Clade-specific genes and the evolutionary origin of novelty; new tools in the toolkit

Longjun Wu (1) and J. David Lambert

Department of Biology
University of Rochester
Rochester NY, USA 14627

dlamber2@ur.rochester.edu

(1) Current address: Department of Ocean Science, Hong Kong University of Science and Technology, Clearwater Bay, Hong Kong SAR

Abstract

Clade-specific (a.k.a. lineage-specific) genes are very common and found at all taxonomic levels and in all clades examined. They can arise by duplication of previously existing genes, which can involve partial truncations or combinations with other protein domains or regulatory sequences. They can also evolve *de novo* from non-coding sequences, leading to potentially truly novel protein domains. Finally, since clade-specific genes are generally defined by lack of sequence homology with other proteins, they can also arise by sequence evolution that is rapid enough that previous sequence homology can no longer be detected. In such cases, where the rapid evolution is followed by constraint, we consider them to be ontologically non-novel but likely novel at a functional level. In general, clade-specific genes have received less attention from biologists but there are increasing numbers of fascinating examples of their roles in important traits. Here we review some selected recent examples, and argue that attention to clade-specific genes is an important corrective to the focus on the conserved developmental regulatory toolkit that has been the habit of evo-devo as a field. Finally, we discuss questions that arise about the evolution of clade-specific genes, and how these might be addressed by future studies. We highlight the hypothesis that clade-specific genes are more likely to be involved in synapomorphies that arose in the stem group where they appeared, compared to other genes.

Introduction

Before the cloning of *Hox* genes and other developmental regulatory genes from flies and then vertebrates starting in the mid-1980s, it was generally assumed that animals with profoundly different body plans and developmental modes would use different genes in development (reviewed Gilbert et al. 1996). One prominent example of this commonly held view was articulated by Mayr in 1966: "Much that has been learned about gene physiology makes it evident that the search for homologous genes is quite futile except in very close relatives'' (Mayr 1966). From this perspective, it would be assumed that understanding development in different animals would involve studying genes and proteins that were restricted to their clade.

This changed dramatically as one developmental regulator after another was found to be conserved between distantly related model organisms (e.g. Carrasco et al. 1984; McGinnis et al. 1984; reviewed in Gilbert et al. 1996). These findings led to one of the central conceptual foundations of evolutionary developmental biology—that, in general, animal genomes have the same basic set of developmental regulatory proteins. Known as the "shared toolkit" (Wilkins 2013), this observation has guided many important lines of investigation in the field, where the role of orthologous genes are compared in different species to understand how they participate in morphological evolution. More generally, the idea of the shared toolkit has influenced the intellectual development of the field, by focusing attention on a set of developmental regulators which are tacitly assumed to play central roles in the origin of novelty, either through regulatory or structural evolution. This view has been extremely successful in terms of the breadth and depth of mechanistic understanding that it has generated in the last three decades. Most importantly, it clearly focused attention on what was feasible to study, especially in the pre-genomic era—the genes that were similar enough between species to be cloned and recognized based on homology. It also found broad success because workers in different animals found a

shared language in the genes that were being interrogated in their respective systems, allowing broader interest in these results. Similarly, the fact that the human genome also contains the shared toolkit genes meant that biomedically focused funding agencies were at least somewhat more predisposed to support funding in non-vertebrate systems.

However, despite all these advantages and successes, the shared-toolkit paradigm leaves us blind to some potential sources and causes of evolutionary change. In this review we will describe a special kind of exception to the shared-toolkit–conserved clade-specific genes–and their role in the evolution of novelty. As full genome sequences from diverse organisms have accumulated, it has become clear that many genes are not shared across clades, and these novel genes are found at all phylogenetic levels, from species to phyla, to super-phyletic groups (reviewed in, for instance, Khalturin et al. 2009; Kaessmann 2010; Tautz and Domazet-Loso 2011). We refer to such genes as clade-specific genes, which is equivalent to the term lineage-specific genes. Novel genes can form in several different ways (described in more detail below, and reviewed extensively including (Chen et al. 2013; Klasberg et al. 2016; Rödelsperger et al. 2019)). In general, these modes of origin exist on a continuum, from simple gene duplications of broadly conserved gene families, where the functionality of the duplicate is similar or identical to the original parent locus, all the way to completely novel sequences that have no detectable homology to proteins in other clades. Intermediates between these two include domain shuffling, and duplication followed by partial deletion.

Gene duplication is extremely common in evolution, and has long been recognized to be an important source of raw material for evolutionary innovation (Ohno 1970; Lynch and Katju 2004). Simple gene duplications do create novel genes that can potentially become clade-specific, but these will nonetheless be part of the shared toolkit. Duplications can be more

complex, with partial truncations, domain shuffling, and the addition of completely novel protein sequence, for instance when a frameshift is introduced into part of a coding sequence. Genes that originate via these processes may have novel functional capacities that are not in the shared toolkit.

At the other end of the continuum are genes that are apparently completely novel—these are usually recognized because they do not have detectable homology outside the clade of interest. These too can be found at every phylogenetic level, but we have been especially interested in those that are older than species or genera, because their conservation since their putative origin implies purifying selection and thus functional importance. How do such completely novel genes arise? They are often assumed to arise *de novo* from non-coding sequence. This does happen, but it is important to acknowledge that since novel genes are usually defined by the lack of sequence homology to other genes outside of the clade of interest, they can also arise by rapid sequence evolution of a gene that is present in outgroups, such that the gene has no detectable homology outside the clade in question (see Weisman et al. 2020). While the latter scenario results in a gene that is not *sensu strictu* novel, we argue that this level of sequence divergence can create a gene with the potential for novel function. This is especially true if the period of rapid evolution is followed by relatively slow sequence evolution, indicating purifying selection.

Finally, although it is conceptually somewhat different, clade-specific genes can also arise by lateral gene transfer from a distantly related clade (Chen et al. 2013; Rödelsperger et al. 2019). In this case, the domain is not strictly limited to the clade that acquired the gene, but it may be found in one group of animals and not others, for instance, and thus potentially involved in clade-specific novelties in that clade.

Often novel clade-specific genes arise by various combinations of these different processes; one result of this is that a clade-specific protein sequence may have some domains that are not detectable outside the clade and some that are. In this review, we are focused on putative clade-specific genes that have at least one protein domain that is apparently novel, since those are most likely to represent a gene whose function is outside of the toolkit. Moreover, these are also relatively easier to find in a bioinformatic screen using homology-based search pipelines. Below, we highlight some recent cases where clade-specific genes have been implicated in novel traits that arose at the same phylogenetic node as the novel gene. As expected if this is a pervasive evolutionary phenomena, they are diverse in terms of taxa, evolutionary age, aspects of biology, mechanism of origin, and levels of biological organization. Because of this broad diversity and disparity, any organization would be somewhat arbitrary; we have organized them roughly by evolutionary age, from genes that are specific to a metazoan super-phyletic group that arose in the Cambrian, to genes that are implicated in the evolution of modern humans. The main examples and their modes of origin are summarized in Table 1 and Figure 1.

Novel protein domains implicated in clade-specific traits of an animal superphylum

We became interested in the role of clade-specific genes in the origin of novelty after our recent discovery of two novel genes that are expressed in the ciliary bands of diverse spiralian taxa (Wu et al. 2020). The Spiralia is an ancient and diverse clade of invertebrate protostome animals including molluscs, annelid worms, nemertean worms, flatworms and rotifers. One prominent trait in most spiralians are ciliary bands. These bands not only have important functions such as

locomotion and feeding but also are key characters in phylogenetic and taxonomic discussions. In a bioinformatic screen for sequences that were present in multiple spiralian genomes, but not detectable in other animal genomes, we found 37 such genes. We examined the expression of 20 in a model mollusc embryo, and discovered two genes with specific expression in the larval ciliary bands of this species. Based on comparative studies from five different spiralian phyla, we and our collaborators showed that these two genes are broadly specific to ciliary bands in these other taxa as well. We named these genes *Trochin* and *Lophotrochin*.

