ARTICLE (Discoveries)

**TITLE: Family saga: 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], Andrea Pozzi[1], Alexander S Mikheyev[1,3]

1. Ecology and Evolution unit, Okinawa Institute of Science and Technology, Japan
2. Australian Venom Research Unit, Department of Pharmacology and Therapeutics, University of Melbourne, Australia
3. Evolutionary Genomics Research group, Australian National University, Canberra, Australia

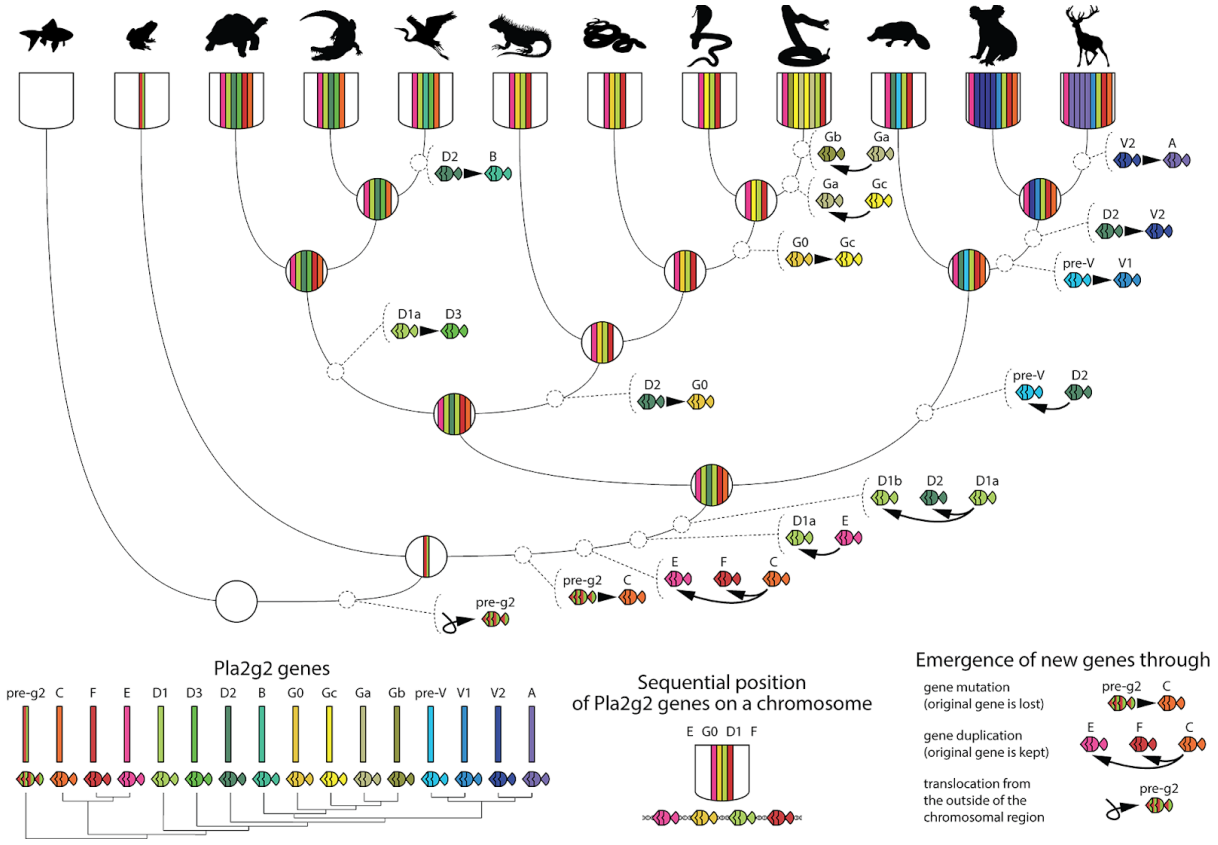
Corresponding authors:

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. Numerous alternative models have been formulated, however, including several (such as "subfunctionalization" and "escape from adaptive constraints") 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 in 92 genomes 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 a novel method to annotate gene clusters, which overcomes many limitations of automatic annotation. 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 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, allowing each copy to specialise for a particular function; in the latter, a genuinely 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 constraints", in which a novel function emerges following the partitioning of the ancestral function; and of neofunctionalization, such as "modified duplication" in which duplication is itself positively selected, driving the accumulation of redundant copies in gene family networks which may become hotspots for functional novelty (Innan and Kondrashov 2010).

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 (Hughes 1994). Thus, in these scenarios (which include subfunctionalization models), acquisition of a novel function occurs prior to duplication.

As recognised 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 centres 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.

Virtually all studies of gene families have taken a comparative approach focusing on statistical patterns such as copy number variation (CNV) over time. This approach is powerful and general and can propose or 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 generally unusual cases, and hard to generalize. As a result, it is not clear whether the processes of gene family evolution fit global patterns. Therefore there is a pressing need to reconstruct gene families in detail, a task that is nonetheless 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 g2 (Pla2g2) – a family of enzymes with multiple interaction partners that exhibits great CNV in vertebrate genomes. Pla2g2 are additionally interesting in that they possess multiple functional roles and have been differentially neofunctionalized in divergent vertebrate lineages – mammalian Pla2g2A is an important component of innate immunity (Nevalainen 2007; Birts, Barton, and Wilton 2010), whilst Pla2g2G are a major component of the venom of viperid snakes (Kini 2003). 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 the 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.

### New Approach

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. In particular, genomic annotations remain a problematic aspect. The most accurate annotations are produced by aligning a properly processed (e.g. masked) genome with comprehensive set of RNA-seq data from the same specimen or at least a member of the same species (Yandell and Ence 2012; Ekblom and Wolf 2014). Thus, genomes that don’t have complementary RNA-seq data must be annotated using predictive and homology-based algorithms. Though these algorithms are rapidly advancing, they 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 proves erroneous.

Another weakness of commonly used homology-based annotation pipelines is that they attempt to align an entire protein or mRNA sequence with a genomic sequence, a practice that is very 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).

Incorrect annotations can create a ripple effect, with predicted genes and proteins being added to a diverse range of online databases, impacting any study that uses ontology databases as well as future genomic annotations (Klimke et al. 2011; Schnoes et al. 2009). To address these

concerns, we have developed a precision approach to annotating tandem-array duplications 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 section and Fig. 8 there). Currently, this approach works only for targeted gene families and cannot be used to annotate entire genomes. It is most effective when used as a complementary tool in concert with automated annotations, in order to refine annotations of problematic regions. When used to annotate genomes for which RNA-seq data wasn't available, this approach recovered functional genes with a higher fidelity than published annotations, as well as providing information on pseudogenes and orphan exons (Fig. 8).

In the present study, we utilized this method to reconstruct the evolutionary history of this gene family. By examining over 90 genomes from 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. 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, with birds, snakes and mammals deriving proteins with novel functions from the same ancestral gene.

## Results and Discussion

### *Pla2g2 gene cluster synteny is conserved in amniotes*

In the present study, we used previously published genomes (see SM2 for the full list) and manual re-annotations to examine the genomic region in which the Pla2g2 gene cluster is located. We discovered remarkable synteny in this region: upstream and downstream regions flanking the Pla2g2 cluster share more than ten genes in almost exactly similar positions across the entire Tetrapoda clade (Fig. 2 and SM1-1 and SM1-2). This allowed us to reconstruct duplication events spanning 300 million years of the family's evolutionary history. Interestingly, against this background of conservation, several unrelated species (e.g., *Gecko japonicus, Pelodiscus sinensis*) exhibit substantial rearrangements in this region. In addition to species-specific rearrangements, the most prominent long-term rearrangement is shared by all squamate reptiles (lizards and snakes), making them more divergent from crocodylians in this region than crocodylians are from humans. Thus, phylogenetic distance is not necessarily an accurate predictor of syntenic conservation in this genomic region.

All analysed genes, excepting those of amphibians, fall into two major clades that diverged from one another some time following the split of Amniota from the amphibians and prior to the evolution of the inferred most recent common ancestor (MRCA) of extant amniotes. These two clades contain Pla2g2s E, F & C; and Pla2g2 D, respectively. Members of the "EFC clade" occupy the flanks of the cluster with g2E genes positioned next to OTUD3 and g2F and/or g2C positioned next to UBXN10 (Fig. 2). Members of the EFC clade are single-copy genes with a tendency to become pseudogenes and are under relaxed selection in all taxonomic lineages (see SM3 for gene maps and SM8 for selection analysis logs). On the other hand, members of the "D

clade" occupy the centre of the cluster and are involved in all subsequent expansion and neofunctionalization events within the Pla2g2 family (ibid and Fig. 3). Both viperid toxins and mammalian antimicrobial g2A derive from within this clade.

