

A duplicated Hedgehog ligand and a duplicated Patched receptor show different constraint and genomic context in land snails and slugs

Zhi-Jie Xu

zjxmolls@outlook.com

Donghua University <https://orcid.org/0009-0003-3855-0921>

Research Article

Keywords: Hedgehog signalling, Patched, Smoothened, gene duplication, whole-genome duplication, Stylommatophora, genome synteny

Posted Date: October 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11512040/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

A duplicated Hedgehog ligand and a duplicated Patched receptor show different constraint and genomic context in land snails and slugs

Zhi-Jie Xu¹

¹College of Biological Science and Medical Engineering, Donghua University, Shanghai, China

Corresponding author:
Zhi-Jie Xu¹

Email address: zjxmolls@outlook.com

ABSTRACT

Gene duplication is a major source of evolutionary novelty, but duplicated genes do not share a single fate. We compared the duplication history of the *Hedgehog* ligand and its receptor *Patched* across land snails and slugs (Stylommatophora), a group with a debated ancestral whole-genome duplication. In eleven genomes, *Hedgehog* and *Patched* are each retained in two copies, whereas *Smoothened* is strictly single-copy. The two *Hedgehog* paralogs form two deeply separated clades; one evolves under consistently relaxed constraint, accumulating amino-acid substitutions faster in all ten species (median ratio 1.86), with a higher d_N/d_S (0.222 versus 0.156) and shorter introns. The two *Patched* copies are both strongly constrained (ω 0.116 and 0.126). At the site level, the *Hedgehog* copies show significant rate-shift (Type I) divergence concentrated in the C-terminal Hog domain, whereas the *Patched* copies show none. The genomic contexts differ too. In two *Cepaea* species the *Patched* paralogs lie on a chromosome-scale duplicate pair retaining 74–95% two-sided collinear coverage, and their neighbourhoods stay syntenic in eight of nine genomes; the *Hedgehog* pair retains 20% and 53%, and its neighbourhood test is significant in one of eight. *Hedgehog* collinearity is reduced but not absent; the data do not distinguish a different mechanism from faster local evolution. Two public *Hedgehog* models are defective, and *Gli*, reported as single-copy from proteomes, has a second partially degenerated copy in *Lissachatina immaculata*. Proteome-level inventories alone are an unreliable guide to copy number.

Keywords: Hedgehog signalling; Patched; Smoothened; gene duplication; whole-genome duplication; Stylommatophora; genome synteny

INTRODUCTION

Hedgehog (Hh) signalling is one of the core intercellular signalling systems of bilaterians. It was first characterised through the segment-polarity mutants of *Drosophila melanogaster* (Nüsslein-Volhard and Wieschaus, 1980) and is now known to control patterning, proliferation and differentiation in organisms as diverse as insects, annelids, molluscs and vertebrates (Ingham and McMahon, 2001; Briscoe and Théron, 2013). The pathway is short and structurally simple: a secreted ligand (Hh) binds the twelve-pass transmembrane receptor Patched (Ptch), which in the absence of ligand represses the seven-pass transmembrane protein Smoothened (Smo); ligand binding relieves this repression and allows downstream signal transduction through the Gli family of zinc-finger transcription factors (Briscoe and Théron, 2013).

Because the pathway is short, its stoichiometry is unusually exposed to perturbation. The gene-balance hypothesis predicts that genes whose products participate in stoichiometric complexes or in dose-sensitive

44 signalling cascades are retained in duplicate more often than genes whose products act alone (Birchler and
45 Veitia, 2012; Veitia et al., 2008). A signalling pathway in which a ligand, its receptor and an intracellular
46 transducer are all subject to duplication therefore provides a compact test case for how dosage constraints
47 shape duplicate-gene retention.

48 Land snails and slugs (Mollusca: Gastropoda: Stylommatophora) are an attractive system for such
49 a comparison. The group contains more than 30,000 described species and has colonised terrestrial
50 habitats repeatedly and independently. A whole-genome duplication (WGD) at or near the base of the
51 Stylommatophora has been proposed on the basis of chromosome-scale synteny, K_s distributions and
52 duplicated Hox clusters in the giant African snail *Achatina immaculata* (Liu et al., 2021), in the slug
53 *Arion vulgaris* (Chen et al., 2022), and across ten Stylommatophoran species (McHale et al., 2025). The
54 estimated date is roughly 70–110 million years ago, and the event is thought to be shared by the group as
55 a whole (Liu et al., 2021; Chen et al., 2022; McHale et al., 2025).

56 The WGD hypothesis is not settled. A recent comparative analysis of 238 lophotrochozoan genomes
57 found no distinct peak in the intermediate K_s range in 52 taxa, and interpreted the duplication signal
58 in *Arion vulgaris* as a weak, shallow peak consistent with a low likelihood of an ancient WGD (Nath
59 et al., 2025). Duplicated BUSCO proportions in Stylommatophora (16–21%) are nonetheless an order
60 of magnitude higher than in other molluscs (0.5–1.5%) (McHale et al., 2025), and chromosome-scale
61 1:2 synteny between *Biomphalaria* and *Poecilozonites* has been reported (Winingear et al., 2026). The
62 dispute turns largely on how K_s distributions are interpreted.

63 Against this background we asked three questions. First, how many copies of each Hedgehog pathway
64 component are present in Stylommatophoran genomes, and are those copies orthologous across species?
65 Second, do the duplicates evolve under different constraint? Third, do they occupy different genomic
66 contexts—in particular, are the *Patched* duplicates associated with the chromosome-scale duplication that
67 the WGD hypothesis predicts, and are the *Hedgehog* duplicates not?