Trochin has no detectable sequence similarity to any non-spiralian genes or protein domains even using very sensitive search methods (*i.e.* multiple rounds of PSI-BLAST or using HMMER), and thus appears to be the result of *de novo* gene formation or rapid evolution in the spiralian ancestor. Sensitive search algorithms showed that part of *Lophotrochin* has similarity to an uncharacterized domain found in some non-spiralian genes (DUF4476 domain), but a motif in the C-terminal part of the protein is specific to spiralians and strongly conserved. This indicates that *Lophotrochin* evolved from a DUF4476 domain-containing protein in the spiralian common ancestor that underwent rapid evolution to generate the new C-terminal spiralian-specific motif, or a fusion event with the novel C-terminal motif. For both genes, the strong purifying selection on the sequence over the more than 500 MY since the spiralian common ancestor in the Cambrian indicates that these genes have significant functional roles. The conservation of expression pattern in ciliary bands indicates a function in this key spiralian structure. This finding highlighted for us the importance of clade-specific genes or protein domains in the evolution of major clade-specific traits.

Clade-specific genes in ancient novel cell types of Cnidarians

An even older animal group than Spiralians are the Cnidarians, a large and diverse phylum that predate the origin of bilaterian animals. Perhaps one of the best characterized instances of clade-specific genes and novel traits involve the cnidocyte, a novel cell type that is the hallmark of cnidarians. Cnidocytes, or stinging cells, are one of the most specialized cell types in the Metazoa, and a classic example of a complex clade-specific trait. They contain an organelle called the nematocyst: when discharged this structure rapidly ejects a harpoon-like filament that can pierce prey and inject toxin (David et al. 2008). The discharge is thought to be one of the fastest cellular processes known (Holstein and Tardent 1984).

The importance of novel proteins in cnidocyte structure and function has long been appreciated (e.g. Kurz et al. 1991; Koch et al. 1998), and this cell type has become a leading model for understanding the role of novel proteins in a complex novelty (reviewed in Khalturin et al. 2009; Babonis and Martindale 2014). Several recent studies have systematically looked for cnidocyte-specific genes, and found that many are restricted to cnidarians. For instance Hwang et al. 2007 found that 24/51 cnidocyte-specific genes in *Hydra* were apparently cnidarian specific (*i.e.* found in *Hydra* and the distantly related *Nematostella* but not outside cnidarians); this includes those with some domains that were detectable outside the phylum and some that were not. Similarly, Milde et al. 2009 found a set of 5 putatively clade-specific and cnidocyte-specific genes in the genus *Hydra*, of which 2 were found in *Nematostella*, indicating that they are possibly cnidarian-specific. Hwang and colleagues 2008 and 2010 showed that two cnidarian-specific cnidocyte proteins, nematocilin and nematogalactin are localized to the tubule structure of the nematocyst itself. Having these genes in this key structure of the cnidocytes implicates clade-specific genes in the evolution of this novel cnidarian cell type.

Another cnidarian-specific trait is striated muscle, which is thought to have evolved convergently in bilaterians, cnidarians and ctenophores (Steinmetz et al. 2012). Within cnidarians, the Medusozoa, which consists of groups with medusae, or jellyfish, generally have striated muscles, and the main group lacking medusae (Anthozoa) does not, except for two species that are reported to have sarcomeres (Leclere and Rottinger 2017). This pattern suggests that striated muscles may have evolved at the base of the Medusozoa, perhaps to drive the strong muscle contraction of the swimming medusa. Khalturin et al. 2019 reported two medusozoan-specific gene families that are restricted to striated muscles; one is a WD-40 domain protein family and the other is a family with novel and myosin tail domain structure, indicating possible interactions with actin fibers. The authors suggest that they may play a structural role in the striated muscle, perhaps similar to titin and troponin, two bilaterian striated muscle proteins that are not found in cnidarian striated muscle.

There are other examples of clade specific genes in novel cell types. The bacteriocytes of aphids evolved to host the mutualistic endosymbiotic bacteria *Buchnera sp.* (Shigenobu and Stern 2013). Deep mRNA sequencing of bacteriocytes revealed an enrichment of aphid-specific genes in bacteriocytes: 10/30 of the most strongly expressed genes in bacteriocytes were aphid-specific. They concluded that these genes play a role in the bacteriome, and could help mediate the endosymbiosis with *Buchnera*, an important ecological adaptation.

Clade-specific genes in key synapomorphies of vertebrate subgroups

The vertebrate lineage is well-studied in terms of the mechanisms and genetic bases of important innovations, and this clade has been thoroughly sampled by full-genome sequencing efforts.

Together, these observations suggest that this clade may be a particularly good place to find known examples of clade-specific genes that are involved in traits that arose at the same evolutionary node–and this is indeed the case. The myelinated fiber is a hallmark of gnathostome vertebrate nervous systems. It is an axon that is sheathed in lipid-rich processes from glial cells, and this covering acts as electrical insulation that improves the speed and efficiency of action potentials in the axon. Myelinated fibers are absent in all invertebrates, as well as the agnathan vertebrates, but found in all jawed vertebrates (Gnathostomes), indicating that this trait arose at the base of this clade. Myelin largely consists of lipids and proteins. The dominant myelin proteins in all gnathostome peripheral nervous system (PNS) fibers are myelin basic protein (MBP) and myelin protein zero (MPZ). In tetrapods, these remain the dominant proteins in the PNS myelin, but in the central nervous system (CNS), MPZ was replaced by the Proteolipid protein-1 protein (PLP1). Gould et al. 2008 reported that MBP and MPZ are ubiquitous among gnathostomes, but absent in all invertebrates and agnathans examined and concluded that these genes likely arose in early gnathostomes coordinately with myelination. *PLP1* is related to the lipophilin gene family which is widespread across metazoans. However, a novel exon arose in PLP1 at the base of the tetrapods, suggesting that the origin of this protein sequence could have allowed PLP1 to replace MPZ in myelin of the central nervous system (Gould et al. 2008; Mobius et al. 2008). Indeed, transgenic experiments indicate the additional sequence improves PLP1's ability to enter myelin and myelin stability (Wang et al. 2008). These results are striking to us because they go beyond the relatively common pattern where a single clade-specific gene is implicated in a novelty, and instead indicate a more general pattern—both of the dominant proteins in myelin that arose at the base of the gnathostomes are novel proteins that arose at the same time as myelination itself. In addition, the later replacement

of MPZ with PLP1 in tetrapod CNS was associated with the origin of novel protein sequence within an ancestral lipophilin-like gene at the same phylogenetic node, further supporting the importance of novel proteins for the origin of complex new traits.

Bones, including a cranium joined by sutures, are clade-specific traits in the Euteleostomi, or bony vertebrates. Cranial sutures and several other bones of the head are formed by a process called intramembranous ossification. The *Mn1* gene was originally cloned as a possible cause of meningioma, but it was later shown that its normal biological role is in skull formation, especially intramembranous ossification (Meester-Smoor et al. 2005). *Mn1* is a bony vertebrate (Euteleostomi)-specific gene, indicating that it arose at roughly the same time as the intramembranous ossification (Pallares et al. 2015). In addition, genetic variation in *Mn1* was found to be important in craniofacial shape variation in an outbred mouse population (Pallares et al. 2015). *Mn1* thus has crucial roles in craniofacial development and is an example of an clade-specific gene that is implicated in a clade-specific apomorphy.