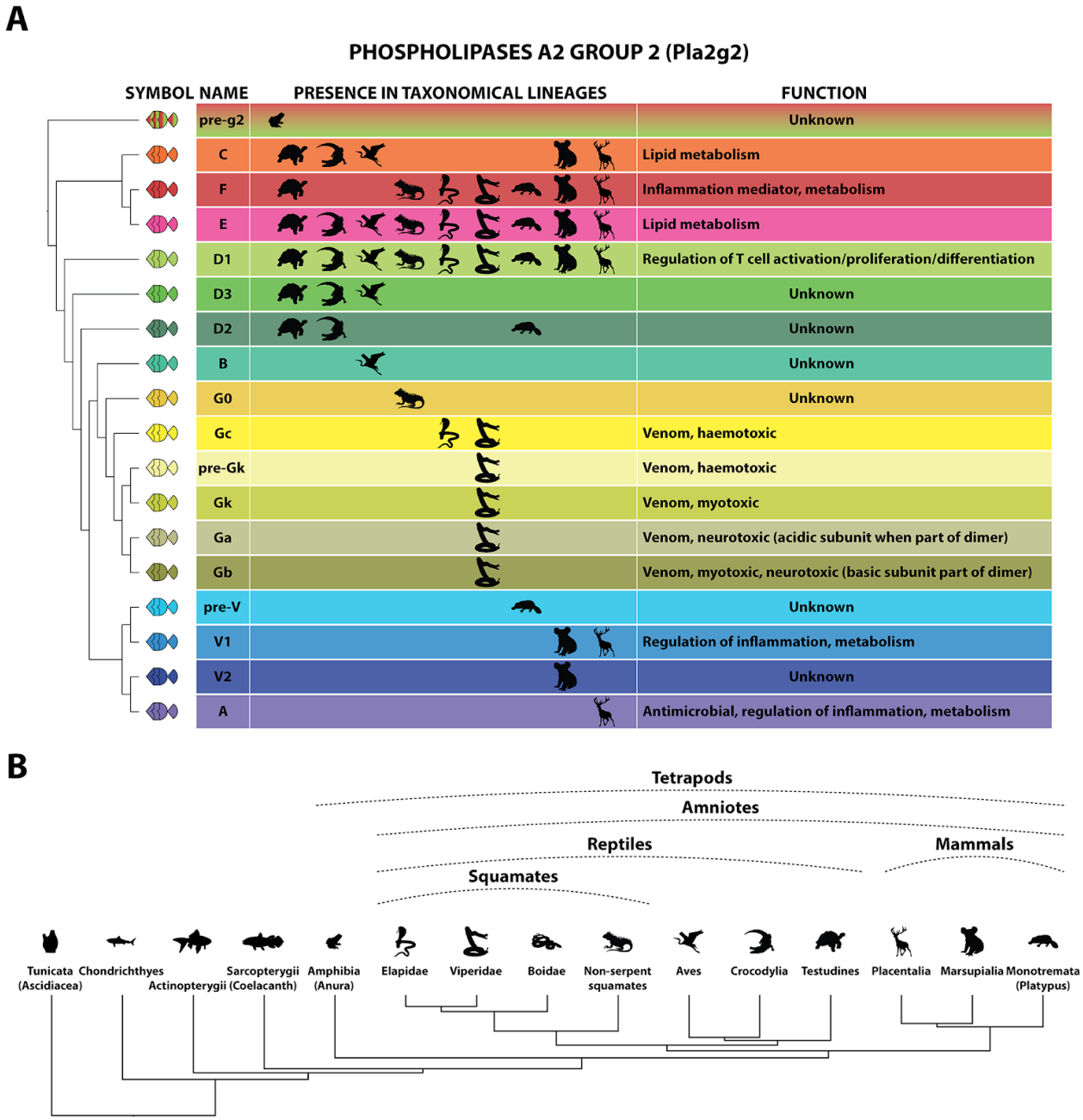



**Fig. 1. A:** Key to colour-coding of Pla2g2 clades, along with phylogenetic relationships, functions and taxonomic presence. Functions as in (Petryszak et al. 2016; Thul et al. 2017; Six and Dennis 2000) and (B. Fry 2015). **B:** Phylogenetic tree of taxa surveyed in this study and their respective symbols used throughout the article.

With our grouping 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 interpreting them in light of the evolutionary history revealed by our analyses (see discussion section for detailed review of Pla2 nomenclature). We expanded the definition of

each Pla2g2 group to include all homologues where possible and changed some of the names to reflect evolutionary relationships where it was deemed necessary (Fig. 1). Thus, chicken g2A and g5 were renamed to be g2D and g2C respectively, and mammalian group 5 became g2V since it evolved from a group 2 precursor and in its turn gave rise to g2A. In addition, g2V has no distinct structural features that would justify its position as a separate Pla2 group (the features previously used to describe it are in fact shared by several other g2 genes, that acquired it independently – SM1-3).

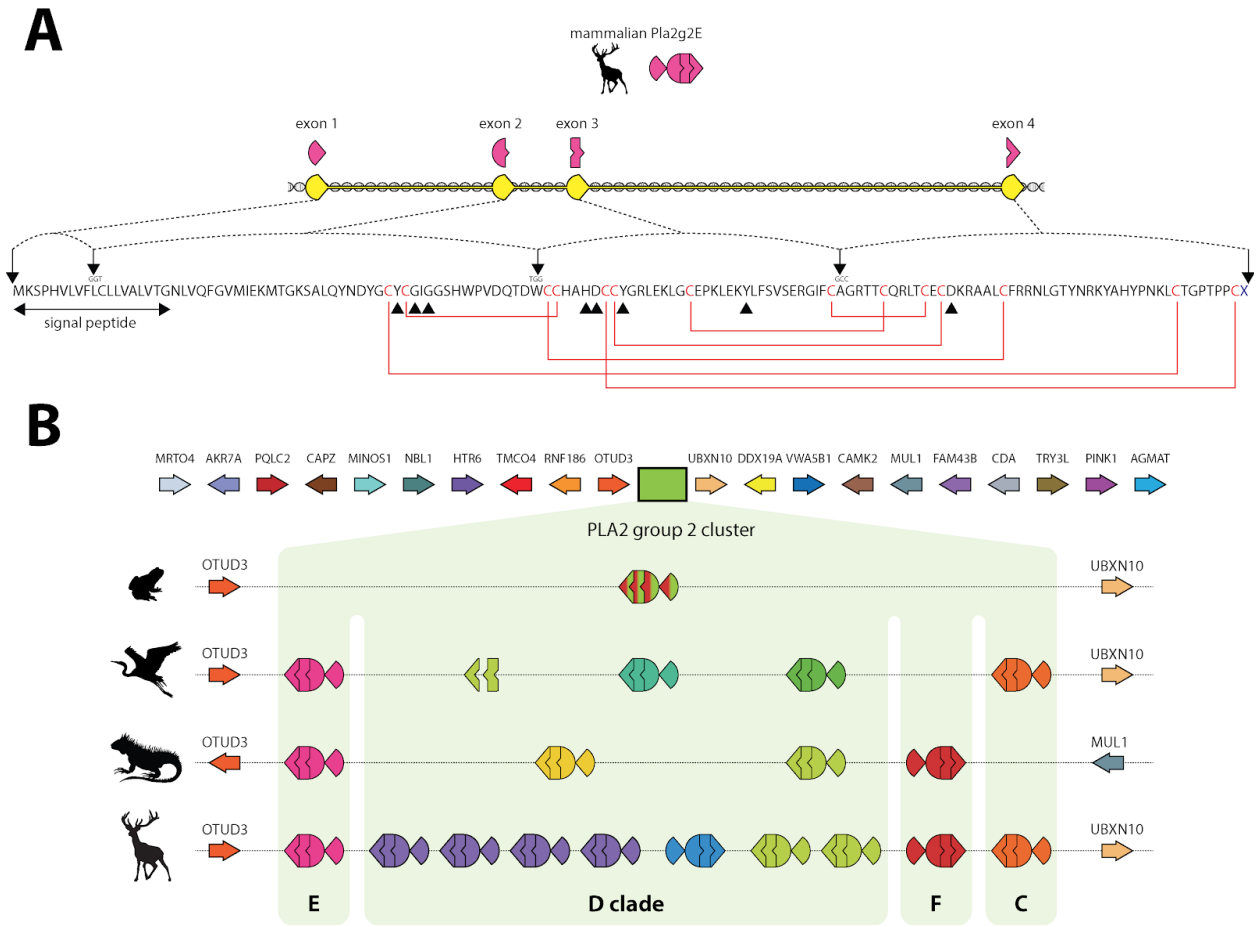



**Fig. 2. A:** Structure of a typical Pla2g2 gene and the way it corresponds to protein sequence, using *Homo sapiens* g2E as an example (intron lengths are up to scale, cysteines and disulfide bonds are in red, triangles mark catalytic sites, arrows mark the position of the splice sites with respect to codons). A typical Pla2g2 gene has 4 exons, first of which encodes half of the signal peptide, while three other exons encode what becomes a mature protein, hence us graphically representing the former as a "tail" and latter as a "body". For some genes in our analysis we were unable to recover first exon, and other genes might have multiple exons coding for signal peptide or transmembrane domain (e.g. mammalian g2F). However, in all the cases the three exons that encode mature peptide retain the same organisation. Number of cysteines (and therefore disulfide bonds) varies between the genes from 12 to 15, but the most common and likely ancestral condition is 14. **B:** Location of Pla2g2 cluster and schematic of its content in Amphibia, Aves, Squamata and Mammalia. Pla2g2 genes are colour-coded by group (see Fig. 1 for details). Notice the remarkable synteny in this region – upstream and downstream regions flanking the Pla2g2 cluster share more than ten genes in almost exactly similar positions across the entire Tetrapoda clade (see SM1-1 and SM1-2 for more details). All analysed genes, excepting those of amphibians, fall into two major clades that diverge in the inferred amniote MRCA. These two clades contain Pla2g2 E (pink), F (red) & C (orange); and Pla2g2 D, respectively. Members of the EFC clade occupy the flanks of the cluster with g2E genes positioned next to

OTUD3 and g2F and/or g2C positioned next to UBXN10. Members of the EFC clade are single-copy genes with a tendency to become pseudogenes and are under relaxed selection in all taxonomic lineages. On the other hand, members of the D clade occupy the centre of the cluster and are involved in all subsequent expansion and neofunctionalization events within the Pla2g2 family. Both viperid toxins and mammalian antimicrobial g2A derive from within this clade.

### *Evolution of the Pla2g2 cluster*

**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 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 genomic region. This region underwent duplication, possibly during the whole genome duplications that, according to the 2R hypothesis, occurred early in vertebrate evolution, cf. (Van de Peer, Maere, and Meyer 2010). Subsequent to that, a genomic rearrangement resulted in two regions with one Pla2 each – the ancestral g2 gene and Pla2 otoconin-22-like gene. Each of these genes are present in a single copy in amphibians, however the g2 gene has a tendency to disappear in much of the Amphibia clade. In contrast, g2 persisted and presumably gained functional significance in the Amniota clade – by the time of the inferred amniote MRCA it had undergone ancestral expansion to form a cluster of 5-6 genes (Fig. 3).

Based on our analysis, the ancestral g2 gene possessed a structure similar to that of modern g2C. This gene was likely triplicated via tandem inversion (TID, Fig. 4). The evidence for that is the close relationship between the sequences of E, F and C genes that flank the cluster; their direct-reversed-direct position characteristic of the result of TID event; and the presence of a short palindromic sequence upstream of the g2 gene in *Xenopus*, which is necessary to facilitate such an event (Reams and Roth 2015). While the g2E gene is present in all species studied, there's a clear taxonomic bias concerning the preservation of the g2F or g2C gene. Only turtles and mammals have kept both and in the case of mammals the explanation for this may be the fact that the mammalian g2F gene has evolved a transmembrane domain and thus no longer encodes a secretory protein (Thul et al. 2017; Petryszak et al. 2016).

In other lineages, sequence similarity between C and F genes may have conferred functional redundancy leading to the elimination of either one or the other. Early in amniote evolution one of the ancestral EFC genes, likely the g2E of the Amniota MRCA based on its genomic position, duplicated to create the g2D gene which in turn spawned two or three additional copies (Fig. 3, although g2C is closer in sequence to g2D than g2E is). This is indicated by the fact that all extant lineages have at least one g2D gene and one or two differentiated D-clade genes.