68 We addressed these questions in eleven Stylommatophoran genomes, combining phylogenetics, codon-
69 based selection analysis and genome-internal synteny. In the course of the work we encountered, and
70 corrected, several annotation errors that materially affect the conclusions, and we report them because
71 they illustrate a general problem with proteome-level gene-family inventories.

72 MATERIALS AND METHODS

73 Genomes and proteomes

74 Protein and coding sequences were obtained from Ensembl Rapid Release for eleven Stylommatophoran
75 species: *Albinaria teres*, *Ariolimax columbianus*, *Arion vulgaris*, *Candidula unifasciata*, *Cepaea hort-*
76 *ensis*, *Cepaea nemoralis* (hap1), *Cornu aspersum*, *Deroceras lasithionense*, *Lissachatina immaculata*,
77 *Meghimatium bilineatum* and *Oreohelix idahoensis*. A twelfth genome, the chromosome-level *Achatina*
78 *immaculata* assembly ASM976088v1 (Liu et al., 2021), was used where noted; *Achatina immaculata* and
79 *Lissachatina immaculata* refer to the same species.

80 For genome-level searches, the corresponding Ensembl assemblies were queried through the Ensembl
81 BLAST service, and the *Achatina immaculata* assembly ASM976088v1 and the *Bradybaena similaris*
82 assembly were included for comparison.

83 Gene identification and copy-number assignment

84 Protein sets were searched with HMMER 3 (Eddy, 2011) against Pfam profile HMMs for the Hedge-
85 hog N-terminal domain (PF01085), the Hog domain (PF01079), the Patched domain (PF02460), the
86 Frizzled/Smoothed domain (PF01534) and the Gli zinc-finger domain, using the gathering thresholds
87 (`--cut_ga`). Domain hits were merged per gene rather than per protein, because alternative transcripts of
88 a single gene can produce several apparently independent hits; counting by protein inflates copy number
89 two-fold in several species.

90 *Smoothed* was distinguished from *Frizzled* by reciprocal best hits against human SMO, and con-
91 firmed by bitscore: genuine Smoothed matches scored 481–636 bits, whereas the Frizzled background
92 scored 49–245 bits.

93 Reconstruction of two defective gene models

94 Two *Hedgehog* models were inconsistent with the underlying genomic sequence and were rebuilt. The
95 *Lissachatina immaculata* model ENSLHEG00000006157 encoded only a 1.4 kb terminal fragment; the

Ariolimax columbianus model ENSIBKG00005010863 encoded two separate polypeptides of 197 and 275 amino acids in place of a single gene.

For each locus, the corresponding chromosome interval was extracted and subjected to reference-guided annotation with miniprot 0.17 (protein-to-genome alignment with splice-site awareness), supplemented by tblastn searches against the interval. Exons supported by either method were merged into a single model; the 5' end was extended to the first in-frame ATG downstream of the last upstream stop codon, and the extension length was forced to a multiple of three. Rebuilt models were validated by the absence of internal stop codons, the presence of a start codon, and the presence of a signal peptide and both Hedgehog domains. Corrected models were then substituted into all downstream analyses, which were re-run in full.

Phylogenetic analysis

The data set comprised the 21 in-group *Hedgehog* proteins and 15 outgroups. The outgroups were the three vertebrate Hedgehogs *Homo sapiens* SHH (UniProtKB Q15465), IHH (Q14623) and DHH (O43323); *Drosophila melanogaster* Hh (Q02936); and eleven lophotrochozoan sequences: *Biomphalaria glabrata* (B6RSZ6), *Patella vulgata* (Q8MWG6), *Octopus vulgaris* (A0A4Y1LTG3), *Euprymna scolopes* (Q49QW4), *Antalis entalis* (A0A1J0M5M7), *Nucula tumidula* (A0A1J0M5N3), *Gymnomenia pellucida* (A0A1J0M5N0), two *Magallana gigas* paralogues (A0A482CVZ8 and A0A8W8NMQ6), and *Acanthochitona crinita* Hh (A0A1J0M5P1) with its Lophohh-related paralogue (A0A1J0M5N9). All outgroup accessions are UniProtKB; none was taken from the Ensembl Rapid Release genomes analysed here.

Proteins were aligned with MAFFT 7 (Katoh and Standley, 2013) using the automatic strategy with up to 1,000 refinement iterations (--auto --maxiterate 1000). Maximum-likelihood trees were inferred with IQ-TREE 3.0.1 (Wong et al., 2026); the substitution model and rate heterogeneity were selected by ModelFinder under the Bayesian information criterion, which chose WAG+F+I+G4, and branch support was assessed with 1,000 ultrafast bootstrap replicates and 1,000 SH-aLRT replicates. The tree was rooted on the outgroup block, which was recovered as monophyletic. For display only, the 15 outgroup tips were collapsed to a single tip whose branch is drawn dashed and is not to scale; all inferences used the uncollapsed tree.

Selection analysis

The codon alignment contained the 21 in-group coding sequences only. The outgroups were excluded because no comparably annotated coding sequence is available for them, and because the long-branch outgroups would dominate the rate estimates. Each coding sequence was translated, the translations were aligned with MAFFT 7, and the alignment was back-translated to codons so that every column corresponds to a complete codon; the final alignment is 678 codons long.

Substitution rates were estimated with codeml in PAML 4.10 (Yang, 2007) on a tree pruned of outgroups and rooted at the split between clades A and B. Codon frequencies were estimated from the data (F3x4) and the transition/transversion ratio was estimated by maximum likelihood. Three models were fitted: one-ratio (model 0; a single ω for all branches), two-ratio (model 2; a separate ω for the clade A branches, which were marked as foreground), and free-ratio (model 1; one ω per branch). Model 2 adds one parameter to model 0 and was compared with it by a likelihood-ratio test of twice the log-likelihood difference against a χ^2 distribution with one degree of freedom.