Prod1 is a salamander-specific gene that was derived by a duplication of a three-finger family protein, and salamander *Prod1* proteins share a distinctive 12-amino acid motif, indicating that the protein evolved after duplication and before radiation of salamanders (Geng et al. 2015). Salamanders are unique among tetrapods in their ability to regenerate their limbs as adults (Brockes and Gates 2014); since some fish also regenerate their limbs, it is not clear whether salamanders retain an ancestral regeneration ability that was lost in other tetrapod lineages, or whether it arose again in the salamander common ancestor. *Prod1* forms a gradient in the regenerating limb, suggesting that it may be involved in pattern formation (da Silva et al. 2002; Kumar et al. 2007a). Focal *Prod1* overexpression in the regenerating limb disrupts the spatial organization of the developing tissue and the process of regeneration (Echeverri and

Tanaka 2005). The *Ag2* protein was identified as a ligand for Prod1, and local expression of *Ag2* in a regenerating limb that has been denervated can rescue the innervation requirement for limb regeneration (Kumar et al. 2007b). Together, these results indicate that *Prod1* is a salamander-specific gene that is functionally important for adult limb regeneration, a possible salamander novelty.

Salamander limbs share an evolutionary novelty called preaxial dominance. In other amphibians and amniotes, the posterior digits develop before the anterior ones, but in salamanders, the two most anterior digits develop first (Frobisch and Shubin 2011). Disruption of *Prod1* with TALEN gene editing blocks the outgrowth of the anterior digits, showing that this gene is required for preaxial dominance, as well as limb regeneration—an unexpected mechanistic link mediated by a clade-specific gene (Kumar et al. 2015).

A defining synapomorphy of mammals is lactation. Milk provides vital proteins, carbohydrates, vitamins, salts and lipids to the mammalian neonate. Caseins are the most abundant proteins in milk, where they facilitate calcium transport for bone and tooth growth, and provide protein. There are two kinds of caseins in mammals, calcium sensitive (*i.e.* CSN1S1, CSN1S2, and CSN2) and calcium insensitive (*i.e.,*CSN3). These proteins are highly disordered and fast-evolving at the amino acid level, however Kawasaki et al. 2011 have used intron-exon structure to trace the ancestry and origin of these proteins. They are all members of the secretory calcium-binding phosphoprotein (*SCPP)* gene family, and in particular, both kinds of casein protein derived ultimately from the odontogenic ameloblast–associated (*ODAM*) gene, which is involved in the control of calcium deposition in tooth development. In particular, calcium sensitive caseins appear to have evolved from a duplication of the SCPP-Pro-Gln-rich 1 (*SCPPPQ1*) gene, which in turn evolved from a duplication of *ODAM*. The calcium insensitive

caseins seem to have evolved from a duplication of the follicular dendritic cell secreted peptide (*FDCSP*) gene, which also seems to be derived from a duplication of an ancestral *ODAM* gene. Notably for our argument, both types of casein genes appeared at the base of the mammals, when lactation evolved. Caseins were identified biochemically, without regard for their conservation across animals, and turn out to be excellent examples of clade-specific genes that are crucial for a clade-specific trait. We note that given the high sequence divergence among these genes, and between them and their ancestral genes, this appears to be an example of a "novel" protein family that has evolved by evolution from an identifiable ancestral gene. Such cases, given additional divergence, could evolve so far as to lose detectable homology.

Butterflies and beets: clade-specific genes with roles in pigmentation at intermediate taxonomic levels

Pigmentation is a conspicuous trait that can have obvious ecological significance. In addition, the regulatory and biochemical pathways that control pigmentation are often well-understood, making such traits a potentially fruitful area to look for clade-specific genes. Butterfly wing patterning is a classic model for developmental evolution. There is a tremendous diversity of lepidopteran wing patterning, but many of the elements seem to be part of an ancestral patterning ground plan, because they are found in divergent taxa and lost and gained repeatedly in evolution (reviewed in Martin and Reed 2010). The *aristaless* (*al*) homeobox gene is a conserved component of wing patterning in insects, but Martin and Reed discovered a lepidopteran-specific duplication of *aristaless* (*al2*; Martin and Reed 2010). The expression pattern of the new paralog diverged dramatically from *al*, and came to be expressed in the location of one of the conserved

elements of the wing patterning ground plan, called DII. This patterning element functions via wingless signaling, and Martin and Reed were able to show that the expression of *al2* in the DII element preceded the co-option of *wingless* expression into the element. This indicates that the novel *al2* gene created a pre-pattern that facilitated the evolution of localized expression of *wingless* in the pattern element in some clades. This is a remarkably clear case for a novel duplicated gene potentiating a novelty in the clade where it arose.

The Caryophyllales is a diverse order of flowering plants that includes cacti, carnations and beets. Nearly all Caryophyllales use a class of tyrosine-derived yellow and red pigments called the betalains, which are distinct from the anthocyanin pigments in other land plants (Brockington et al. 2015). Betalains are thought to have evolved at the base of the Caryophyllales. Brockington et al. 2015 reported that two key enzymes in the betalain synthesis pathway, DODA and CYP76AD1 originated through Caryophyllales-specific gene duplication, in the stem clade for this group, at the same phylogenetic node as the origin of betalain pigmentation. They conclude that clade-specific duplication and the subsequent neofunctionalization of these enzymes were important evolutionary steps in the origin of betalain pigmentation.

Novel genes in genera-specific organs

Genera-level synapomorphies are often striking, and the often relatively close evolutionary distance to outgroups can make them compelling models of evolutionary developmental change. The propelling fan is a clade-specific trait in water strider genus *Rhagovelia*, which facilitates the rhagovelia species' locomotion on the surface of running water. This is a novel ecological niche

which other related organisms are unable to exploit. Using gene expression, functional and behavioral studies, Santos et al. 2017 reported that two clade-specific genes—*geisha* and *mother-of-geisha*—have a significant roles in the development of propelling fans. *Mother-of-geisha* is a novel gene that is likely specific to Hemiptera (the true bugs), an order-level clade that includes *Rhagovelia* (it was also detected in a termite, an isopteran, but the significance of this is unclear). *Geisha* arose as a duplicate of this ancestral copy at the base of *Rhagovelia,* when the fan evolved. Both genes are expressed at the site of the developing fan, and knockdown of both together prevents fan formation. They showed that *geisha* and *mother-of-geisha* gene knockdown individuals have poorer locomotion performance in fast-running water. They concluded that these newly duplicated genes could have a central role in the evolution of this novel trait, and thus the adaptation to new environmental niches. One interesting question about the role of clade-specific gene in the origin of novelty is whether they are, once fixed, more likely to be co-opted for evolution later in evolution. This case is interesting on this point since *mother-of-geisha* arose and was fixed much earlier than the origin of the propelling fan. It would be interesting to know the function of *mother-of-geisha* in the outgroups to *Rhagovelia,* before it was co-opted to participate in the development of the propelling fan.

Nematosomes are a novel trait and the defining character of the sea anemone genus *Nematostella*. Nematosomes are a circulating and free-floating tissue type consisting of multicellular masses of cells within the gastrovascular cavity. Babonis et al. 2016 reported evidence that nematosomes have a role in the immune system. They also found that *Nematostella* genus-specific genes are overrepresented in the set of genes that are upregulated in nematosomes compared to two other tissues, mesentery and tentacle. Cnidarian-specific genes were also over-represented in the nematosomes. Interestingly, the same pattern—over

representation of *Nematostella*-genus and cnidarian-specific genes—was found in the other two tissues, which are not *Nematostella*-specific. In fact, the tentacles had a higher fraction of *Nematostella*-specific genes than nematosomes. Thus, these results suggest that clade-specific genes may be important for nematostomes, but that they are also important in other cell types whose origin is not thought to correlate with the origin of the genes.

