. The evolution of duplicate gene expression
1042 in mammalian organs. Genome Res. 27:1461–1474.

1043 Guttinger, S., 2018. A Process Ontology for Macromolecular Biology. Everything Flows:
1044 Towards a Processual Philosophy of Biology, p.303. Oxford University Press.

1045 Hargreaves AD, Swain MT, Hegarty MJ, Logan DW, Mulley JF. 2014. Restriction and
1046 recruitment-gene duplication and the origin and evolution of snake venom toxins.
1047 Genome Biol. Evol. 6:2088–2095.

1048 Hedges SB, Kumar S. 2009. The Timetree of Life. Oxford University Press

1049 Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. Proc. Biol.
1050 Sci. 256:119–124.

1051 Hull DL. 1988. Science as a Process: An Evolutionary Account of the Social and Conceptual
1052 Development of Science. The University of Chicago Press.

1053 Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing
1054 between models. Nat. Rev. Genet. 11:97–108.

1055 Kaern M, Elston TC, Blake WJ, Collins JJ. 2005. Stochasticity in gene expression: from theories
1056 to phenotypes. Nat. Rev. Genet. 6:451–464.

1057 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
1058 improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

1059 Kini RM. 2003. Excitement ahead: structure, function and mechanism of snake venom
1060 phospholipase A2 enzymes. Toxicon 42:827–840.

1061 Koludarov I, Aird SD, 2019. Snake venom NAD glycohydrolases: primary structures, genomic
1062 location, and gene structure. PeerJ 7:e6154.

1063 Kumar S, Stecher G, Suleski M, and Hedges SB, 2017. TimeTree: a resource for timelines,
1064 timetrees, and divergence times. Molecular Biology and Evolution 34: 1812-1819.

1065 Lan X, Pritchard JK. 2016. Coregulation of tandem duplicate genes slows evolution of
1066 subfunctionalization in mammals. Science 352:1009–1013.

1067 Lynch VJ. 2007. Inventing an arsenal: adaptive evolution and neofunctionalization of snake
1068 venom phospholipase A2 genes. BMC Evol. Biol. 7:2.

1069 Mackessy SP. 2010. Evolutionary trends in venom composition in the western rattlesnakes
1070 (*Crotalus viridis* sensu lato): toxicity vs. tenderizers. Toxicon 55:1463–1474.

1071 Margres MJ, Bigelow AT, Lemmon EM, Lemmon AR, Rokyta DR. 2017. Selection To Increase
1072 Expression, Not Sequence Diversity, Precedes Gene Family Origin and Expansion in
1073 Rattlesnake Venom. Genetics 206:1569–1580.

1074 Margres MJ, Wray KP, Seavy M, McGivern JJ, Herrera ND, Rokyta DR. 2016. Expression
1075 Differentiation Is Constrained to Low-Expression Proteins over Ecological Timescales.
1076 Genetics 202:273–283.

1077 Martinson EO, Mrinalini, Kelkar YD, Chang C-H, Werren JH. 2017. The Evolution of Venom by
1078 Co-option of Single-Copy Genes. Curr. Biol. 27:2007–2013.e8.

1079 Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013.
1080 FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol. Biol.
1081 Evol. 30:1196–1205.

1082 Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting
1083 individual sites subject to episodic diversifying selection. PLoS Genet. 8:e1002764.

1084 Nevalainen TJ. 2007. Phospholipase A2 in acute pancreatitis: review. Am. J. Surg. 194:S28–
1085 S32.

1086 Nevalainen TJ, Graham GG, Scott KF. 2008. Antibacterial actions of secreted phospholipases
1087 A2. Review. Biochim. Biophys. Acta 1781:1–9.

1088 Nobre T, Campos MD, Lucic-Mercy E, Arnholdt-Schmitt B. 2016. Misannotation Awareness: A
1089 Tale of Two Gene-Groups. Front. Plant Sci. 7:868.

1090 Ohno S. 1970. Evolution by Gene Duplication. Springer.

1091 Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, Füllgrabe A, Fuentes AM-
1092 P, Jupp S, Koskinen S, et al. 2016. Expression Atlas update – an integrated database of
1093 gene and protein expression in humans, animals and plants. Nucleic Acids Res.
1094 44:D746–D752.

1095 Phillips CM. 2002. Sea snake envenomation. Dermatol. Ther. 15:58–61.

1096 Piatigorsky J, Wistow G. 1991. The recruitment of crystallins: new functions precede gene
1097 duplication. Science 252:1078–1079.

1098 Popper KR. 1963. Science as falsification. Conjectures and refutations. London: Routledge and
1099 Keagan Paul.

1100 Reams AB, Roth JR. 2015. Mechanisms of gene duplication and amplification. Cold Spring
1101 Harb. Perspect. Biol. 7:a016592.

1102 Renn J. 2020. The Evolution of Knowledge: Rethinking Science for the Anthropocene. Princeton
1103 University Press.

1104 Reza MA, Swarup S, Kini MR. 2005. Two parallel prothrombin activator systems in Australian
1105 rough-scaled snake, *Tropidechis carinatus*. Thromb. Haemost. 93:40–47.