Two further comparisons did not use the codon alignment. Mean within-clade patristic distances were computed for all 45 species pairs. Those 45 ratios are not independent observations, because each of the ten species contributes to nine of them. We therefore tested the ratio on a per-species basis, averaging the nine pair ratios that involve each species so that the ten species enter as ten observations, and used a two-sided sign test on those ten values; the per-pair test is reported only for description. Total intron length per gene, taken from the exon annotation, was compared between clades by a label permutation test.

Functional divergence

Type I (site-specific rate shift) and Type II (site-specific radical change) functional divergence between the two copies were tested with DIVERGE v4 (Chen et al., 2025), which implements the Gu99 and Gu2001 tests for Type I (Gu, 1999; Gu and Vander Velden, 2002) and the Type2 test for Type II. The input was the protein alignment of both copies together with one neighbour-joining tree per copy. Alignment ends were trimmed where more than half the sequences carried a gap, leaving 528 columns for *Hedgehog*

149 and 1,743 for *Patched*. θ was tested by likelihood ratio with one degree of freedom, and sites with a
150 posterior probability of at least 0.5 were counted and mapped to the Hedgehog and Hog domains through
151 a reference sequence.

152 **Genome-internal synteny**

153 Genome-internal collinearity was assessed with MCScan implemented in jcv (Tang et al., 2008). Coding
154 sequences were aligned with LAST in frameshift- tolerant mode (Kielbasa et al., 2011), and anchors were
155 retained after reciprocal-best-hit filtering and block screening. Because reciprocal-best-hit filtering retains
156 only one partner per gene and therefore loses most anciently diverged duplicates, we additionally scanned
157 all significant hits and retained, for each gene, its best non-self hit.

158 For each chromosome pair we computed the longest monotone subsequence of anchors (allowing
159 either orientation) and the fraction of 5 Mb bins on each chromosome touched by that chain. Chromosome
160 pairs were considered chromosome-scale duplicates when both one-sided coverages exceeded 45%. The
161 genome-wide circle plot (Figure 4) was drawn with circlize (Gu et al., 2014).

162 **Micro-synteny at the four loci**

163 For each species, the genomic neighbourhood of each locus was defined as the fifteen genes upstream
164 and downstream. The proteins encoded by these neighbourhoods were searched against the complete
165 proteome with BLASTP, and all significant hits were retained rather than only the best hit. For a pair of
166 loci, the score was the number of significant hits connecting the two neighbourhoods. The null distribution
167 was obtained not by simulation but by exhaustive enumeration: the same score was computed for every
168 window of equal gene number across the genome, and the empirical P -value is the rank of the observed
169 window among all windows.

170 This design was validated with *Patched* as a positive control. Because *Patched* paralogs are known
171 from the chromosome-level analysis to lie on a chromosome-scale duplicate pair, the test must recover a
172 signal at the *Patched* loci; it does. An earlier version of the test that used only each gene's single best hit
173 failed this control and was discarded.

174 **Genome-level copy-number verification**

175 Because proteome-level inventories cannot detect pseudogenes or genes missing from an annotation, copy
176 number was verified directly against genomic sequence. Full-length coding sequences of *Smoothed*
177 (eleven queries) and *Gli* (seven queries) were searched against each target genome with BLASTN through
178 the Ensembl BLAST service ($E \leq 1 \times 10^{-20}$, identity $\geq 70\%$). Hits were clustered within 100 kb on a
179 chromosome, and each cluster was scored for the union of query intervals that it covered. A cluster was
180 accepted as a locus if it covered at least 30% of the query and scored at least 200 bits.

181 Loci were further required to be collinear: hits belonging to a single query were sorted by genomic
182 position, and the Spearman correlation between genomic position and query position was required to
183 exceed 0.9 in absolute value. Collinearity, rather than coverage alone, distinguishes a genuine partial
184 duplicate from scattered domain-level similarity.

185 **Software**

186 Analyses used HMMER 3.3, MAFFT 7, IQ-TREE 3.0.1, PAML 4.10, BLAST+ 2.13 (Camacho et al.,
187 2009), miniprot 0.17 (Li, 2023), jcv 1.6.7 and Python 3.12.

188 **RESULTS**

189 **Two Hedgehog gene models in public annotations are defective**

190 Two of the twenty-two *Hedgehog* models retrieved from Ensembl Rapid Release were inconsistent with
191 their genomic loci (Table 1).

192 The *Lissachatina immaculata* model ENSLHEG00000006157 encoded only a 1.4 kb terminal frag-
193 ment. Reference-guided reannotation of the locus recovered a single gene spanning 188 kb and encoding
194 473 amino acids, with a start codon, no internal stop codons, a predicted signal peptide, and both the
195 Hedgehog and Hog domains.

196 The *Ariolimax columbianus* model ENSIBKG00005010863 had been split into two models encoding
197 197 and 275 amino acids respectively. Reannotation recovered a single gene of 74 kb encoding 481 amino
198 acids, again complete and internally consistent.

Neither locus is a pseudogene. Both are full-length, and the apparent degradation was an artefact of annotation. All downstream analyses were repeated with the corrected models; the qualitative conclusions were unchanged and the statistical support generally increased.

Species	Original model	Defect	Rebuilt model
<i>Lissachatina immaculata</i>	ENSLHEG00000006157	1.4 kb terminal fragment	188 kb, 473 aa
<i>Ariolimax columbianus</i>	ENSIBKG00005010863	split into 197 + 275 aa	74 kb, 481 aa

Table 1. Two defective *Hedgehog* models and their reconstruction.