Clade-specific genes in the origin of species-level adaptations: evolution of the human brain

Perhaps the most significant human-specific adaptation is our unique cognitive ability, which is linked to changes in brain size and structure in the human clade. In recent years there have been a number of human-specific gene duplications that have been implicated in the development of human-specific aspects of brain development (reviewed in Heide and Huttner 2021). Here we focus on three families, *SRGAP2*, *ARHGAP11*, and *NOTCH2NL*.

The *SRGAP2* proteins (Slit-Robo Rho GTPase activating protein 2) encode regulators of neuronal development, with roles in neuron migration and neurite outgrowth (Guerrier et al. 2009). The *SRGAP2* protein is strongly conserved across mammals, and there are no extant duplicates in any mammalian clade, but there has been a complex sequence of duplications in our own ancestral clade (Dennis et al. 2012). After the split from the chimpanzee clade, there was an initial duplication of the ancestral *SRGAP2A* gene around 3.4 MYA, to create *SRGAP2B*. Importantly, this duplication included the promoter and 5' end of the coding sequence of the ancestral gene but was truncated at the 3' end—it contained most of the F-BAR domain, which is involved with homodimerization, but lacked some 3' sequence, including several important functional domains. *SRGAP2B* was then duplicated twice, creating *SRGAP2C* and *SRGAP2D*.

The authors estimated that the duplication of *SRGAP2C*, which is now thought to be the most important duplicate for human brain evolution, occurred around 2.4 MYA based on molecular clock data. This is significant, because this is around the time that *Homo erectus* diverged from *Australopithcines,* a milestone in human evolution because the origin of *Homo erectus* was associated with increased brain size and advances in tool fabrication.

The ancestral gene (*SRGAP2A*) has a role in dendritic spine maturation in the cortex and the regulation of neuronal migration (Guerrier et al. 2009; Charrier et al. 2012). The *SRGAP2C* duplicate is strongly expressed in the developing cortex, in a pattern that largely overlaps *SRGAP2A*. SRGAP2C dimerizes with SRGAP2A, and blocks all of its known functions, at least in part by reducing the stability of SRGAP2A (Schmidt et al. 2019). The expression of SRGAP2C inhibits the ability of SRGAP2A to cause neurite branching, and this is associated with increased migration of neural precursors in the developing cortex (Charrier et al. 2012). The downregulation of SRGAP2A by SRGAP2C also causes an increase in synapse density in cortical pyramidal neurons, and lengthens the maturation period of these neurons, two human-specific neural traits. The functional effects of SRGAP2C are in part due to the C-terminal truncation, but there are also several non-synonomous mutations in this paralog that make it a stronger negative regulator of SRGAP2A. Thus, the story that has emerged is that the truncated *SRGAP2C* paralog arose around the time that the *Homo* genus was diverging from other hominins, and because of the truncation, and the shared regulatory sequence, it was immediately able to inhibit the ancestral protein, which contributed to to some human-specific changes in the cortex. Subsequently, this inhibition was strengthened by the accumulation of amino acid substitutions in the new paralog. This is a dramatic example of how a protein with a novel molecular function can arise and potentially have an immediate phenotypic effect.

The *ARHGAP11B* gene was also derived by partial gene duplication, and the human-specific form has been shown to increase rates of basal progenitor cell proliferation in the developing cortex (Florio et al. 2016). The ancestral gene, *ARHGAP11A* was recently identified as a key component of cytokinesis in animal cells (Zanin et al. 2013). The mitotic spindle positions the contractile ring by limiting the spatial activity of RhoA on the cortex. There are many genes in animal cells with G-protein activating activity that could, in principle, limit RhoA to its proper, narrow band of activity, but Zanin and colleagues identified *ARHGAP11A* as the RhoGAP that performs this essential function. The duplication of *ARHGAP11A* to create the *ARHGAP11B* gene has been dated to around 5 MYA in our clade, after the split with the clade leading to chimpanzees. The original duplication was truncated, but still contained the full RhoGAP domain; this original duplicate protein sequence does not promote basal progenitor proliferation (Florio et al. 2015). A subsequent C->G nucleotide substitution created a novel splice acceptor site, which results in a frameshift and a human-specific 47 amino acid C' terminal protein sequence. This novel protein does not encode RhoGAP activity, and it does increase basal cell progenitor proliferation (Florio 2016). Gain of function studies in several different mammals have shown that human *ARHGAP11B* has widespread effects in the developing cortex that generally increase proliferation in the cortex and make it more human-like (Florio et al. 2015; Kalebic et al. 2018). In one especially remarkable recent study, Heide et al. 2020 expressed human *ARHGAP11B* under the control of the human promoter in fetal marmoset brains. This resulted in an increase of neocortex size, and induced human-like folding of the neocortex which is normally smooth at the stages examined. They also observed other effects that reflect human-specific aspects of neocortical development, including thickening of the

cortical plate, and increase in basal progenitor cells in the subventricular zone that is thought to be particularly important for human neocortical expansion.

Importantly for the topic of this review, the molecular function of *ARHGAP11B* seems to be totally different from the function of its parent gene, *ARHGAP11A* (Namba et al. 2020). The human-specific gene is localized to the mitochondria rather than the nucleus, and it interacts with the mitochondrial adenine nucleotide translocase (ANT) and inhibits the opening of the permeability transition pore (mPTP). The inhibitory interaction with the mPTP is mediated by the novel human-specific sequence that was generated by the novel splice site, and results in elevated Ca++ levels. This in turn promotes glutominolysis, which is an alternate metabolic pathway where glutamine is used for the tricarbocylic acid (TCA, a.k.a. Krebs or CAC) cycle. Inducing glutaminolysis increases basal progenitor cell proliferation, indicating that *ARHGAP11B*'s effects on cortical proliferation are mediated by mPTP inhibition and increases in glutominolysis (Namba et al. 2020).

The human-specific family *NOTCH2NL* also has a complex origin story that occured largely in the clade leading to modern humans (Suzuki et al. 2018; Fiddes et al. 2018; reviewed in Heide and Huttner 2021). In this case, the original event was a duplication of the *NOTCH2* gene to create the *NOTCH2NLR* gene, which exists as a pseudogene in the chimpanzee and gorilla genomes. Sometime after the chimpanzee-human common ancestor, in the human clade, there was a gene conversion with *NOTCH2* that restored the functional sequence of *NOTCH2NLR*. This includes the sequence derived from the *NOTCH2* promoter, and most of the extracellular domain, but not the transmembrane domain or intracellular domain of *NOTCH2*. In addition there are 20 amino acids that are specific to the human-specific genes. This was

followed by three duplications, to generate the *NOTCH2NLB*, *NOTCH2NLA* and *NOTCH2NLC*, in that order.

All of these three genes are expressed in developing human neocortex, but *NOTCH2NLB* is the only one with a signal peptide sequence and shows the highest expression levels. They are broadly expressed in the ventricular zone, in apical progenitor cells and also in some basal progenitor cells in the subventricular zone. The effects of perturbing these genes during cortical development in mouse or mouse cortical organoids have been examined (Florio et al. 2018; Suzuki et al. 2018; Fiddes et al. 2018). In general, loss of function and gain of function experiments together indicate that the role of these genes is to promote more proliferation by progenitor cells in the developing cortex. Molecularly, they have been shown to bind to canonical Notch receptors, and upregulate Notch signaling (Suzuki et al. 2018; Fiddes et al. 2018). *NOTCH2LB* has specifically been shown to bind to Delta-like 1 ligands on *NOTCH2LB* expressing cells, and increase the expansion of progenitor cells (Suzuki et al. 2018).