1106 Six DA, Dennis EA. 2000. The expanding superfamily of phospholipase A2 enzymes:
1107 classification and characterization. Biochimica et Biophysica Acta (BBA) – Molecular and
1108 Cell Biology of Lipids 1488:1–19.

1109 Surm JM, Smith HL, Madio B, Undheim EAB, King GF, Hamilton BR, Burg CA, Pavasovic A,
1110 Prentis PJ. 2019. A process of convergent amplification and tissue-specific expression
1111 dominates the evolution of toxin and toxin-like genes in sea anemones. Mol. Ecol.
1112 181:60.

1113 Suryamohan K, Krishnankutty SP, Guillory J, Jevit M, Schröder MS, Wu M, Kuriakose B,
1114 Mathew OK, Perumal RC, Koludarov I, Goldstein LD, et al. 2020. The Indian cobra
1115 reference genome and transcriptome enables comprehensive identification of venom
1116 toxins. Nat Genet doi:10.1038/s41588-019-0559-8

1117 Taylor JS, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas.
1118 Annu. Rev. Genet. 38:615–643.

1119 Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, Alm T, Asplund A, Björk L,
1120 Breckels LM, et al. 2017. A subcellular map of the human proteome. Science 356.

1121 True JR, Carroll SB. 2002. Gene co-option in physiological and morphological evolution. Annu.
1122 Rev. Cell Dev. Biol. 18:53–80.

1123 Van de Peer Y, Maere S, Meyer A. 2010. 2R or not 2R is not the question anymore. Nat. Rev.
1124 Genet. 11:166.

1125 Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. 2018. Leveraging multiple
1126 transcriptome assembly methods for improved gene structure annotation. GigaScience.
1127 7, 8.

1128 Vonk FJ, Casewell NR, Henkel CV, Heimberg AM, Jansen HJ, McCleary RJR, Kerkkamp HME,
1129 Vos RA, Guerreiro I, Calvete JJ, et al. 2013. The king cobra genome reveals dynamic
1130 gene evolution and adaptation in the snake venom system. Proc. Natl. Acad. Sci. U. S.
1131 A. 110:20651–20656.

1132 Wang J, Al-Ouran R, Hu Y, Kim S-Y, Wan Y-W, Wangler MF, Yamamoto S, Chao H-T, Comjean
1133 A, Mohr SE, et al. 2017. MARRVEL: Integration of Human and Model Organism Genetic
1134 Resources to Facilitate Functional Annotation of the Human Genome. Am. J. Hum.
1135 Genet. 100:843–853.

1136 Whittington AC, Mason AJ, Rokyta DR. 2018. A Single Mutation Unlocks Cascading Exaptations
1137 in the Origin of a Potent Pitviper Neurotoxin. Mol. Biol. Evol. 35:887–898.

1138 Wistow G, Piatigorsky J. 1987. Recruitment of enzymes as lens structural proteins. Science
1139 236:1554–1556.

1140 Wong ESW, Belov K. 2012. Venom evolution through gene duplications. Gene 496:1–7.

1141 Woods HA. 2014. Mosaic physiology from developmental noise: within-organism physiological
1142 diversity as an alternative to phenotypic plasticity and phenotypic flexibility. J. Exp. Biol.
1143 217:35–45.

1144 Yamaguchi K, Chijiwa T, Ikeda N, Shibata H, Fukumaki Y, Oda-Ueda N, Hattori S, Ohno M.
1145 2014. The finding of a group IIE phospholipase A2 gene in a specified segment of
1146 *Protobothrops flavoviridis* genome and its possible evolutionary relationship to group IIA
1147 phospholipase A2 genes. Toxins 6:3471–3487.

1148 Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. Nat. Rev.
1149 Genet. 13:329–342.

1150 Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate
1151 lyso- zyme evolution. Mol. Biol. Evol. 15:568–573.

1152 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–
1153 1591.

1154 Yang Z, Nielsen R. 1998. Synonymous and nonsyn- onymous rate variation in nuclear genes of
1155 mam- mals. J. Mol. Evol. 46:409–418.

1156 Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites
1157 under positive selection. Mol. Biol. Evol. 22:1107–1118.

1158 Zallot R, Harrison KJ, Kolaczkowski B, de Crécy-Lagard V. 2016. Functional Annotations of
1159 Paralogs: A Blessing and a Curse. Life 6.

1160 Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW,
1161 et al. 2014. Comparative genomics reveals insights into avian genome evolution and
1162 adaptation. Science 346:1311–1320.

1163 Zhang L, Cai X, Wu J, Liu M, Grob S, Cheng F, Liang J, Cai C, Liu Z, Liu B, et al. 2018.
1164 Improved Brassica rapa reference genome by single-molecule sequencing and
1165 chromosome conformation capture technologies. Hortic Res 5:50.