# Expansion of the Pla2g2 cluster in Tetrapoda

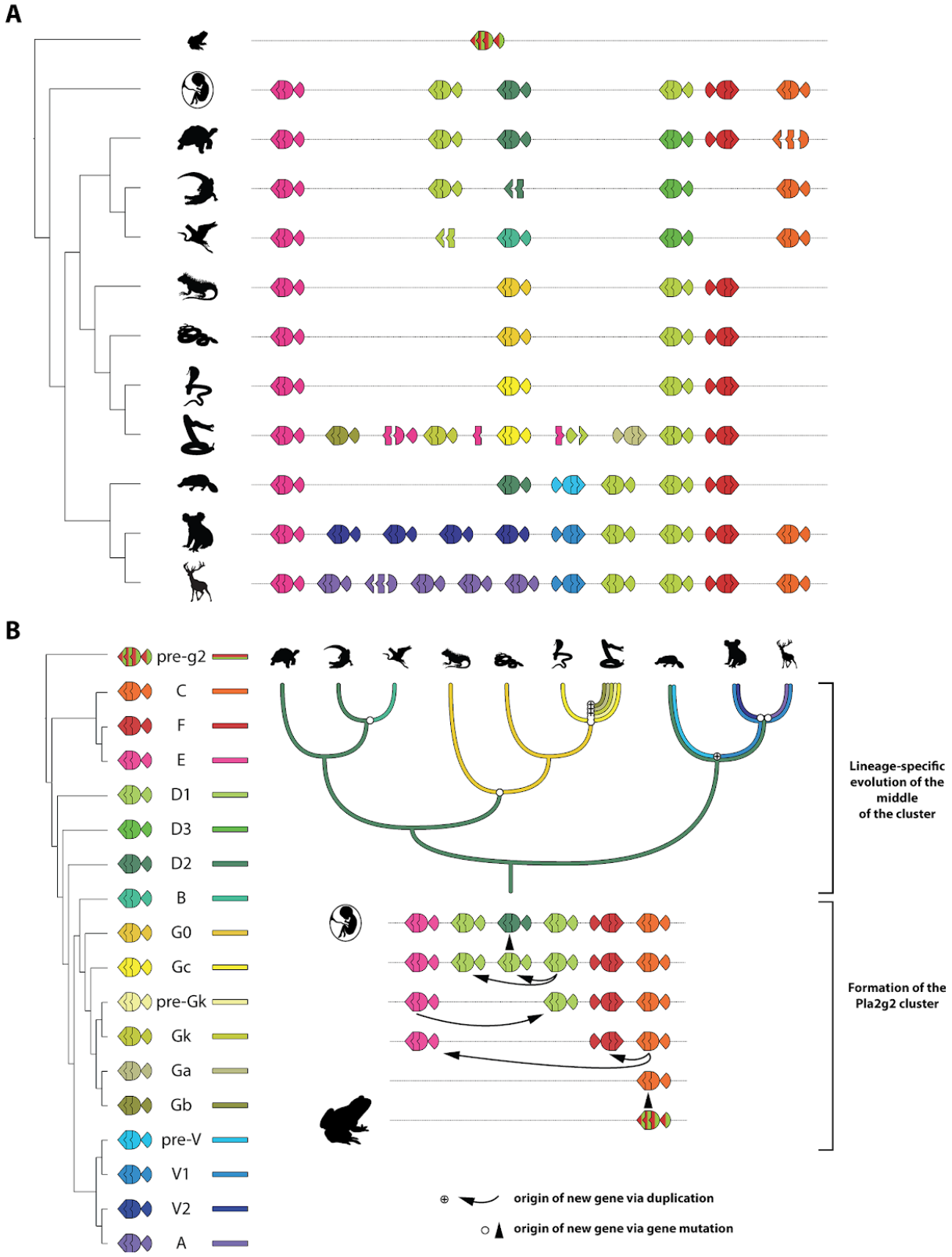


**Fig. 3. Pla2g2 cluster in Tetrapoda. A:** schematic representation of the cluster for each of the taxonomic lineages studied. While the g2E gene is present in all species studied, there's a clear taxonomic bias concerning the preservation of the g2F or g2C gene. Only turtles and mammals have kept both and in the case of mammals the explanation for this may be the fact that the mammalian g2F gene has evolved a transmembrane domain and thus no longer encodes a secretory protein (Thul et al. 2017; Petryszak et al. 2016). In other lineages, sequence similarity between C and F genes may have conferred functional redundancy leading to the elimination of either one or the other; **B:** Evolution of the cluster, initially illustrating the expansion events undergone following the split of Amniota from Amphibia (frog icon), prior to the evolution of the amniote MRCA (human embryo icon). Based on our analysis, the ancestral g2 gene possessed a structure similar to that of modern g2C. This gene was likely triplicated via tandem inversion (TID). Early in amniote evolution one of the ancestral EFC genes, likely the g2E, duplicated to create the g2D gene which in turn spawned two or three additional copies. The cladogram stemming from the g2D form (dark green) of the Amniota MRCA illustrates the subsequent lineage-specific expansions of g2D that have occurred within Amniota, as well as mutational origins of new genes that have occurred in the absence of duplication. The major differences in Pla2g2 clusters between different lineages of Amniota concern the evolution of new D-clade genes unique to each taxonomic lineage. All of these genes appear to be descendants of the same ancestral g2D2 gene, which is still present in a plesiomorphic form in crocodiles and turtles (SM3). The ancestral g2D2 appears to have undergone gene mutation independently in mammals, birds, and squamate reptiles. These mutated derivations of g2D2 are the ancestors of the g2V, g2B, and g2G clades, respectively. These mammalian (g2V), avian (g2B), and squamate reptilian (g2G) forms all have unique structures. The avian g2B protein has an N-terminal region unique to this gene that sets it apart from almost all other g2 proteins. Its gene is always present only in a single copy and its function remains unknown. The mammalian and squamate forms are virtually the only ones to have undergone expansion after the ancestral amniote duplication. According to selection analyses, both of these are evolving under the influence of positive selection (SM8) however both of them evolved their unique structures prior to duplication.

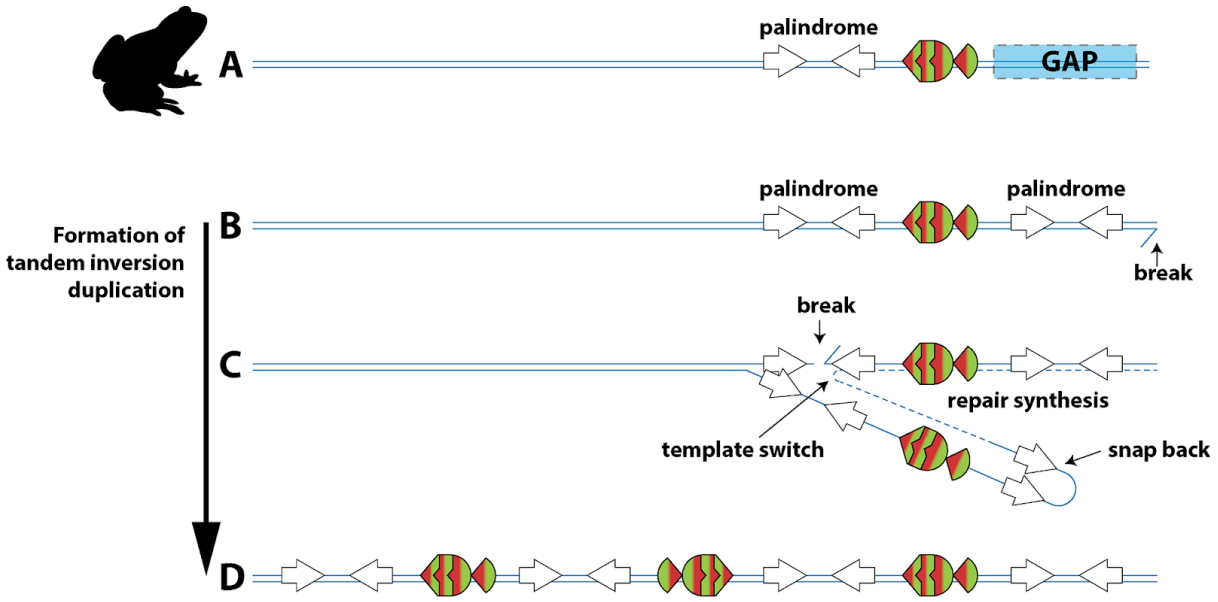



**Fig. 4. Illustration of the tandem-inversion duplication mechanism hypothesised to have created the EFC clade of Pla2g2 (adapted from** (Reams and Roth 2015)**. A** illustrates the extant condition of *Xenopus laevis* and *X. tropicalis,* where GAP indicates a sequencing gap present in published genomes of both species. In **B** and **C**, following a DNA break the presence of flanking palindromic sequences (parts of the transposon "Kolobok T2") results in template switching leading to the incidental multiplication of genes during the repair process. The final result, illustrated in D, is 3 genes, with two copies in the same orientation as the parent gene and one reversed.

**Expansion of D-clade genes is associated with lineage-specific neofunctionalization.** The major differences in Pla2g2 clusters between different lineages of Amniota concern the evolution of new D-clade genes unique to each taxonomic lineage. All of these genes appear to be descendants of the same ancestral g2D2 gene, which is still present in a plesiomorphic form in crocodiles and turtles (Fig. 3, SM3). The ancestral g2D2 appears to have undergone mutation independently in mammals, birds, and squamate reptiles (Fig. 3). These mutated derivations of g2D2 are the ancestors of the g2V, g2B, and g2G clades, respectively. These mammalian (g2V), avian (g2B), and squamate reptilian forms all have unique structures. The avian g2B gene has an N-terminal region unique to this gene that sets it apart from almost all other g2 genes. It is always present only in a single copy and its function remains unknown. The mammalian and squamate forms are virtually the only ones to have undergone expansion after the ancestral amniote duplication. According to selection analyses, both of these are evolving under the influence of positive selection (SM8) however both of them evolved their unique structures prior to duplication.

*Evolution of snake venom PLA2 genes*

The squamate reptile g2G gene is present in lizards and *Python molurus bivittatus* in a single copy. Since it is the ancestral (non-toxic) form, we have labelled it g2G0 (Fig. 1). In the ancestor of "advanced snakes" (Colubroidea – i.e. the shared ancestor of cobras and vipers), whilst remaining a single copy, it underwent structural change (becoming g2Gc) and this may have exapted (Gould and Vrba 1982) it for its subsequent functional recruitment into the venom arsenal of vipers. This form, unique to advanced snakes, was recovered from the genomes of the natricid snake *Thermophis baileyi,* the elapid snakes *Notechis scutatus*, *Ophiophagus hannah*, and *Pseudonaja textilis*, and all viperid species (i.e. all colubroid snakes with genomes of sufficient quality), indicating that it is likely a synapomorphy of the clade.

The colubroid ancestor was likely venomous (Jackson et al. 2017), and thus the potential exists at that early stage for positive selection acting upon genes encoding orally secreted toxins. It is unclear, however, whether g2Gc was in fact utilised as a venom toxin by early colubroid snakes and whether this function may have provided the selection pressure leading to the fixation of this form in the colubroid ancestor. Regardless, it is selectively expressed in the venom gland of extant viperid snakes (Aird et al. 2017), indicating that it was "recruited" in a venom gland in the common ancestor of viperids, if not before. The gene (which is 94% similar to the viperid form - Fig. 5) is not expressed in the *O. hannah* venom gland or accessory gland, indicating that it is not utilised as a toxin by this species; it is also expressed at extremely low levels in pooled tissues, which may be indicative of its incipient toxicity (Vonk et al. 2013). Whilst elapid snakes do utilise phospholipases as toxins, all known elapid Pla2 venom toxins are members of group 1, which is unrelated to group 2, the subject of the present study (Fry 2015). Group 1 Pla2s exhibit quite a different evolutionary pattern to group 2, and are the subject of a follow-up study (Koludarov et al., forthcoming).