Copy number across the pathway

Across the eleven Ensembl genomes, *Hedgehog* and *Patched* were each present in two copies in every species in which both could be assessed, whereas *Smoothened* was present in a single copy in all eleven (Table 2). *Gli* was detected in only six of eleven proteomes; we return to this below, because the genome-level search shows that the proteome inventory is incomplete.

Within each species the two *Hedgehog* copies differ in length and in domain composition only trivially; both retain the Hedgehog and Hog domains. The two *Patched* copies likewise both retain the Patched domain.

Species	<i>Hedgehog</i>	<i>Patched</i>	<i>Smoothened</i>	<i>Gli</i>
<i>Albinaria teres</i>	2	2	1	0*
<i>Ariolimax columbianus</i>	2	2	1	1
<i>Arion vulgaris</i>	1 [†]	2	1	1
<i>Candidula unifasciata</i>	2	2	1	0*
<i>Cepaea hortensis</i>	2	2	1	1
<i>Cepaea nemoralis</i>	2	2	1	1
<i>Cornu aspersum</i>	2	2	1	0*
<i>Deroceras lasithionense</i>	2	2	1	1
<i>Lissachatina immaculata</i>	2	2	1	0*
<i>Meghimatium bilineatum</i>	2	2	1	0*
<i>Oreohelix idahoensis</i>	2	2	1	1

Table 2. Copy number of Hedgehog pathway components in eleven Stylommatophoran proteomes. *No hit in the proteome; genome-level searches show that at least one of these species retains a second *Gli* copy (see below). [†]Only one copy retrieved from the proteome; the second is not represented in this annotation.

One further genome, *Bradybaena similaris*, was searched but is not part of the eleven-genome survey. Its NCBI proteome (29,227 proteins) contains no *Hedgehog* at all: no Hedge domain hit reaches the gathering threshold, and the single Hog hit is an AIG1+Hint fusion rather than a Hedgehog. A tblastx search of the same assembly recovered two complete *Hedgehog* genes, on chromosomes 6 and 23 ($E = 3.2 \times 10^{-130}$ and 8.7×10^{-129}), each covering both the Hedge and the Hog domain. This is the same two-chromosome pattern as in *Albinaria teres*, and it shows that the absence from the proteome is an annotation failure rather than a loss of the gene.

The two Hedgehog paralogs form two clades with asymmetric constraint

The two *Hedgehog* copies from ten species with both copies resolved into two deeply separated clades, which we refer to as A and B (Figure 1). Every species with two copies contributes one sequence to each clade; no species has two sequences in the same clade. This topology is consistent with a single duplication predating the diversification of the sampled species, rather than with independent lineage-specific duplications.

The two clades differ in constraint, and the difference is a shift of the whole clade rather than of a few branches. Clade A is farther from clade B in every one of the 45 species pairs (geometric mean $d_A/d_B = 1.70$), and because those pairs are not independent the ratio was also averaged within each species: all ten species-level means exceed one (range 1.32–2.02, median 1.86; two-sided sign test

227 $p = 0.0020$). The free-ratio model gives median per-branch d_N/d_S of 0.208 for clade A and 0.148 for
 228 clade B, and the two-ratio model gives 0.222 versus 0.156 ($2\Delta\ln L = 25.21$, $p = 5.1 \times 10^{-7}$; Figure 2).
 229 Gene structure differs in the same direction. The median individual intron is 6,748 bp in clade A and
 230 22,442 bp in clade B, a 3.33-fold difference; the median total intron length per gene is 27.9 kb against
 231 88.8 kb, a 3.18-fold difference (label permutation test $p = 0.0007$; Figure 1B). The two ratios are quoted
 232 separately because they are different statistics on the same data, not two estimates of one number.