Interestingly, each of the *NOTCH2NL* genes has evolved with a nearby partner gene that is part of the the human-specific *HNBF* family, indicating that the initial duplication included an *HNBF* gene, and each successive duplication also duplicated that loci (Fiddes et al. 2019). These *HNBF* genes contain repeated protein domains, called Olduvai domains, that are the most amplified protein domains in the human lineage. These proteins also have been implicated in cortical growth, and also neuronal disease in humans. Together these four gene families provide further examples of novel gene duplicates, with novel protein sequence and domain structure, that have been implicated in a dramatic clade-specific synapomorphy—the morphological expansion and elaboration of the human brain.

Discussion

Our goal has been to highlight compelling cases of clade-specific genes that seem to be important for characters that arose around the same time as the genes were fixed in the ancestral clade. Our motivation is to broaden appreciation of the functional roles that such genes often have, for two reasons. First, we hope that this will inform choices that biologists make when they are looking for genes that control interesting traits. For instance, QTL studies often identify large chromosomal intervals containing multiple genes, and so choices sometimes need to be made about which candidate loci to pursue. In these cases, known gene families are perhaps more compelling, but knowing that a novel gene arose around the same time as the trait might be an equally or more compelling reason to prioritize it. This could also be true when evaluating lists of genes from RNA-seq experiments or other similar exploratory approaches.

Second, we hope that these examples might motivate others to directly screen for clade-specific genes in clades that they are interested in, and screen those genes for roles in clade-specific traits—an approach that could be called a clade-specific gene screen. It is relatively straightforward to screen for novel protein sequences that are conserved within a clade but not detectable outside of it. For example, in our recent screen for spiralian-specific sequences (Wu et al. 2020), we used high quality genomes from three spiralian phyla, and looked for sequences that were in all the three spiralian phyla with strong conservation but not detectable in any of the ten outgroup animal genomes we used nor the NCBI nr database. As indicated above, this recovered 37 sequences, which is not an unreasonable number for follow-up screens and validation of clade-specificity with more sensitive homology searching. In our case, the

subsequent *in situ* hybridization screen identified two (of 20 screened) genes with similar expression in ciliary bands; these were analyzed in depth with more sensitive search strategies. These results indicate that for many ancient clades, stringent homology screens for clade-specific protein sequences are feasible and fruitful. Nevertheless, most clade-specific genes for any clade are gene duplications, and these are more challenging to reliably detect using whole genome scale analyses, though some methods are available (Li et al. 2003; reviewed in Lallemand et al. 2020). We note that, depending on the downstream screening strategy, a bioinformatic screen with a somewhat high false positive rate for clade-specific duplications would also be workable if the subsequent screen reduced the number of candidate duplicates involved in a novel trait down to a reasonable number for in depth sequence analysis.

Although we still have only a limited number of examples of clade-specific genes participating in clade specific traits, there are a number of general questions that arise given the examples we do have. Among the most compelling issues is the relationship between clade-specific genes and clade-specific traits. We propose that, in clades that are sufficiently old that the sequence conservation of clade-specific genes can be observed after their origin, such conserved clade-specific genes are in fact *more likely* than other genes to be involved in clade-specific traits. On one level, this is self-evidently true, even if it might be sometimes trivially so: if a gene duplicates and diverges at the base of a clade, and the duplicate takes one part of the parent gene's function or a new function, then the molecular basis of that function is novel by virtue of being carried out by a slightly different protein in that clade. However, we are proposing that, when *complex new* traits evolve, a gene that arose and was fixed at the same evolutionary node is more likely than other genes to be involved, simply because the trait and the fixation event occurred at roughly the same evolutionary interval. Perhaps this is more true of

new genes that arise apparently *de novo*, compared to clade-specific duplications: when an existing gene duplicates, there are likely many processes that it could be co-opted into, because it is a functional protein with existing molecular interactions, while the fixation of a completely novel protein sequence may be more likely to occur in the context of the origin of a complex trait that might require new molecular functions.

As recently proposed by Jockusch and Fischer 2021, there are ways that this question (and related hypotheses) can be tested systematically. In general, this would involve comparing the expression (and/or function) of clade-specific genes with a set of genes that arose at an earlier node in the phylogeny, and a set that arose at a later node. By comparing the involvement of these sets of genes in one or more clade-specific traits, and in control traits that arose later and earlier in evolution, we could test questions like: 1) whether clade-specific genes are more or less likely to be involved in a clade-specific trait, compared to older genes, as we predict above; 2) whether new genes that arise after the origin of a trait might also be more likely to be involved in that trait; 3) whether *de novo* clade-specific genes are more likely to be involved in clade specific traits compared to duplications. This approach should be facilitated by the improvement in single-cell RNA seq methods in diverse organisms, which could allow the roles of clade-specific genes and control groups to be measured in many cell types and organs simultaneously.

There are other compelling questions about the role of clade-specific genes in novelties that will likely have to wait until there are more examples that have been functionally characterized. For instance, are clade-specific genes more likely to be involved in some kinds of clade-specific phenotypes, e.g. morphological vs. metabolic vs. neural? It is tempting to predict that clade-specific genes are particularly likely to become involved with novel biochemical or structural traits, since these may fall outside of the repertoire of existing proteins. The clade-

specific genes involved with myelination, and the caseins seem to be examples of this, but we also have good examples of clade-specific genes that are clearly regulatory, like *aristaless2, NOTCH2NLB* and *SRGAP2C*. Similarly, are there some molecular functions that are more likely to be carried out by particular kinds of clade-specific genes? For instance, are *de novo* clade-specific genes more likely to be structural, and clade specific duplications more likely to be regulatory? This could be a particularly difficult question given the challenges of determining the molecular function of *de novo* clade-specific genes, which by definition have no homology to other proteins.

Conclusion

We have tried to make the case that clade-specific genes are broadly involved in evolutionary novelties that are of interest to evolutionary developmental biologists. We think such genes clearly deserve attention when they are implicated in traits of interest, and we also encourage others to directly screen for them and examine their functions in particular clades. The techniques developed to examine conserved tool-kit genes in non-genetic systems, like RNAi and morpholinos, are equally suited to perturbing clade-specific genes. At the same time, some new techniques like single-cell RNA-seq promise to be particularly useful for efficiently implicating clade-specific genes in clade-specific cell types and organs. We hope that the next few years see a rapid accumulation of new insights into the functions of clade-specific genes.

We will end with an acknowledgement that there is a biologically interesting limitation to the kinds of approaches we are describing here: for most clades, there are certainly many clade-specific novelties that we are unaware of, so that even if a clade-specific gene does not appear to

function in one of the known clade-specific traits, it might be involved in another, unknown synapomorphy. This does not limit the utility of comparing the clade-specific genes to older genes in one known novelty, but it does mean that the number of clade-specific genes that are involved in any clade-specific trait will be underestimated. This bias is conservative for our proposal above. It also suggests, if clade-specific genes are disproportionately likely to function in clade specific traits, then unbiased surveys of conserved clade-specific genes will not only find those that are involved with known synapomorphies of the group, but will also find those that turn out to be involved with previously unknown traits that arose at the same phylogenetic node.

Acknowledgements

We thank Scott Gilbert for useful insights. This work was supported by N.S.F. grants IOS 1656558 and 2053371 to J.D.L.