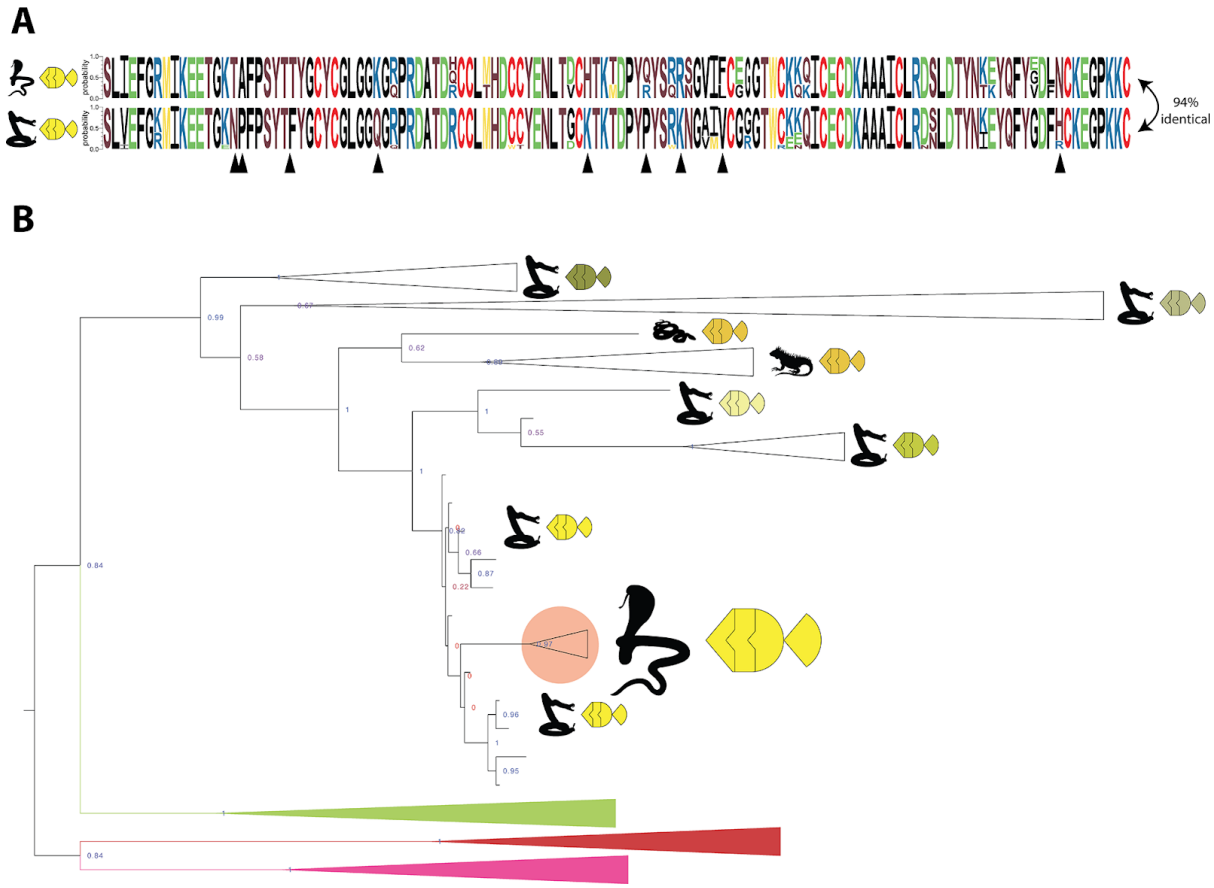



**Fig. 5. A:** Consensus sequences of elapid g2Gc and viperid g2Gc, with differences between them indicated by black triangles. Overall, the sequences are 94% similar. The size of the letters is proportional to the probability of a member of the lineage having it, while boldness of the letter is proportional to how often this position occurs between the species; **B:** Phylogenetic tree of squamate Pla2g2 genes, with the orange circle highlighting the nesting of elapid g2Gc sequences within viperid venom sequences, a consequence of the extreme similarity of these forms.

Pla2g2Gc is the founder member of the "G clade", which all viperid PLa2g2 toxins are members of. Given its presence as a single copy in the genomes of the other colubroid snakes (a clade which is ancestrally venomous), we have uncovered no evidence of duplication associated with the acquisition of a toxic function in venom for this gene. Thus we infer that this novel function arose *prior* to duplication, possibly via a shift in tissue-specific expression patterns which resulted in its selective expression in the venom gland. This remains an inference as we lack transitional forms within Viperidae, and duplication of this gene occurred sometime between the split of viperid snakes from the main stem of Colubroidea and the origin of the MRCA of extant Viperidae, which possessed additional copies (Fig. 6).

The alternative possibility is that duplication occurred prior to "recruitment" to the venom system, giving rise to the new gene g2Ga. This new gene's product, possessing by chance a greater toxicity than that of its parent gene (g2Gc), was selected for venom-gland-specific expression and the parent gene was co-expressed due to the co-regulation of neighbouring genes. This alternate scenario is further complicated by the fact that g2Gc, initially a passively co-expressed (unselected) gene in the venom system later evolves (in a Crotalinae specific derivation) into the myotoxic g2Gk. Thus this alternate hypothesis requires two "recruitment"

events - one (of g2Ga) for the initial addition to the venom arsenal, and a second one associated with the mutation of g2Gc into g2Gk. In either case, changes in gene expression, which are untraceable at this level of analysis (and possibly lost to the sands of time) are crucially important in the initial acquisition of the novel, toxic function. Given the presence of additional "random" (unselected) steps – which may also be viewed at the theoretical level as ad hoc assumptions – in the latter scenario (duplication precedes novel function), we prefer the former (novel function precedes duplication - see below for a more detailed discussion). However, additional research is required to definitively differentiate between these scenarios.

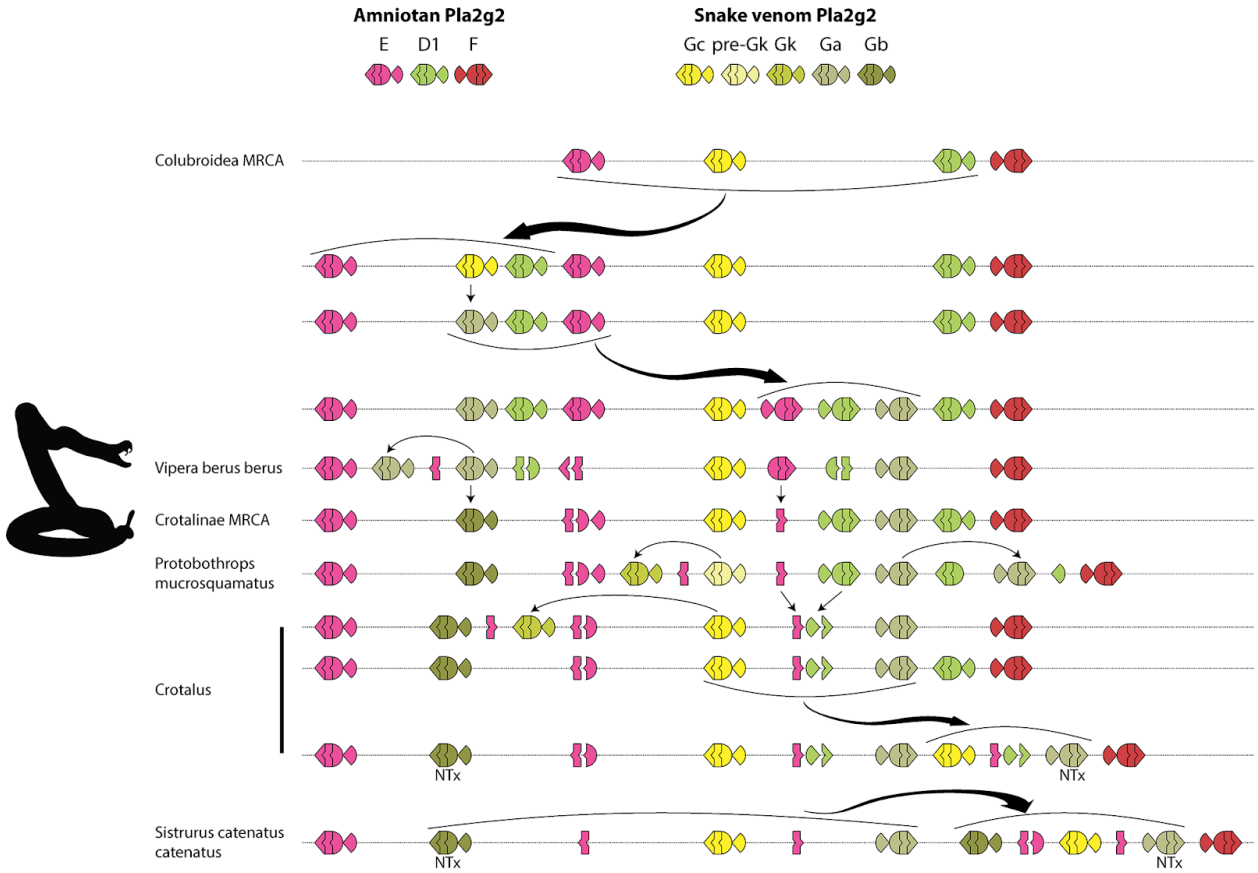



**Fig. 6. Evolution of Pla2g2 cluster in Viperidae.** In the colubroid snake MRCA, Pla2g2Gc exists in a single copy. This ancestral snake was likely venomous (B. Fry 2015; Jackson et al. 2017) though Pla2g2Gc may or may not have been deployed as a toxin at this stage. After the split of the viperid lineage, however, this gene was unequivocally recruited as a toxin. Subsequently the region containing it and g2E and g2D genes duplicated, giving rise first to the g2Ga (acidic) and after the second round of duplication to g2Gb (basic) venom Pla2s. Afterwards g2Gc experienced individual duplication events and gave rise to additional forms in various viperid snake genera, most notably myotoxic g2Gk49 and the intermediate form between it and plesiotypic g2Gc – pre-g2Gk49 that lacks key K49 mutation. Heterodimeric neurotoxins arise convergently in *Crotalus* and *Sistrurus* via distinct multiplication pathways. It is the presence of multiple fragments of "exonic debris" (mainly from g2E and g2D) that make possible the reconstruction of the evolutionary history, including all duplication events, of this genomic region in viperid snakes. It is important to note that most of the duplication events do not involve single genes (unlike g2A of mammals) but rather small groups which are duplicated as units, typically composed of g2G gene flanked by parts of g2E and g2D.