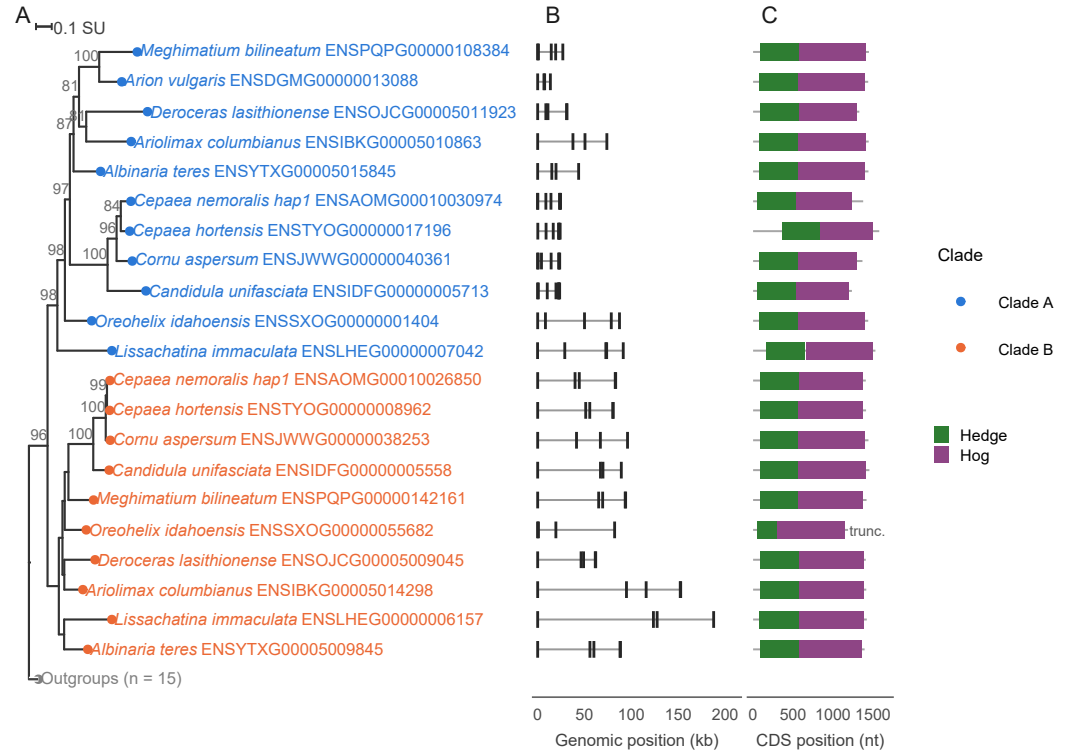


Figure 1. Phylogeny, exon distribution and domain architecture of the 21 Stylommatophoran *Hedgehog* paralogs. (A) Maximum-likelihood phylogram inferred with IQ-TREE 3 under WAG+F+I+G4 from 21 in-group and 15 outgroup proteins and rooted on the outgroups; outgroup accessions are listed in the Methods. The 15 outgroup tips are collapsed to a single tip for display and that branch is dashed and not to scale. Tip labels give the species (*italic*) and the Ensembl gene identifier; colour denotes clade A or B. Node labels are ultrafast bootstrap percentages, shown where ≥ 80 ; the scale bar is in substitutions per site (SU) and applies to the in-group. (B) Exon positions on the genomic axis; each vertical tick marks one exon and the grey line spans the annotated gene, so the gaps between ticks are introns. (C) The Hedge (PF01085) and Hog (PF01079) domains on the coding-sequence axis. Rows follow the tip order of the tree. Tick and box widths are fixed and encode nothing; positions are to scale. The Hedge box of *Oreohelix idahoensis* ENSSXOG00000055682 is marked “trunc.”: its canonical transcript lacks a 5' UTR, begins with TTT rather than ATG and yields an 84-aa Hedge domain, so the short box reflects a 5'-truncated annotation rather than an in-frame deletion.

233 We note that an earlier version of this analysis, performed before the two models above were rebuilt,
 234 attributed the difference to a small number of high- d_N/d_S branches driven by the two fragments. After
 235 reconstruction the difference is visible in the median and not in the upper tail, and the conclusion is
 236 correspondingly simpler and stronger.

237 By contrast, both *Patched* paralogs are strongly constrained. On alignments built the same way and
 238 fitted with the same one-ratio model, ω is 0.116 for the A copy and 0.126 for the B copy, against 0.209
 239 and 0.159 for the two *Hedgehog* clades and 0.148 for *Smoothed* (Figure 2C). Both *Patched* copies
 240 therefore sit below both *Hedgehog* copies. Fitting a two-ratio model to the two *Patched* copies jointly (20
 241 sequences, 5,928 codons, one sequence per species per copy) gives $\omega = 0.134$ for the A copy and 0.104