References

Babonis, Leslie S., and Mark Q. Martindale. "Old cell, new trick? Cnidocytes as a model for the evolution of novelty." *American Zoologist* 54, no. 4 (2014): 714-722.

Babonis, Leslie S., Mark Q. Martindale, and Joseph F. Ryan. "Do novel genes drive morphological novelty? An investigation of the nematosomes in the sea anemone Nematostella vectensis." *BMC evolutionary biology* 16, no. 1 (2016): 1-22.

Brockes, Jeremy P., and Phillip B. Gates. "Mechanisms underlying vertebrate limb regeneration: lessons from the salamander." (2014): 625-630.

Brockington, Samuel F., Ya Yang, Fernando Gandia-Herrero, Sarah Covshoff, Julian M. Hibberd, Rowan F. Sage, Gane KS Wong, Michael J. Moore, and Stephen A. Smith. "Lineage-specific gene radiations underlie the evolution of novel betalain pigmentation in Caryophyllales." *New Phytologist* 207, no. 4 (2015): 1170-1180.

Carrasco, Andrés E., William McGinnis, Walter J. Gehring, and Eddy M. De Robertis. "Cloning of an X. laevis gene expressed during early embryogenesis coding for a peptide region homologous to Drosophila homeotic genes." *Cell* 37, no. 2 (1984): 409-414.

Charrier, Cécile, Kaumudi Joshi, Jaeda Coutinho-Budd, Ji-Eun Kim, Nelle Lambert, Jacqueline De Marchena, Wei-Lin Jin et al. "Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation." *Cell* 149, no. 4 (2012): 923-935.

Chen, Sidi, Benjamin H. Krinsky, and Manyuan Long. "New genes as drivers of phenotypic evolution." *Nature Reviews Genetics* 14, no. 9 (2013): 645-660.

da Silva, Sara Morais, Phillip B. Gates, and Jeremy P. Brockes. "The newt ortholog of CD59 is implicated in proximodistal identity during amphibian limb regeneration." *Developmental cell* 3, no. 4 (2002): 547-555.

David, Charles N., Suat Özbek, Patrizia Adamczyk, Sebastian Meier, Barbara Pauly, Jarrod Chapman, Jung Shan Hwang, Takashi Gojobori, and Thomas W. Holstein. "Evolution of complex structures: minicollagens shape the cnidarian nematocyst." *Trends in genetics* 24, no. 9 (2008): 431-438.

Dennis, Megan Y., Xander Nuttle, Peter H. Sudmant, Francesca Antonacci, Tina A. Graves, Mikhail Nefedov, Jill A. Rosenfeld et al. "Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication." *Cell* 149, no. 4 (2012): 912-922.

Echeverri, Karen, and Elly M. Tanaka. "Proximodistal patterning during limb regeneration." *Developmental biology* 279, no. 2 (2005): 391-401.

Engel, Ulrike, Suat Oezbek, Ruth Engel, Barbara Petri, Friedrich Lottspeich, and Thomas W. Holstein. "Nowa, a novel protein with minicollagen Cys-rich domains, is involved in nematocyst formation in Hydra." *Journal of cell science* 115, no. 20 (2002): 3923-3934.

Fiddes, Ian T., Alex A. Pollen, Jonathan M. Davis, and James M. Sikela. "Paired involvement of human-specific Olduvai domains and NOTCH2NL genes in human brain evolution." *Human genetics* 138, no. 7 (2019): 715-721.

Fiddes, Ian T., Gerrald A. Lodewijk, Meghan Mooring, Colleen M. Bosworth, Adam D. Ewing, Gary L. Mantalas, Adam M. Novak et al. "Human-specific NOTCH2NL genes affect Notch signaling and cortical neurogenesis." *Cell* 173, no. 6 (2018): 1356-1369.

Florio, Marta, Mareike Albert, Elena Taverna, Takashi Namba, Holger Brandl, Eric Lewitus, Christiane Haffner et al. "Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion." *Science* 347, no. 6229 (2015): 1465-1470.

Florio, Marta, Michael Heide, Anneline Pinson, Holger Brandl, Mareike Albert, Sylke Winkler, Pauline Wimberger, Wieland B. Huttner, and Michael Hiller. "Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex." *elife* 7 (2018): e32332.

Florio, Marta, Takashi Namba, Svante Pääbo, Michael Hiller, and Wieland B. Huttner. "A single splice site mutation in human-specific ARHGAP11B causes basal progenitor amplification." *Science advances* 2, no. 12 (2016): e1601941.

Fröbisch, Nadia B., and Neil H. Shubin. "Salamander limb development: integrating genes, morphology, and fossils." *Developmental Dynamics* 240, no. 5 (2011): 1087-1099.

Geng, Jie, Phillip B. Gates, Anoop Kumar, Stefan Guenther, Acely Garza-Garcia, Carsten Kuenne, Peng Zhang, Mario Looso, and Jeremy P. Brockes. "Identification of the orphan gene Prod 1 in basal and other salamander families." *Evodevo* 6, no. 1 (2015): 1-4.

Gilbert, Scott F., John M. Opitz, and Rudolf A. Raff. "Resynthesizing evolutionary and developmental biology." *Developmental biology* 173, no. 2 (1996): 357-372.

Gould, Robert M., Todd Oakley, Jared V. Goldstone, Jason C. Dugas, Scott T. Brady, and Alexander Gow. "Myelin sheaths are formed with proteins that originated in vertebrate lineages." *Neuron glia biology* 4, no. 2 (2008): 137-152.

Guerrier, Sabrice, Jaeda Coutinho-Budd, Takayuki Sassa, Aurélie Gresset, Nicole Vincent Jordan, Keng Chen, Wei-Lin Jin, Adam Frost, and Franck Polleux. "The F-BAR domain of srGAP2 induces membrane protrusions required for neuronal migration and morphogenesis." *Cell* 138, no. 5 (2009): 990-1004.

Heide, Michael, Christiane Haffner, Ayako Murayama, Yoko Kurotaki, Haruka Shinohara, Hideyuki Okano, Erika Sasaki, and Wieland B. Huttner. "Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset." *Science* 369, no. 6503 (2020): 546-550.

Heide, Michael, and Wieland B. Huttner. "Human-Specific Genes, Cortical Progenitor Cells, and Microcephaly." *Cells* 10, no. 5 (2021): 1209.

Holstein, T., & Tardent, P. (1984). An ultrahigh-speed analysis of exocytosis: nematocyst discharge. *Science*, *223*(4638), 830-833.

Hwang, Jung Shan, Hajime Ohyanagi, Shiho Hayakawa, Naoki Osato, Chiemi Nishimiya-Fujisawa, Kazuho Ikeo, Charles N. David, Toshitaka Fujisawa, and Takashi Gojobori. "The evolutionary emergence of cell type-specific genes inferred from the gene expression analysis of Hydra." *Proceedings of the National Academy of Sciences* 104, no. 37 (2007): 14735-14740.

Hwang, Jung Shan, Yasuharu Takaku, Jarrod Chapman, Kazuho Ikeo, Charles N. David, and Takashi Gojobori. "Cilium evolution: identification of a novel protein, nematocilin, in the mechanosensory cilium of Hydra nematocytes." *Molecular biology and evolution* 25, no. 9 (2008): 2009-2017.

Hwang, Jung Shan, Yasuharu Takaku, Tsuyoshi Momose, Patrizia Adamczyk, Suat Özbek, Kazuho Ikeo, Konstantin Khalturin et al. "Nematogalectin, a nematocyst protein with GlyXY and galectin domains, demonstrates nematocyte-specific alternative splicing in Hydra." *Proceedings of the National Academy of Sciences* 107, no. 43 (2010): 18539-18544.