Viperid snakes diverged early from the main colubroid lineage (which includes elapid snakes; the front-fanged lamprophiids *Atractaspis* and *Homoroselaps*; and many non-front-fanged venomous species) and the most striking synapomorphy of the family is the possession of large, hollow fangs which are the sole tooth located on a mobile maxillary bone (B. G. Fry et al. 2012). These fangs, like those of other front-fanged snakes, are connected to the venom gland by an enclosed duct, and the gland itself is surrounded by compressor musculature which contracts during venom delivery. Thus viperids are in possession of a "high-pressure" venom delivery system and, moreover, were the first lineage of snakes in which such a system evolved. That Pla2g2Gc apparently only became specialised for use as a venom toxin after the divergence of Viperidae from other advanced snakes suggests that the acquisition of this function may have been associated with the evolution of a delivery system capable of inoculating the toxin directly into the muscle tissue of potential prey organisms. This hypothesis is consistent with the subsequent diversification of the subfamily in viperid snakes, including the evolution of specialised myotoxic and presynaptically neurotoxic forms, which would be more effective if delivered intramuscularly – a feat that non-front-fanged snakes, and even many front-fanged elapid snakes, are unlikely to be capable of.

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. In viper venoms these forms are more abundant than the plesiotypic g2Gc form (Aird et al., 2017, SM6, SM7). Subsequently, g2G venom genes were duplicated in several lineages independently and via different mechanisms, to produce genes that became subunits of heterodimeric neurotoxins in several *Crotalus* and *Sistrurus* species (Fig. 6). The heterodimeric neurotoxins thus arose independently in these two genera (cf. (Dowell et al. 2016), an example of convergent evolution explained 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 pre-g2Gk49 form in Crotalinae (pit vipers), and an additional duplication of this form became the non-catalytic myotoxin (g2Gk49) (Fig. 6).

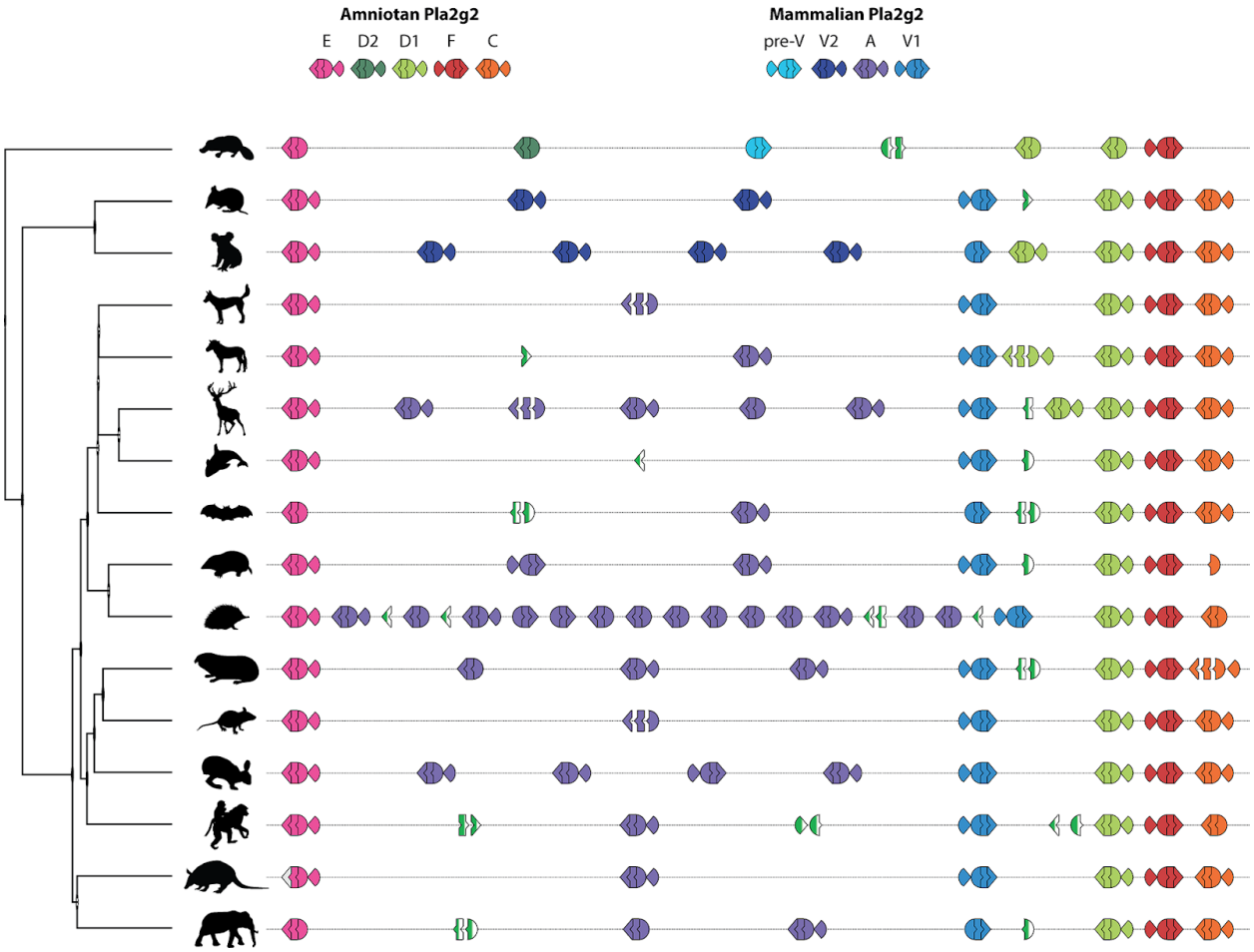



**Fig. 7. The g2V clade exhibits a convoluted recent evolutionary history in mammals.** An early duplication of g2D produced "pre-g2V" (present in the platypus - Monotremata) and the gene that later became g2V1 (a.k.a. Pla2g5), which is present in both placental (Eutheria) and marsupial (Metatheria) mammals. While g2V1 is always present in a single copy, pre-g2V evolved into a form similar to g2V2 of marsupials and then underwent multiple independent expansions (also: Fig. 3, SM3). In marsupials, this gene exists in several copies, all of which structurally resemble an intermediate form between g2V1 and g2A of placentals. In placentals, the same gene evolved into the g2A form which then underwent lineage-dependent multiplication at the level of individual families and species, with four shared copies in Bovidae and 14 copies in the hedgehog. However, several mammalian families retained a single-copy ancestral state and in some species (or in some strains, like the one for which reference genome of the mouse was made) the gene has been rendered non-functional. Particularly noteworthy is the hedgehog, with 14 copies of g2A, for many of which a signal peptide encoding exon was undetectable (exon 1) and thus they may be non-functional, though this may also be an artefact of the stringency of our search parameters.

The g2V clade exhibits a similarly convoluted recent evolutionary history in mammals. An early duplication of g2D produced "pre-g2V" (present in the platypus - Monotremata) and the gene that later became g2V1 (a.k.a. Pla2g5), which is present in both placental (Eutheria) and marsupial (Metatheria) mammals. While g2V1 is always present in a single copy, pre-g2V evolved into a form similar to g2V2 of marsupials and then underwent multiple independent expansions (Fig. 3, Fig. 7, SM3). In marsupials, this gene exists in several copies, all of which structurally resemble an intermediate form between g2V1 and g2A of placentals. In placentals, the same gene evolved into the g2A form which then underwent lineage-dependent multiplication at the level of individual families and species, with four shared copies in Bovidae

and 14 copies in the hedgehog (Fig. 6). However, several mammalian families retained a single-copy ancestral state and in some species (or in some strains, like the one for which reference genome of the mouse was made) the gene has been rendered non-functional.

Thus, a single ancestral gene, g2D2, evolves into a new structural form independently in squamates, birds and mammals. In all the three cases (g2B, g2G and g2V), these genes evolve new protein structures without prior duplication. Whilst the avian g2B remains in a single copy, in squamate reptiles and mammals an expansion of the group takes place. Interestingly, g2A and g2G genes aren't just the only Pla2g2 genes to multiplicate, but also the only ones that show evidence of evolving under the influence of positive selection (SM8). This pattern suggests that the acquisition of a novel activity, associated with structural change, perhaps in concert with an appropriate pattern of tissue-specific expression, was the change that facilitated the accumulation of duplicates at this locus. Subsequent to the expansion of these gene networks, the locus became a neofunctionalization hotspot, particularly within viperid snakes.

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

Our results indicate that duplication is not necessary for the acquisition of a novel function, a conclusion most clearly supported in the case of the g2Gc gene of colubroid snakes. Whether or not this gene's product was first deployed as a venom toxin in an ancestral colubroid or an ancestral viper, duplication does not appear to have been a prerequisite for the acquisition of this novel "exochemical" (or "exophysiological", being deployed to work outside of organism's inner "chemistry") function.

It is well-documented that duplication of existing toxin genes can drive changes in gene expression, and provide raw material for future evolution ((Margres et al. 2016)). Rather than duplication being a prerequisite, 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.

Change of context may occur in multiple ways:

- following a change in expression pattern that sees a gene being expressed in a novel tissue;
- 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);
- 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 in 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; B. Fry 2015), Pla2g2 genes may have played 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, g2V in mammals and g2G in squamates. The role of antimicrobial mammalian V-clade genes in the innate immune response is well known (Nevalainen, Graham, and Scott 2008). Amongst squamate G-clade genes, however, only viperid snake toxins forms are well-studied and the function of avian g2B is unclear.

The pattern we have observed, in which both the emergence of novel functions and subsequent gene family expansion take place at a single locus in distantly related taxa, suggests that such loci have a deep ancestral propensity for mutation and duplication, or at least for their subsequent preservation and fixation. The propensity for duplication is likely determined by genomic structure, as it is well-understood that 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, that duplications occur continuously in such regions but that all resultant genes are deleted seems implausible. This is for two reasons: 1) because exonic/intronic debris is typically evident following deletion (unless the deletions are extremely ancient events); and 2) because down-regulation ("dosage sharing") or silencing with methylation may facilitate the long-term preservation of segmental duplications in genomes despite the predictions of the dosage balance hypothesis (Assis and Bachtrog 2015; Lan and Pritchard 2016; Guschanski, Warnefors, and Kaessmann 2017). Another alternative is that individuals in which deleterious duplications occur are strongly selected against and thus no evidence of these duplications persists in sequenced genomes, but this also seems an unnecessarily extreme speculation as it requires that such duplications be immediately and invariably lethal.