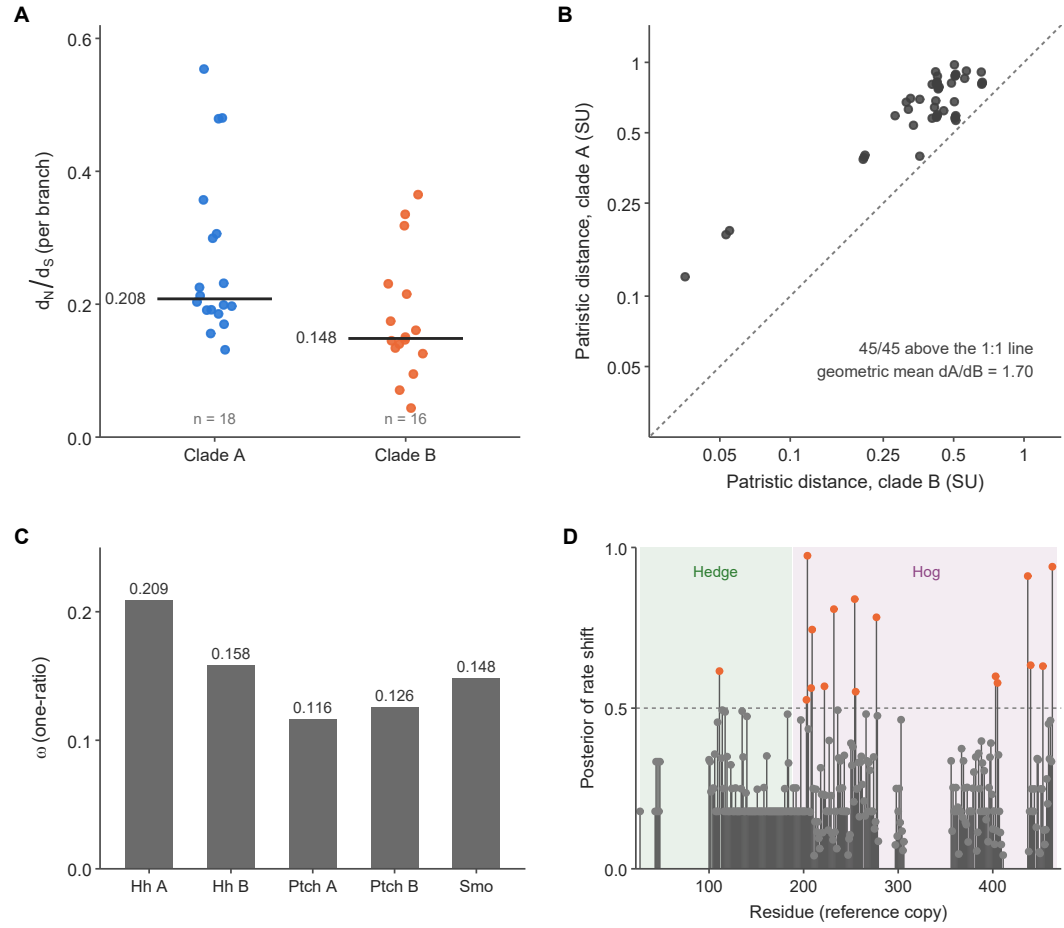


Figure 2. Selection on the two *Hedgehog* clades and on the pathway they belong to. (A) Per-branch d_N/d_S under the free-ratio model (PAML model 1); each point is one branch and the horizontal bar is the clade median (clade A, $n = 18$; clade B, $n = 16$). (B) Patristic distances of the two clades for all 45 species pairs, on log-log axes; the dashed line is the 1:1 expectation. Every pair lies above it, with a geometric mean d_A/d_B of 1.70. Because the 45 pairs are not independent, the test is on the ten species-level means (median 1.86, two-sided sign test $p = 0.0020$). (C) One-ratio ω for each gene set, estimated on alignments built the same way and each containing one sequence per species. Removing *Oreohelix*, whose *Patched* models are fragments, gives 0.206, 0.153, 0.114, 0.130 and 0.148 respectively and does not change the ordering. (D) Posterior probability of a Type I rate shift at each alignment column, plotted against the corresponding residue of the reference copy, with the Hedgehog and Hog domains shaded. Orange points exceed 0.5. The two *Patched* copies have no column above 0.5 and are not plotted.

for the B copy, against a single 0.118 for both ($2\Delta\ln L = 22.68$, $df = 1$, $p = 9.6 \times 10^{-7}$). The asymmetry is therefore real, but it is smaller than at *Hedgehog*: the ratio between the two copies is 1.28 against 1.42, and the absolute difference is 0.030 against 0.046.

Site-level analysis separates the two duplicates further. Between the two *Hedgehog* clades there is significant Type I divergence, a shift in evolutionary rate at a subset of sites ($\theta = 0.246 \pm 0.052$; LRT = 22.46, $df = 1$, $p = 2 \times 10^{-6}$), and 16 of the 274 scored alignment columns exceed a posterior probability of 0.5. Fifteen of those 16 lie in the Hog domain, against a background of 66% of scored columns (binomial $p = 0.011$; Figure 2D), so the rate shift is concentrated in the C-terminal autoprocessing domain rather than in the N-terminal signalling domain. Type II divergence is not detected ($\theta_{II} = 0.016 \pm 0.076$, $p = 0.18$). Between the two *Patched* copies neither type is detected (Type I $\theta = 0.001$, LRT ≈ 0 ; Type II $\theta_{II} = 0.001 \pm 0.079$, $p = 0.066$). The divergence between the two *Hedgehog* copies is therefore not uniform along the protein but concentrated in one domain, whereas the two *Patched* copies show no site-level divergence of either type. Removing *Oreohelix*, whose two *Patched* models are fragments, changes these values by less than 0.005 and does not change the ordering.

The *Patched* paralogs lie in a chromosome-scale duplication; the *Hedgehog* paralogs do not

Genome-internal collinearity was assessed for all 22 chromosomes of *Cepaea nemoralis* and *Cepaea hortensis*. Three chromosome pairs in each species showed two-sided collinear coverage above 45%, spanning nearly the whole of both chromosomes and in inverted orientation (Figure 3). One of these pairs carries the two *Patched* paralogs in each species: *C. nemoralis* OZ071526 and OZ071532 (74% and 95% coverage), and *C. hortensis* OY992885 and OY992893 (78% and 80%). The two species are separate assemblies, so the association is not an artefact of one assembly; they are sister species, however, so this is not independent phylogenetic replication.

The two chromosomes carrying the *Patched* paralogs share 8,361 significant hits in *C. nemoralis*, and a chain of 160 collinear anchors traces an inverted path across their full length.

The *Hedgehog* paralogs show a much weaker, but not absent, association. In *C. nemoralis* they lie on OZ071525 and OZ071538, a pair with 118 best hits, a longest ordered chain of 22 anchors among those best hits, and 20% and 53% two-sided coverage. This is roughly a quarter of the coverage seen at *Patched*, and individual anchor segments do connect the two loci (Figure 3, bottom row; Figure 4), but no chain spans the chromosomes.

We emphasise that this is a quantitative difference in retained collinearity, not evidence that the two *Hedgehog* paralogs arose by a mechanism different from the one that produced the *Patched* paralogs. A duplication shared by the two chromosomes, followed by faster local evolution or more extensive anchor loss around the *Hedgehog* locus, would give the same pattern. Distinguishing the two explanations would require an independent anchor set, for example from a more closely related outgroup or from long-read assemblies of additional *Cepaea* populations.

Micro-synteny confirms the contrast

To ask whether the neighbourhoods of the loci are conserved, we compared the fifteen genes flanking each locus with every other window of equal size in the genome, using all significant protein hits and an exhaustively enumerated null distribution.

The test has power at the *Patched* loci. The score exceeds the genome-wide mean in all nine genomes examined and does so significantly in seven (empirical $p \leq 0.0035$, enrichment 31- to 406-fold). At the *Hedgehog* loci the same score is significant in one of eight genomes (*Deroceras lasithionense*, $p = 0.0027$) and marginal in one more (*Cornu aspersum*, $p = 0.039$); in the remaining six it does not separate from the null (Table 3).

The neighbourhoods of the *Patched* duplicates are therefore conserved far more often than those of the *Hedgehog* duplicates. As above, this is a statement about retained sequence similarity, not about the duplication mechanism: the same contrast would arise if the *Hedgehog* region had evolved faster after a shared duplication.

Genome-level copy number: *Smoothed* is single-copy, *Gli* is not

Because proteome-level inventories cannot detect unannotated or degenerating copies, we searched full-length coding sequences against genomic sequence.