Jockusch, Elizabeth L., and Cera R. Fisher. "Something old, something new, something borrowed, something red: the origin of ecologically relevant novelties in Hemiptera." *Current opinion in genetics & development* 69 (2021): 154-162.

Kaessmann, Henrik. "Origins, evolution, and phenotypic impact of new genes." *Genome research* 20, no. 10 (2010): 1313-1326.

Kawasaki, Kazuhiko, Anne-Gaelle Lafont, and Jean-Yves Sire. "The evolution of milk casein genes from tooth genes before the origin of mammals." *Molecular biology and evolution* 28, no. 7 (2011): 2053-2061.

Kalebic, Nereo, Carlotta Gilardi, Mareike Albert, Takashi Namba, Katherine R. Long, Milos Kostic, Barbara Langen, and Wieland B. Huttner. "Human-specific ARHGAP11B induces hallmarks of neocortical expansion in developing ferret neocortex." *Elife* 7 (2018): e41241.

Khalturin, Konstantin, Chuya Shinzato, Maria Khalturina, Mayuko Hamada, Manabu Fujie, Ryo Koyanagi, Miyuki Kanda et al. "Medusozoan genomes inform the evolution of the jellyfish body plan." *Nature ecology & evolution* 3, no. 5 (2019): 811-822.

Khalturin, Konstantin, Georg Hemmrich, Sebastian Fraune, René Augustin, and Thomas CG Bosch. "More than just orphans: are taxonomically-restricted genes important in evolution?" *Trends in Genetics* 25, no. 9 (2009): 404-413.

Klasberg, Steffen, Tristan Bitard-Feildel, and Ludovic Mallet. "Computational identification of novel genes: current and future perspectives." *Bioinformatics and Biology insights* 10 (2016): BBI-S39950.

Koch, Alexander W., Thomas W. Holstein, Carola Mala, Eva Kurz, Jürgen Engel, and Charles N. David. "Spinalin, a new glycine-and histidine-rich protein in spines of Hydra nematocysts." *Journal of Cell Science* 111, no. 11 (1998): 1545-1554.

Kumar, Anoop, Phillip B. Gates, and Jeremy P. Brockes. "Positional identity of adult stem cells in salamander limb regeneration." *Comptes rendus biologies* 330, no. 6-7 (2007a): 485-490.

Kumar, Anoop, James W. Godwin, Phillip B. Gates, A. Acely Garza-Garcia, and Jeremy P. Brockes. "Molecular basis for the nerve dependence of limb regeneration in an adult vertebrate." *science* 318, no. 5851 (2007b): 772-777

Kumar, Anoop, Phillip B. Gates, Anna Czarkwiani, and Jeremy P. Brockes. "An orphan gene is necessary for preaxial digit formation during salamander limb development." *Nature Communications* 6, no. 1 (2015): 1-8.

Kurz, Eva M., Thomas W. Holstein, Barbara M. Petri, Jürgen Engel, and Charles N. David. "Mini-collagens in hydra nematocytes." *The Journal of cell biology* 115, no. 4 (1991): 1159-1169..

Lallemand, Tanguy, Martin Leduc, Claudine Landès, Carène Rizzon, and Emmanuelle Lerat. "An overview of duplicated gene detection methods: Why the duplication mechanism has to be accounted for in their choice." *Genes* 11, no. 9 (2020): 1046.

Leclère, Lucas, and Eric Röttinger. "Diversity of cnidarian muscles: function, anatomy, development and regeneration." *Frontiers in cell and developmental biology* 4 (2017): 157.

Li, Li, Christian J. Stoeckert, and David S. Roos. "OrthoMCL: identification of ortholog groups for eukaryotic genomes." *Genome research* 13, no. 9 (2003): 2178-2189.

Lynch, Michael, and Vaishali Katju. "The altered evolutionary trajectories of gene duplicates." *TRENDS in Genetics* 20, no. 11 (2004): 544-549.

Martin, Arnaud, and Robert D. Reed. "Wingless and aristaless2 define a developmental ground plan for moth and butterfly wing pattern evolution." *Molecular biology and evolution* 27, no. 12 (2010): 2864-2878.

Mayr, Ernst. “Animal species and Evolution.” Harvard University Press, Cambridge (1966)

McGinnis, William, Richard L. Garber, Johannes Wirz, Atsushi Kuroiwa, and Walter J. Gehring. "A homologous protein-coding sequence in Drosophila homeotic genes and its conservation in other metazoans." *Cell* 37, no. 2 (1984): 403-408.

Meester-Smoor, Magda A., Marcel Vermeij, Marjolein JL van Helmond, Anco C. Molijn, Karel HM van Wely, Arnold CP Hekman, Christl Vermey-Keers, Peter HJ Riegman, and Ellen C. Zwarthoff. "Targeted disruption of the Mn1 oncogene results in severe defects in development of membranous bones of the cranial skeleton." *Molecular and cellular biology* 25, no. 10 (2005): 4229-4236.

Milde, Sabine, Georg Hemmrich, Friederike Anton-Erxleben, Konstantin Khalturin, Jörg Wittlieb, and Thomas CG Bosch. "Characterization of taxonomically restricted genes in a phylum-restricted cell type." *Genome biology* 10, no. 1 (2009): 1-16.

Möbius, Wiebke, Julia Patzig, Klaus-Armin Nave, and Hauke B. Werner. "Phylogeny of proteolipid proteins: divergence, constraints, and the evolution of novel functions in myelination and neuroprotection." *Neuron glia biology* 4, no. 2 (2008): 111-127.

Namba, Takashi, Judit Dóczi, Anneline Pinson, Lei Xing, Nereo Kalebic, Michaela Wilsch-Bräuninger, Katherine R. Long et al. "Human-specific ARHGAP11B acts in mitochondria to expand neocortical progenitors by glutaminolysis." *Neuron* 105, no. 5 (2020): 867-881.

Ohno S. “Evolution by Gene Duplication.” Berlin-Heidelberg-New York: Springer-Verlag; 1970.

Pallares, Luisa F., Peter Carbonetto, Shyam Gopalakrishnan, Clarissa C. Parker, Cheryl L. Ackert-Bicknell, Abraham A. Palmer, and Diethard Tautz. "Mapping of craniofacial traits in outbred mice identifies major developmental genes involved in shape determination." *PLoS genetics* 11, no. 11 (2015): e1005607.

Rödelsperger, Christian, Neel Prabh, and Ralf J. Sommer. "New gene origin and deep taxon phylogenomics: opportunities and challenges." *Trends in Genetics* 35, no. 12 (2019): 914-922.

Santos, M. Emília, Augustin Le Bouquin, Antonin JJ Crumière, and Abderrahman Khila. "Taxon-restricted genes at the origin of a novel trait allowing access to a new environment." *Science* 358, no. 6361 (2017): 386-390.

Schmidt, Ewoud RE, Justine V. Kupferman, Michelle Stackmann, and Franck Polleux. "The human-specific paralogs SRGAP2B and SRGAP2C differentially modulate SRGAP2A-dependent synaptic development." *Scientific reports* 9, no. 1 (2019): 1-8.

Shigenobu, Shuji, and David L. Stern. "Aphids evolved novel secreted proteins for symbiosis with bacterial endosymbiont." *Proceedings of the Royal Society B: Biological Sciences* 280, no. 1750 (2013): 20121952.

Steinmetz, Patrick RH, Johanna EM Kraus, Claire Larroux, Jörg U. Hammel, Annette Amon-Hassenzahl, Evelyn Houliston, Gert Wörheide, Michael Nickel, Bernard M. Degnan, and Ulrich Technau. "Independent evolution of striated muscles in cnidarians and bilaterians." *Nature* 487, no. 7406 (2012): 231-234.