By casting the net widely, we have been able to detect a pattern that does not conform to any one of the common theoretical models for duplicate preservation, but rather subsumes several of them into a temporal series. The single model our results most closely resemble is "subfunctionalization" (Force et al. 1999), however it contains additional processes (e.g. "moonlighting" and "neofunctionalization") not described by that model and may or may not include a period of "degeneration" (see below for further discussion).

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

The initial acquisition of a novel function may occur i) when noisy expression patterns (leaky transcripts) instigate a moonlighting scenario – a single copy gene fulfilling multiple functions by virtue of expression in multiple locations (Copley 2014); or ii) when structural change facilitates interaction with novel partners, whilst maintaining the ancestral function. The novel function may then expose the gene to a distinct selection regime, which may facilitate the accumulation and fixation of further mutations. When a novel function is acquired by a single copy gene, this

may create pressure for the creation of duplicate copies such that the multiple functions can be segregated between those copies, which may then specialise.

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

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

Subsequent to initial duplication, specialisation (a.k.a. “escape from adaptive constraint” – (Hughes 1994; Innan and Kondrashov 2010) may occur, in which one copy of the gene maintains the original function and the other specialises for its exochemical role, e.g. a role in venom in viperid snakes. This specialisation may facilitate a tissue-specific pattern of expression – although it has been suggested that the expression of tandem duplicates is likely to be co-regulated until one copy undergoes chromosomal displacement (Lan and Pritchard 2016) available expression data clearly indicate that Pla2g2G are highly tissue-specific in their expression and that neighbouring genes (Pla2g2E and Pla2g2D) are not expressed in the venom gland (SM7 and (Aird et al. 2017; Vonk et al. 2013).

This specialisation may lead to increased selection on dosage, driving the accumulation of duplicate genes now specifically expressed within the exochemical system. This is particularly likely for systems in which more gene product is “better”, either leading to a more toxic venom (Margres et al. 2016) or more effective response to infection. At this point, classic Ohno-style redundancy occurs, as multiple gene copies represent both a larger target for mutational change (and thus a network for exploring phenotype space) and each becomes less constrained by purifying selection (Aird et al. 2017). This in turn leads to neofunctionalization, in Ohno’s sense of the term, in which specific gene copies evolve interactions with novel partners.

The aforementioned sequence describes a model (and a hypothesis in need of testing) that loosely subsumes moonlighting, specialization/subfunctionalization and neofunctionalization into a single temporal series. Models in which duplication is central to the evolution of functional novelty have dominated discussion in recent years, but assertions that functional novelty may often precede duplication are nothing new. Indeed, they date back at least to the

work of Serebrovsky (1938, referenced in (Taylor and Raes 2004), who discussed the pleiotropic effects of a single gene being distributed between daughter genes following duplication. More recently, Hughes (Hughes 1994) explicitly states that a period of gene sharing precedes duplication-facilitated specialization. Whether these models, or that which we have outlined in the previous paragraph, should be considered "subfunctionalization" in the sense of Force et al. (Force et al. 1999) is perhaps a moot point. The formal subfunctionalization model includes "degeneration" (of regulatory elements and/or functional structures) following duplication. Whilst this may occur, the significant consequence of it, particularly in terms of venom toxins, appears to be "escape from adaptive constraint" (Hughes 1994), which in turn leads to neofunctionalization proper (Ohno 1970). This pattern conforms with the analyses of Assis and Bachtrog (Assis and Bachtrog 2015), who demonstrated that subfunctionalization was rare in comparison to conservation, specialization, or neofunctionalization, and indicated that subfunctionalization may be merely a stage in the evolutionary series leading towards neofunctionalization.

In any case, formal models are rarely more than schematics, and there is little reason to expect real world sequences to conform to them precisely. Thus, whilst we do not believe we have reconstructed a history that conforms to rigorously defined "subfunctionalization", clearly that history resembles this model, just as it resembles elements of several others. Hargreaves et al. (Hargreaves et al. 2014) previously argued that venom toxins likely acquire their toxic functions via subfunctionalization rather than neofunctionalization. In this they were making a point of difference with much of the molecular evolutionary work done in the field of toxinology, in which it had been previously well accepted that Ohno-style neofunctionalization was the dominant process of protein "weaponisation". We agree that Ohno's model does not account for all the details, but (as described above) feel that it describes an important part of the process characteristic of certain venom toxin families, namely the expansion of these families via duplication and the attendant evolution of multiple novel functions. We further recommend that the term "neofunctionalization" not be too narrowly defined, as it, etymologically, merely refers to the origin of novel functions. Ohno's initial coinage was a catchy one and we would like the usage of this term to be legitimate, despite the fact that in its narrow definition is does not capture all the details. Those that have read Ohno's monumental publication of 1970 (Ohno 1970), know that his thought was expansive and that he described processes akin to subfunctionalization working alongside the neofunctionalization for which he is remembered. In this sense he was like Darwin, whose thoughts on evolution extended beyond Natural Selection and the conceptual tools of what became, in the 20th Century, Neo-Darwinism. Thus "Darwinism" is more expansive than "Neo-Darwinism" and "neofunctionalization" may be legitimately considered more expansive than its formal definition suggests.

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

*A note on nomenclature and the history of Pla2g2 research*

Despite the fact that research on Pla2s was initiated in the early 1900s (with studies of cobra venom and pancreatic juice), the nomenclature of this family wasn't formalised until the late 1990s, when many mammalian forms were cloned and studied. The earliest attempt at a systematic Pla2 nomenclature was that of Heinrikson et al. (Heinrikson, Krueger, and Keim 1977), who compared all available Pla2 sequences and, based on their structural features, proposed to lump mammalian pancreatic and elapid snake venom Pla2s together as Group 1 and viperid snake venom Pla2s as Group 2. Later, Joubert et al. (Joubert, Townshend, and Botes 1983) proposed to split Group 2 into g2A and g2B, the former of which included all known viperid sequences with the sole exclusion of a *Bitis gabonica* (gaboon viper) Pla2 which was classified as the sole member of g2B. Later still, g2A was expanded to include mammalian synovial Pla2, when Davidson and Dennis (Davidson and Dennis 1990) made the first-ever Pla2 phylogeny using 40 protein sequences. Due to computational restrictions, they trimmed their dataset to include only one protein sequence per species, and their nomenclatural conclusions may have differed had their dataset included all sequences available at the time.

By the end of the 20th century all mammalian Pla2g2 subgroups had been discovered (Chen et al. 1994; Ishizaki et al. 1999; Valentin et al. 1999), and they received their names as a continuation of the g2A and g2B series: g2C, g2D, g2E, g2F. The only obvious exclusion was so-called "Group 5", that owed its special status to a reduced number of disulfide bonds (6 instead of 7) and the lack of non-mammalian, non-squamate Pla2g2 sequences that routinely share this feature (SM1-3). Later it was discovered that "Group 5" is located within the chromosomal loci occupied by other Group 2 Pla2 genes in humans, but the nomenclatural distinction persisted.

At the same time, venom researchers started to use "gA" and "gB" to mean "acidic venom Pla2s" and "basic venom Pla2s" (cf. (Whittington, Mason, and Rokyta 2018), which, given the historical grouping of viperid venom Pla2s and mammalian g2A was potentially confusing to researchers aiming to connect different kinds of Pla2s under one system. To address this issue, Dowell et al. (Dowell et al. 2016) proposed to name all viperid venom Pla2s "g2G", with an additional distinction between acidic, basic and other distinct lineages within this subgroup.

Since our study revealed more than 20 new Pla2g2 lineages (tripling in size the known number of subgroups), there was a need to resolve all conflicts within the nomenclature and not create any unnecessary conflicts of our own, while expanding it to include all Pla2g2s from the entire Vertebrata clade. In the interests of putting forth a system that takes into account the evolutionary relationships revealed in this study, we have taken the following steps (summarised in Fig. 1):

- Extension of g2E, g2F and g2C to include all non-mammalian homologs that clearly clustered with their mammalian counterparts both in terms of phylogenetic relationship and chromosomal position.
- Expansion of clade g2D to include all Pla2s that cluster together with mammalian g2D but are not experiencing duplication or visible change of structure (unlike mammalian g2V/g2A, bird g2B or squamate g2G). However, we used indices to mark the deep evolutionary splits within the group (g2D1, g2D2 and g2D3).

- As the grouping of mammalian g2A and viperid venom Pla2s together has long been recognised as dubious (Six and Dennis 2000) and taking into account recent suggestions to label venom Pla2g2s as g2G (Dowell et al. 2016), we decided to use g2A to mean exclusively mammalian Pla2g2 of the distinct clade (see phylogenetic trees in SM1 and SM4).
- Downgrading so-called "group 5" and placing it where it truly belongs, based on all available knowledge – as a part of the group 2, thus labelling it as g2V.
- Acknowledging the difference between those g2V that are present in both marsupials and placentals and those unique to marsupials, we decided to use g2V1 to mean the former and g2V2 to mean the latter. A unique Pla2g2 from platypus that seems to be basal to the entire g2V clade (g2V1, g2V2, g2A) thus received the name of pre-g2V.
- Because historical g2B, reserved solely for *Bitis gabonica*, has long been considered a misnomer (Six and Dennis 2000), and given the necessity to label a distinct clade of Pla2s from birds, the N-terminal region of which is dramatically different from all other Pla2g2 surveyed and is highly basic, we used g2B to include bird basic Pla2g2s.
- For the venom Pla2s, we have largely followed the nomenclature proposed by Dowell et al., only expanding it to include elapid Pla2g2s virtually indistinguishable from their viperid homologues, as well as non-venom (to the best of our knowledge) Pla2g2s from lizards, since they cluster together with g2G. The latter got the name of g2G0 to reflect their incipient state.

## Conclusion

By avoiding error-prone bioinformatic annotation pipelines and utilising a labour-intensive manual re-annotation method for genomic regions of interest across more than 90 genomes, we have been able to reconstruct the evolutionary history of the Pla2g2 gene family in unprecedented detail. We have thus contributed qualitatively to our knowledge of the evolution of this gene family, as well as developed a new method that can be applied to any gene family exhibiting copy number variation and located in a genomic region with a moderate to high level of synteny. 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 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 novel functionality, a process we refer to 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 more rigorous 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 and we may have to acknowledge that there can be no "one size fits all" model for the evolution of functional novelty, and that it's rather a case of "whatever can happen, will happen". Lastly, we have done our best to resolve much of the

confusion that has infused the nomenclature of phospholipases within this family, and sincerely hope that our efforts indeed serve to reduce that confusion, rather than compound it.

## Materials and Methods (Fig. 8)

We used published annotations to find genomic sequences that corresponded to OTUD3-UBXN10 region. When no annotations were available, we used the BLAST feature of ncbi-blast v.2.7.1+ suite to find them, using known sequences as a search database. We used *Protobothrops* and *Crotalus* genomic sequences as starting points and traced homologous regions in non-snake reptiles (lizards, turtles, alligators, birds) as well as mammals based on synteny of the flanking genes. We confirmed and extended the knowledge that snake venom Pla2g2 genes evolve in a highly conserved genomic region that has undergone very little rearrangement across the entire Vertebrata clade.