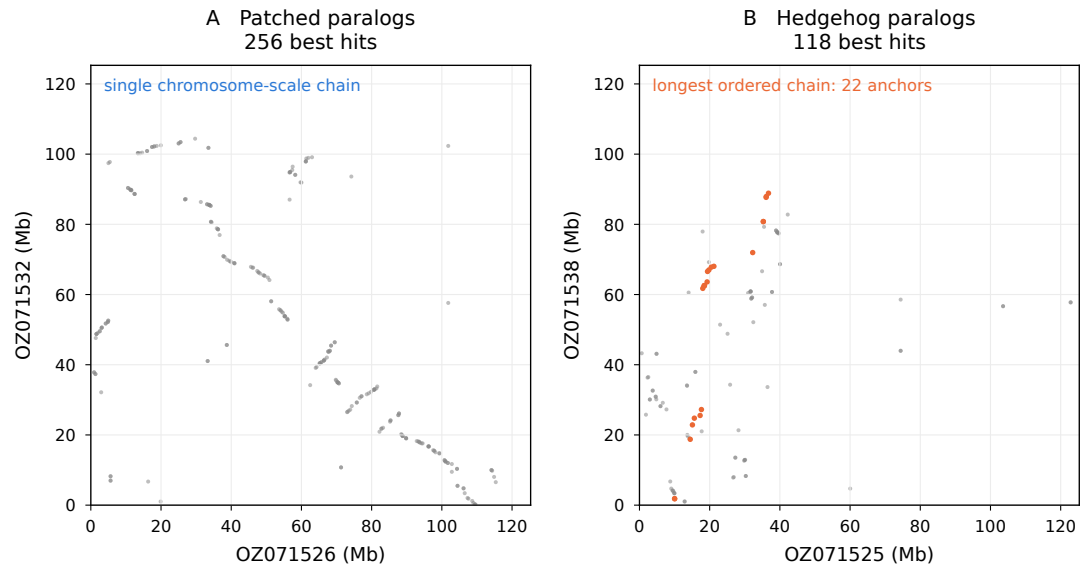


Figure 3. Chromosome-pair collinearity in *Cepaea nemoralis*. (A) The *Patched*-bearing pair OZ071526 and OZ071532, showing a chromosome-scale inverted collinear chain (74% and 95% two-sided coverage). (B) The *Hedgehog*-bearing pair OZ071525 and OZ071538, which retains sparse collinearity (20% and 53% coverage) but no chromosome-spanning chain. Axes are gene order and both panels share one scale; grey points are best non-self protein hits. In (B) the orange points are the longest ordered chain, 22 of the 118 hits, which is what distinguishes reduced collinearity from absent collinearity.

Species	<i>Hedgehog</i>			<i>Patched</i>		
	Score	Enrichment	<i>p</i>	Score	Enrichment	<i>p</i>
<i>Albinaria teres</i>	9	4.6	0.0519	18	39.5	0.0014
<i>Ariolimax columbianus</i>	7	2.0	0.0577	19	30.9	0.00352
<i>Arion vulgaris</i>	–	–	–	30	71.9	0.000271
<i>Cepaea hortensis</i>	2	1.0	0.0645	37	82.6	0.000823
<i>Cepaea nemoralis hap1</i>	3	1.1	0.0831	93	289.3	0.00029
<i>Cornu aspersum</i>	5	1.4	0.0393	61	406.3	0.000436
<i>Deroceras lasithionense</i>	11	31.1	0.00266	9	16.8	0.0192
<i>Lissachatina immaculata</i>	0	0.0	1	42	152.4	0.000249
<i>Meghimatium bilineatum</i>	32	1.1	0.158	26	1.9	0.0877

Table 3. Micro-synteny at the two loci. For each genome the fifteen genes flanking the locus were used as a query against the whole proteome, and the number of significant hits landing in the orthologous window (score) was compared with every window of equal size in the same genome. Enrichment is the score divided by the genome-wide mean; *p* is the fraction of windows scoring at least as high. Dashes: no search available for that genome and locus.