Suzuki, Ikuo K., David Gacquer, Roxane Van Heurck, Devesh Kumar, Marta Wojno, Angéline Bilheu, Adèle Herpoel et al. "Human-specific NOTCH2NL genes expand cortical neurogenesis through Delta/Notch regulation." *Cell* 173, no. 6 (2018): 1370-1384.

Tautz, Diethard, and Tomislav Domazet-Lošo. "The evolutionary origin of orphan genes." *Nature Reviews Genetics* 12, no. 10 (2011): 692-702.

Weisman, Caroline M., Andrew W. Murray, and Sean R. Eddy. "Many, but not all, lineage-specific genes can be explained by homology detection failure." *PLoS biology* 18, no. 11 (2020): e3000862.

Wilkins, Adam S. "“The Genetic Tool-Kit”: The Life-History of an Important Metaphor." *Advances in evolutionary developmental biology* (2013): 1-14.

Wu, Longjun, Laurel S. Hiebert, Marleen Klann, Yale Passamaneck, Benjamin R. Bastin, Stephan Q. Schneider, Mark Q. Martindale, Elaine C. Seaver, Svetlana A. Maslakova, and J. David Lambert. "Genes with spiralian-specific protein motifs are expressed in spiralian ciliary bands." *Nature communications* 11, no. 1 (2020): 1-11.

Zanin, Esther, Arshad Desai, Ina Poser, Yusuke Toyoda, Cordula Andree, Claudia Moebius, Marc Bickle, Barbara Conradt, Alisa Piekny, and Karen Oegema. "A conserved RhoGAP limits M phase contractility and coordinates with microtubule asters to confine RhoA during cytokinesis." *Developmental Cell* 26, no. 5 (2013): 496-510.

Gene name	Clade	Character	Mode of origin	Reference
Trochin	Spiralia	Primary ciliated band	Putative *de novo*	Wu et al. 2020
Lophotrochin	Spiralia	Primary ciliated band	Putative *de novo* origin of domain, combined with older conserved domain	Wu et al. 2020
MBP	Gnathostomes	myelin	Putative *de novo*	Gould et al. 2008
MPZ	Gnathostomes	myelin	Duplication and divergence	Gould et al. 2008
PLP1	Tetrapods	Myelin (CNS)	Duplication and divergence including a novel exon	Gould et al. 2008
Mn1	Euteleostoma	Bony skull	Putative *de novo*	Meester-Smoor et al. 2005
Prod1	Salamanders	Limb development and regeneration	Duplication and divergence, including a novel 12 AA motif	Geng et al. 2015
Aristaless2	Lepidopterans	Wing pattern element DII	Duplication and divergence	Martin and Reed 2010
DODA and CYP76AD1	Caryophyllales	Betalain pigments	Duplication and divergence	Brockington et al. 2015
geisha	*Rhagovelia*	Propelling fan	Duplication from a gene (*mother-of-geisha*) that previously arose at the base of a clade that includes *Rhagovelia*	Santos et al. 2017
SRGAP2C	*Homo sapiens*	Cortical expansion	Duplication with truncation, divergence	Schmidt et al. 2019
ARHGAP11B	Homo sapiens	Cortical expansion	Duplication, frameshift causing loss of RhoGAP domain and new sequence	Florio et al. 2016
NOTCH2NLB	*Homo sapiens*	Cortical expansion	Duplication to form pseudogene, gene conversion resulting in truncation with a novel 20 AA sequence.	Fiddes et al. 2018

Table 1: Examples of clade-specific genes, their characters and origins.

Figure 1.

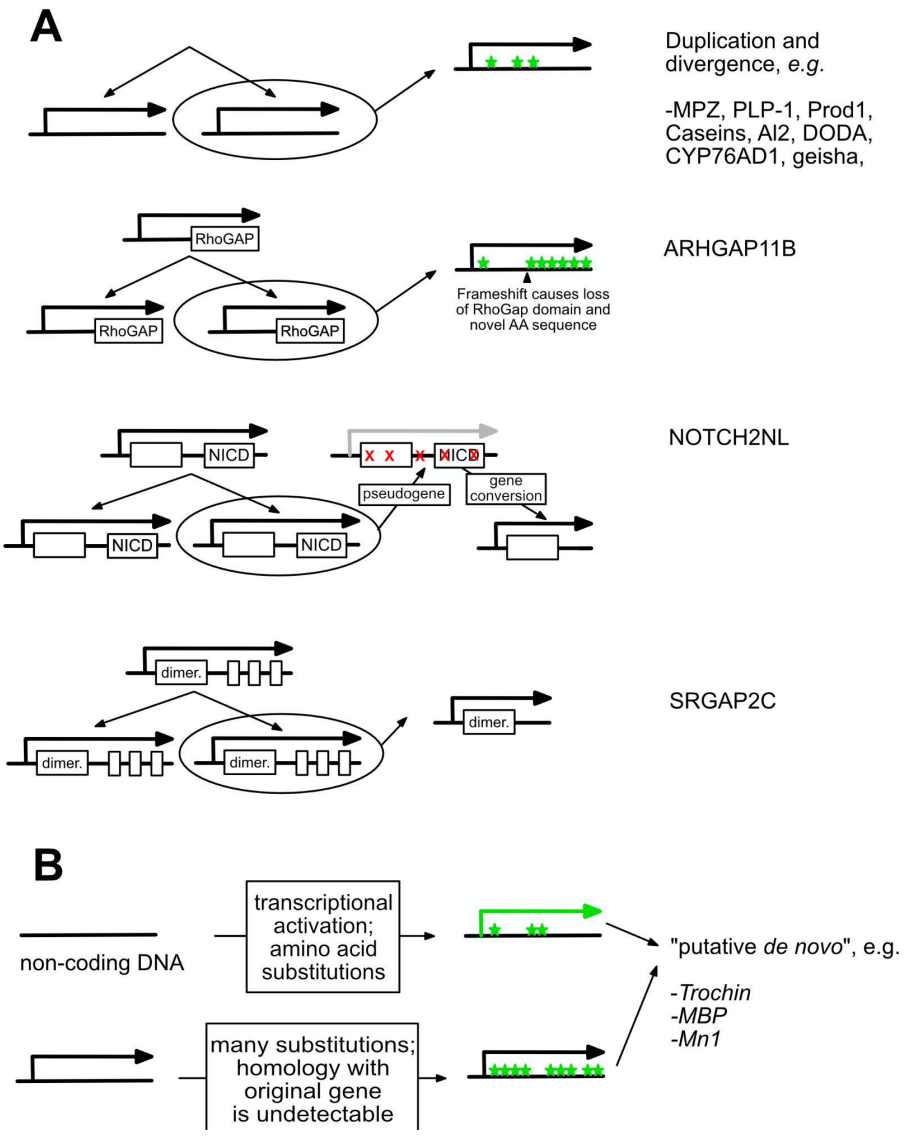

Figure 1: Modes of origin of novel genes described in this review. A) Duplication and divergence. The most common mode of origin for a gene with novel properties is duplication, followed by amino acid divergence in the duplicate of interest. The top panel shows a relatively simple case of this, that applies to some of our examples, and the bottom panels show more complex examples associated with novel human proteins involved in brain development. B) Putative *de novo* genes. These are defined by having no detectable homology outside of the clade, at least in one protein domain. This can arise by, either, *de novo* origin of transcription from a non-coding region (top panel), or by rapid sequence evolution so the homology is not detectable at the sequence level (bottom panel).