We extracted exons that corresponded to Pla2g2 genes according to published annotations and then used BLAST (blastn, e-value of 0.05, default restrictions on word count and gaps) to determine homology of exons. This step was necessary, since many previously ab-initio annotated Pla2g2 genes have more than 5 exons, in some cases – up to 15. As expected, this was an annotation-related artefact and in the final analysis no gene had more than 5 exons. By removing all unique exons, we created the initial exon database that was used to search genomic sequences of all species in this study.

This uncovered exons that were absent from published annotations, and by including those newly found sequences in the search database we refined it and repeated the search using tblastx function of ncbi-blast suite with e-value cutoff of 0.01. This process was repeated until no new exons were discovered. We then manually assessed each result and established exon boundaries using Geneious v11, relying on previously existing transcriptome-verified exon annotations wherever possible. We paid close attention to variations in exon boundaries between the groups of Pla2s and between taxonomic lineages. We were extremely conservative in our predictions and discarded any annotation that had potential frameshift-inducing mutations or otherwise didn't have the structure of a full Pla2g2 exon. Whether the predicted exons are actually transcribed cannot be confirmed without an additional transcriptomic proof. However this issue is largely irrelevant for the purposes of our analysis, which is based on sequence homology and order, rather than whether a sequence is transcribed or not.

Since all previously described Pla2g2 genes have 3 exons that encode the mature protein, we considered triplets of those exons (labelled as 2, 3 and 4 respectively) as a separate Pla2g2 gene if they were located in close proximity to each other. Exons of the 1-type that encode the N-terminal region of the signal peptide proved much harder to locate, and were often present in many copies in a tandem-repeat fashion. This was especially the case in snake genomes with some genes having up to 4 1-type exons, making it impossible to use the full CDR for analysis since it wasn't feasible to tell which of those exons is the one present in mRNA.

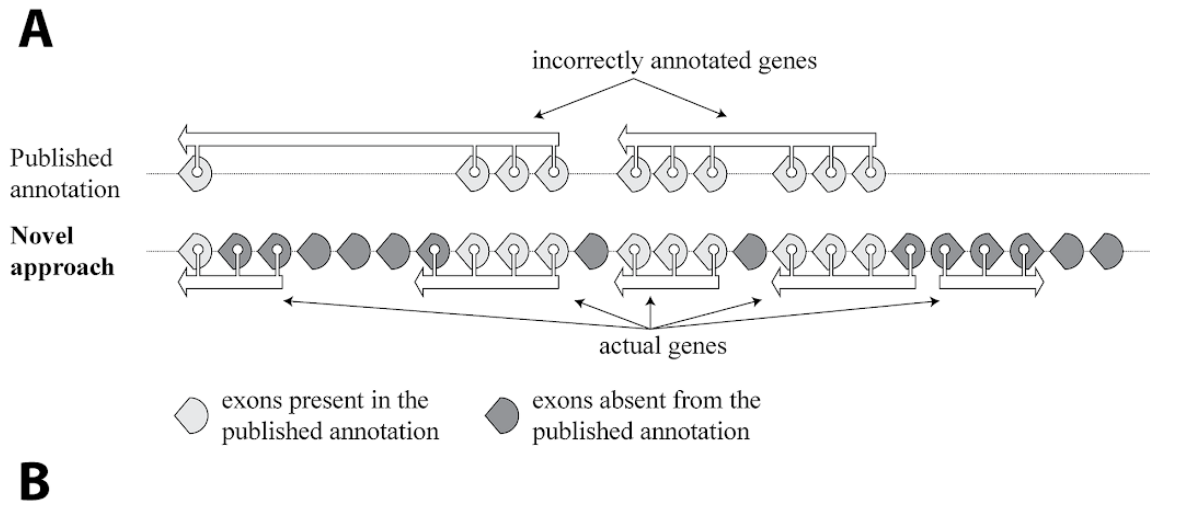

A
incorrectly annotated genes
Published annotation
Novel approach
actual genes
exons present in the published annotation
exons absent from the published annotation
B
1. Isolating all the annotated Pla2g2 genes to form initial dataset
gene as present in the annotation
individual exons
2. Blasting their exons against each other to figure out groups of homology between them
unique exons
groups of homologous exons
3. A typical Pla2g2 gene has four exons
mature peptide
signal peptide
complete Pla2g2 gene
4. Blasting genomic regions using a refined database containing only homologous exons
5. Using genomic position and gene tree to establish gene homology across the entire dataset

**Fig. 8. A:** Side by side comparison of the novel approach and a published annotation (region from *Chelonia mydas* genome); **B:** Schematic representation of the novel approach to genomic annotations developed and implemented in this study. For sequence IDs see SM2, published annotations accessed last time on 10.2018 and therefore this figure may not represent their current state.

It is also worth noting that quality of the initial assembly plays an important role in gene prediction. Many genomes have assembly gaps that may contain exons which may be functional parts of genes. In addition, some methods of assembling a genome are better than others as demonstrated for the *Boa* genome – 12 different assemblies of the same sequencing data resulted in genomes of varying quality ((Bradnam et al. 2013), see SM1-4 for comparison). Of 12 only 2 had complete sequences of all four g2 genes. Therefore the absence of some genes in other genomes could be a consequence of poor assembly.

Since all confirmed Pla2g2 genes have three exons coding mature peptide, we used those to establish individual genes. Then those individual genes were translated and mature peptides they encode used for phylogenetic analysis. The final dataset was trimmed to exclude sequences that might be pseudogenes, and included 442 sequences that we clustered into 17 distinct groups based on their protein sequence similarity and genomic position in respect to other genes. We used this information to further analyse all the partial or non-functional genes in the cluster to create a complete account of its evolutionary history. For the selection analyses we split the dataset by taxonomic clades to avoid saturation.

Protein alignment was done using localpair function of mafft software v7.305 (Katoh and Standley 2013) with 1000 iterations (--localpair --maxiterate 1000). Codon alignment for selection analysis was done using in-built Muscle aligner (v3.8.31, default settings) of AliView (Larsson 2014). In both cases, alignments were refined by hand to make sure that obviously homologous parts of the molecule (like cysteine backbone) are aligned properly.

Phylogenetic analysis was performed using exabayes v1.5 (Aberer, Kobert, and Stamatakis 2014) software with 10M generations of 4 runs and 4 chains running in parallel. Protein model was not specified, which allows software to iterate between different models, until the chains converge on the one that fits the best. Final consensus trees were generated with consense command and 25% burn-in. FigTree v1.4.3 was used to generate tree figures.

Selection was analysed with the use of FEL, MEME and RELAX tools of the HyPhy Datamonkey server (Weaver et al. 2018) (see SM8 for output files). Branch-site analysis was done in slimcodeml (Schabauer et al. 2012) (see SM8 for input and output files).

Transposable elements were annotated with the use of the Repeatmasker web server (Smit, Hubley, and Green 2016). Cross_match search engine and slow mode were selected for both higher sensitivity and for consistency with results from Dowell et al. (Dowell et al. 2016).

For the complete list of sequences used in the study see (SM2).

## Acknowledgements

IK would like to thank all the attendees of Venom Evolution, Function and Biomedical Applications Gordon Research Conference 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.

**References**


Aberer, Andre J., Kassian Kobert, and Alexandros Stamatakis. 2014. "ExaBayes: Massively Parallel Bayesian Tree Inference for the Whole-Genome Era." *Molecular Biology and Evolution* 31 (10): 2553–56.

Aird, Steven D., Jigyasa Arora, Agneesh Barua, Lijun Qiu, Kouki Terada, and Alexander S. Mikheyev. 2017. "Population Genomic Analysis of a Pitviper Reveals Microevolutionary Forces Underlying Venom Chemistry." *Genome Biology and Evolution* 9 (10): 2640–49.

Assis, Raquel, and Doris Bachtrog. 2015. "Rapid Divergence and Diversification of Mammalian Duplicate Gene Functions." *BMC Evolutionary Biology* 15 (July): 138.

Bergthorsson, Ulfar, Dan I. Andersson, and John R. Roth. 2007. "Ohno's Dilemma: Evolution of New Genes under Continuous Selection." *Proceedings of the National Academy of Sciences of the United States of America* 104 (43): 17004–9.

Birchler, James A., and Reiner A. Veitia. 2012. "Gene Balance Hypothesis: Connecting Issues of Dosage Sensitivity across Biological Disciplines." *Proceedings of the National Academy of Sciences of the United States of America* 109 (37): 14746–53.

Birts, Charles N., C. Howard Barton, and David C. Wilton. 2010. "Catalytic and Non-Catalytic Functions of Human IIA Phospholipase A2." *Trends in Biochemical Sciences* 35 (1): 28–35.

Bradnam, Keith R., Joseph N. Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, et al. 2013. "Assemblathon 2: Evaluating de Novo Methods of Genome Assembly in Three Vertebrate Species." *GigaScience* 2 (1): 10.

Chen, Ju, Sandra J. Engle, Jeffrey J. Seilhamer, and Jay A. Tischfield. 1994. "Cloning and Recombinant Expression of a Novel Human Low Molecular Weight Ca (2+)-Dependent Phospholipase A2." *The Journal of Biological Chemistry* 269 (4): 2365–68.

Conant, Gavin C., James A. Birchler, and J. Chris Pires. 2014. "Dosage, Duplication, and Diploidization: Clarifying the Interplay of Multiple Models for Duplicate Gene Evolution over Time." *Current Opinion in Plant Biology* 19 (June): 91–98.

Conant, Gavin C., and Kenneth H. Wolfe. 2008. "Turning a Hobby into a Job: How Duplicated Genes Find New Functions." *Nature Reviews. Genetics* 9 (12): 938–50.

Copley, Shelley D. 2014. "An Evolutionary Perspective on Protein Moonlighting." *Biochemical Society Transactions* 42 (6): 1684–91.

Danchin, Antoine, Christos Ouzounis, Taku Tokuyasu, and Jean-Daniel Zucker. 2018. "No Wisdom in the Crowd: Genome Annotation in the Era of Big Data - Current Status and Future Prospects." *Microbial Biotechnology* 11 (4): 588–605.

Davidson, F. F., and E. A. Dennis. 1990. "Evolutionary Relationships and Implications for the Regulation of Phospholipase A2 from Snake Venom to Human Secreted Forms." *Journal of Molecular Evolution* 31 (3): 228–38.

Dennett, Daniel C. 1995. "Darwin's Dangerous Idea." *The Sciences* 35 (3): 34–40.

Dowell, Noah L., Matt W. Giorgianni, Victoria A. Kassner, Jane E. Selegue, Elda E. Sanchez, and Sean B. Carroll. 2016. "The Deep Origin and Recent Loss of Venom Toxin Genes in Rattlesnakes." *Current Biology: CB* 26 (18): 2434–45.

Ekblom, Robert, and Jochen B. W. Wolf. 2014. "A Field Guide to Whole-Genome Sequencing, Assembly and Annotation." *Evolutionary Applications* 7 (9): 1026–42.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. "Preservation of

Duplicate Genes by Complementary, Degenerative Mutations." *Genetics* 151 (4): 1531–45.
Fry, Bryan. 2015. *Venomous Reptiles and Their Toxins: Evolution, Pathophysiology and Biodiscovery*. Oxford University Press.
Fry, Bryan G., Nicholas R. Casewell, Wolfgang Wüster, Nicolas Vidal, Bruce Young, and Timothy N. W. Jackson. 2012. "The Structural and Functional Diversification of the Toxicofera Reptile Venom System." *Toxicon: Official Journal of the International Society on Toxinology* 60 (4): 434–48.
Gould, Stephen Jay, and Elisabeth S. Vrba. 1982. "Exaptation—a Missing Term in the Science of Form." *Paleobiology* 8 (1): 4–15.
Guschanski, Katerina, Maria Warnefors, and Henrik Kaessmann. 2017. "The Evolution of Duplicate Gene Expression in Mammalian Organs." *Genome Research* 27 (9): 1461–74.
Guttinger, Stephan. 2018. *A Process Ontology for Macromolecular Biology*.
Hargreaves, Adam D., Martin T. Swain, Matthew J. Hegarty, Darren W. Logan, and John F. Mulley. 2014. "Restriction and Recruitment-Gene Duplication and the Origin and Evolution of Snake Venom Toxins." *Genome Biology and Evolution* 6 (8): 2088–95.
Heinrikson, R. L., E. T. Krueger, and P. S. Keim. 1977. "Amino Acid Sequence of Phospholipase A2-Alpha from the Venom of Crotalus Adamanteus. A New Classification of Phospholipases A2 Based upon Structural Determinants." *The Journal of Biological Chemistry* 252 (14): 4913–21.
Hughes, A. L. 1994. "The Evolution of Functionally Novel Proteins after Gene Duplication." *Proceedings. Biological Sciences / The Royal Society* 256 (1346): 119–24.
Innan, Hideki, and Fyodor Kondrashov. 2010. "The Evolution of Gene Duplications: Classifying and Distinguishing between Models." *Nature Reviews. Genetics* 11 (2): 97–108.
Ishizaki, Jun, Noriko Suzuki, Ken-Ichi Higashino, Yasunori Yokota, Takashi Ono, Keiko Kawamoto, Noriko Fujii, Hitoshi Arita, and Kohji Hanasaki. 1999. "Cloning and Characterization of Novel Mouse and Human Secretory Phospholipase A2s." *The Journal of Biological Chemistry* 274 (35): 24973–79.
Jackson, Timothy N. W., Bruce Young, Garth Underwood, Colin J. McCarthy, Elazar Kochva, Nicolas Vidal, Louise van der Weerd, et al. 2017. "Endless Forms Most Beautiful: The Evolution of Ophidian Oral Glands, Including the Venom System, and the Use of Appropriate Terminology for Homologous Structures." *Zoomorphology* 136 (1): 107–30.
Joubert, Francois J., Grace S. Townshend, and Dawie P. Botes. 1983. "Snake Venoms. Purification, Some Properties of Two Phospholipases A2(CM-I and CM-II) and the Amino-Acid Sequence of CM-II fromBitis nasicornis(Horned Adder) Venom." *Hoppe-Seyler´s Zeitschrift Für Physiologische Chemie* 364 (2): 1717–26.
Katoh, Kazutaka, and Daron M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4): 772–80.
Kini, R. Manjunatha. 2003. "Excitement Ahead: Structure, Function and Mechanism of Snake Venom Phospholipase A2 Enzymes." *Toxicon: Official Journal of the International Society on Toxinology* 42 (8): 827–40.
Klimke, William, Claire O'Donovan, Owen White, J. Rodney Brister, Karen Clark, Boris Fedorov, Ilene Mizrachi, Kim D. Pruitt, and Tatiana Tatusova. 2011. "Solving the Problem: Genome Annotation Standards before the Data Deluge." *Standards in Genomic Sciences* 5 (1): 168–93.

Lan, Xun, and Jonathan K. Pritchard. 2016. “Coregulation of Tandem Duplicate Genes Slows Evolution of Subfunctionalization in Mammals.” *Science* 352 (6288): 1009–13.

Larsson, Anders. 2014. “AliView: A Fast and Lightweight Alignment Viewer and Editor for Large Datasets.” *Bioinformatics* 30 (22): 3276–78.

Margres, Mark J., Alyssa T. Bigelow, Emily Moriarty Lemmon, Alan R. Lemmon, and Darin R. Rokyta. 2017. “Selection To Increase Expression, Not Sequence Diversity, Precedes Gene Family Origin and Expansion in Rattlesnake Venom.” *Genetics* 206 (3): 1569–80.

Margres, Mark J., Kenneth P. Wray, Margaret Seavy, James J. McGivern, Nathanael D. Herrera, and Darin R. Rokyta. 2016. “Expression Differentiation Is Constrained to Low-Expression Proteins over Ecological Timescales.” *Genetics* 202 (1): 273–83.

Nevalainen, Timo J. 2007. “Phospholipase A2 in Acute Pancreatitis: Review.” *American Journal of Surgery* 194 (4, Supplement): S28–32.

Nevalainen, Timo J., Garry G. Graham, and Kieran F. Scott. 2008. “Antibacterial Actions of Secreted Phospholipases A2. Review.” *Biochimica et Biophysica Acta* 1781 (1-2): 1–9.

Nobre, Tania, M. Doroteia Campos, Eva Lucic-Mercy, and Birgit Arnholdt-Schmitt. 2016. “Misannotation Awareness: A Tale of Two Gene-Groups.” *Frontiers in Plant Science* 7 (June): 868.

Ohno, Susumu. 1970. *Evolution by Gene Duplication*.

Petryszak, Robert, Maria Keays, Y. Amy Tang, Nuno A. Fonseca, Elisabet Barrera, Tony Burdett, Anja Füllgrabe, et al. 2016. “Expression Atlas Update--an Integrated Database of Gene and Protein Expression in Humans, Animals and Plants.” *Nucleic Acids Research* 44 (D1): D746–52.

Reams, Andrew B., and John R. Roth. 2015. “Mechanisms of Gene Duplication and Amplification.” *Cold Spring Harbor Perspectives in Biology* 7 (2): a016592.

Schabauer, H., M. Valle, C. Pacher, H. Stockinger, A. Stamatakis, M. Robinson-Rechavi, Z. Yang, and N. Salamin. 2012. “SlimCodeML: An Optimized Version of CodeML for the Branch-Site Model.” In *2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops PhD Forum*, 706–14.

Schnoes, Alexandra M., Shoshana D. Brown, Igor Dodevski, and Patricia C. Babbitt. 2009. “Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies.” *PLoS Computational Biology* 5 (12): e1000605.

Six, David A., and Edward A. Dennis. 2000. “The Expanding Superfamily of Phospholipase A2 Enzymes: Classification and Characterization.” *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1488 (1): 1–19.

Smit, A. F. A., R. Hubley, and P. Green. 2016. “RepeatMasker Open-4.0. 2015.” *Google Scholar*.

Taylor, John S., and Jeroen Raes. 2004. “Duplication and Divergence: The Evolution of New Genes and Old Ideas.” *Annual Review of Genetics* 38: 615–43.

Thul, Peter J., Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, et al. 2017. “A Subcellular Map of the Human Proteome.” *Science* 356 (6340). https://doi.org/10.1126/science.aal3321.

Valentin, E., F. Ghomashchi, M. H. Gelb, M. Lazdunski, and G. Lambeau. 1999. “On the Diversity of Secreted Phospholipases A(2). Cloning, Tissue Distribution, and Functional Expression of Two Novel Mouse Group II Enzymes.” *The Journal of Biological Chemistry* 274 (44): 31195–202.

Van de Peer, Yves, Steven Maere, and Axel Meyer. 2010. “2R or Not 2R Is Not the Question

Anymore." *Nature Reviews. Genetics* 11 (2): 166.
Vonk, Freek J., Nicholas R. Casewell, Christiaan V. Henkel, Alysha M. Heimberg, Hans J. Jansen, Ryan J. R. McCleary, Harald M. E. Kerkkamp, et al. 2013. "The King Cobra Genome Reveals Dynamic Gene Evolution and Adaptation in the Snake Venom System." *Proceedings of the National Academy of Sciences of the United States of America* 110 (51): 20651–56.
Wang, Julia, Rami Al-Ouran, Yanhui Hu, Seon-Young Kim, Ying-Wooi Wan, Michael F. Wangler, Shinya Yamamoto, et al. 2017. "MARRVEL: Integration of Human and Model Organism Genetic Resources to Facilitate Functional Annotation of the Human Genome." *American Journal of Human Genetics* 100 (6): 843–53.
Weaver, Steven, Stephen D. Shank, Stephanie J. Spielman, Michael Li, Spencer V. Muse, and Sergei L. Kosakovsky Pond. 2018. "Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes." *Molecular Biology and Evolution*, January. https://doi.org/10.1093/molbev/msx335.
Whittington, A. Carl, Andrew J. Mason, and Darin R. Rokyta. 2018. "A Single Mutation Unlocks Cascading Exaptations in the Origin of a Potent Pitviper Neurotoxin." *Molecular Biology and Evolution* 35 (4): 887–98.
Wistow, G., and J. Piatigorsky. 1987. "Recruitment of Enzymes as Lens Structural Proteins." *Science* 236 (4808): 1554–56.
Yamaguchi, Kazuaki, Takahito Chijiwa, Naoki Ikeda, Hiroki Shibata, Yasuyuki Fukumaki, Naoko Oda-Ueda, Shosaku Hattori, and Motonori Ohno. 2014. "The Finding of a Group IIE Phospholipase A2 Gene in a Specified Segment of Protobothrops Flavoviridis Genome and Its Possible Evolutionary Relationship to Group IIA Phospholipase A2 Genes." *Toxins* 6 (12): 3471–87.
Yandell, Mark, and Daniel Ence. 2012. "A Beginner's Guide to Eukaryotic Genome Annotation." *Nature Reviews. Genetics* 13 (5): 329–42.
Zallot, Rémi, Katherine J. Harrison, Bryan Kolaczkowski, and Valérie de Crécy-Lagard. 2016. "Functional Annotations of Paralogs: A Blessing and a Curse." *Life* 6 (3). https://doi.org/10.3390/life6030039.
Zhang, Lei, Xu Cai, Jian Wu, Min Liu, Stefan Grob, Feng Cheng, Jianli Liang, et al. 2018. "Improved Brassica Rapa Reference Genome by Single-Molecule Sequencing and Chromosome Conformation Capture Technologies." *Horticulture Research* 5 (August): 50.