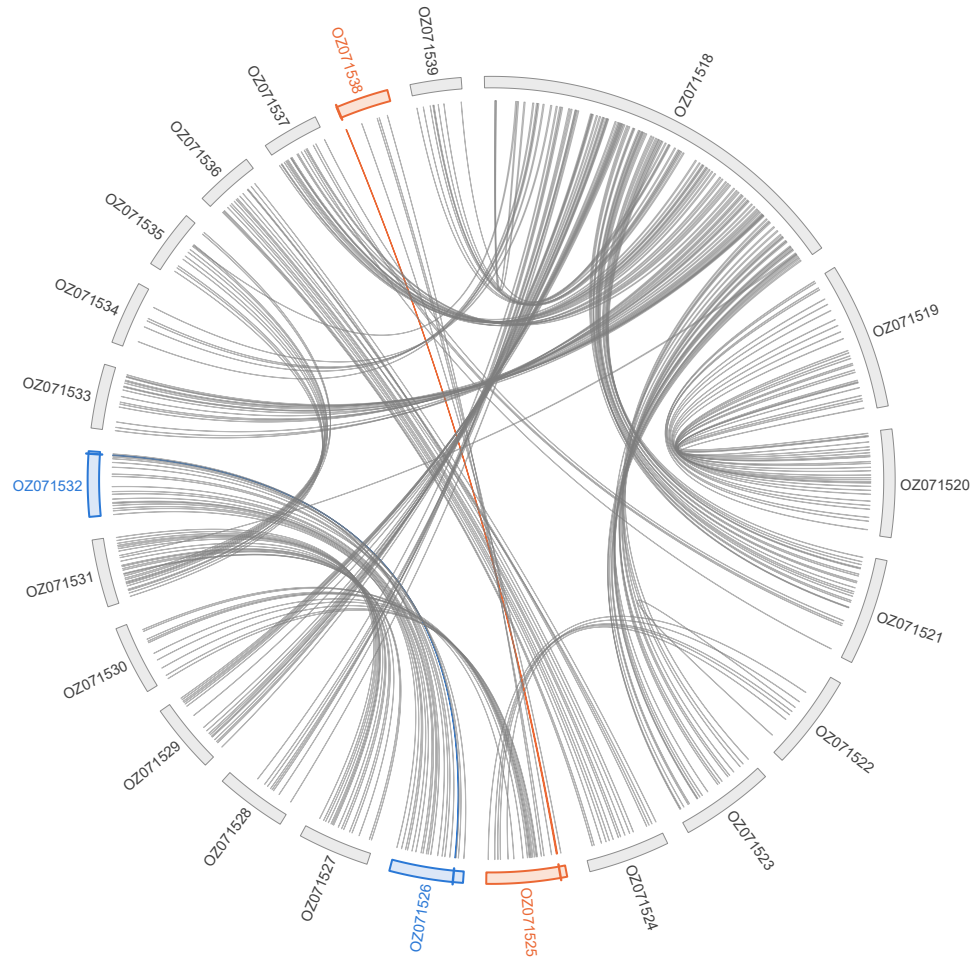


Figure 4. Genome-wide collinearity in *Cepaea nemoralis*. The 22 chromosomes are drawn in proportion to their length. Grey ribbons are the 219 collinear blocks detected between them; the average block is 55 kb, so each ribbon is drawn with a minimum width to stay visible and ribbon width is not to scale. Orange: the two *Hedgehog*-bearing chromosomes OZ071525 and OZ071538, the *Hedgehog* positions (ticks) and the blocks connecting the two loci. Blue: the same for the two *Patched*-bearing chromosomes OZ071526 and OZ071532. The coloured blocks are the gene windows; the chromosome-scale collinear chains are shown in Figure 3.

For *Smoothened*, eleven queries were searched against twelve target genomes. Ten of the eleven queries recovered exactly one collinear locus in every target. The single exception, *Cepaea nemoralis*, showed two loci with identical coverage, identity and score for every query (for example 94% coverage at 67% identity from 65 high-scoring pairs at each of LG14 and LG18), indicating that the same sequence is represented twice in the assembly rather than that a second gene is present. The eleventh query (*Cepaea hortensis* ENSTYOG00000010097) behaved anomalously, returning hundreds of weak partial matches scattered across every genome including its own; because the other ten queries each recover a single locus in *C. hortensis*, the anomaly lies in that gene model and not in the genome. *Smoothened* is therefore single-copy across all twelve assemblies examined.

An independent search of the same eleven queries against the same twelve assemblies, run as tblastx through the Ensembl web service with low-complexity masking and scored on the same criteria, reproduced this result exactly: one locus per query–target pair, two loci in *Cepaea nemoralis*, and the same anomalous behaviour of ENSTYOG00000010097.

For *Gli*, seven queries were searched against twelve target genomes. Most targets showed one collinear locus (Table 4), but *Lissachatina immaculata* (the chromosome-level assembly ASM976088v1, published as *Achatina immaculata*) showed two: a forward-oriented locus on LG1 and an inverted locus on LG12, each recovered by all seven queries independently (LG1 $\rho = +0.99$ to $+1.00$; LG12 $\rho = -0.97$ to -1.00 ; Figure 5). The LG12 locus covers only 33–47% of the query, and its exons are arranged in order over approximately 35 kb of inverted sequence. It is therefore a partially degenerated second copy.

This result directly contradicts the proteome-level inventory, in which *Lissachatina immaculata* had no *Gli* at all. Several other target genomes showed weaker second signals whose interpretation is confounded by haplotype duplication in the assembly; these require case-by-case treatment and we do not interpret them here.

Genome	<i>Smoothened</i>	<i>Gli</i>
<i>Albinaria teres</i>	1	1
<i>Ariolimax columbianus</i>	1	1
<i>Arion vulgaris</i>	1	1
<i>Bradybaena similaris</i>	1	1
<i>Candidula unifasciata</i>	1	1
<i>Cepaea hortensis</i>	1	1
<i>Cepaea nemoralis</i>	2 [†]	2 [†]
<i>Cornu aspersum</i>	1	1
<i>Deroceras lasithionense</i>	1	1
<i>Lissachatina immaculata</i>	1	2 [‡]
<i>Meghimatium bilineatum</i>	1	1
<i>Oreohelix idahoensis</i>	1	1

Table 4. Genome-level copy number from sequence searches. Eleven *Smoothened* queries and seven *Gli* queries were searched against each assembly; the table gives the number of collinear loci recovered, which was the same for every query. The eleven Ensembl assemblies supply ten targets; the chromosome-level *Lissachatina immaculata* assembly ASM976088v1 stands in for that species, and *Bradybaena similaris* was added for comparison. [†]Both loci carry the same sequence with identical scores, so the gene is represented twice in the assembly rather than duplicated in the genome. [‡]The second locus is inverted, on LG12, and covers 33–47% of the query: a partially degenerated copy.

DISCUSSION

A ligand duplicate and a receptor duplicate followed different paths

The central observation of this study is that two duplicated components of the same signalling pathway differ in both constraint and the amount of collinearity they have retained (Figure 6). The two *Hedgehog* paralogs have diverged in constraint, with one clade consistently relaxed across all sampled species, and they sit on a chromosome pair that retains only sparse collinearity and almost no conserved micro-synteny. The two *Patched* paralogs are both strongly constrained and much closer to each other than the two *Hedgehog* clades, sit on a chromosome-scale duplicate pair in two independent species, and retain conserved neighbourhoods in almost every genome examined.

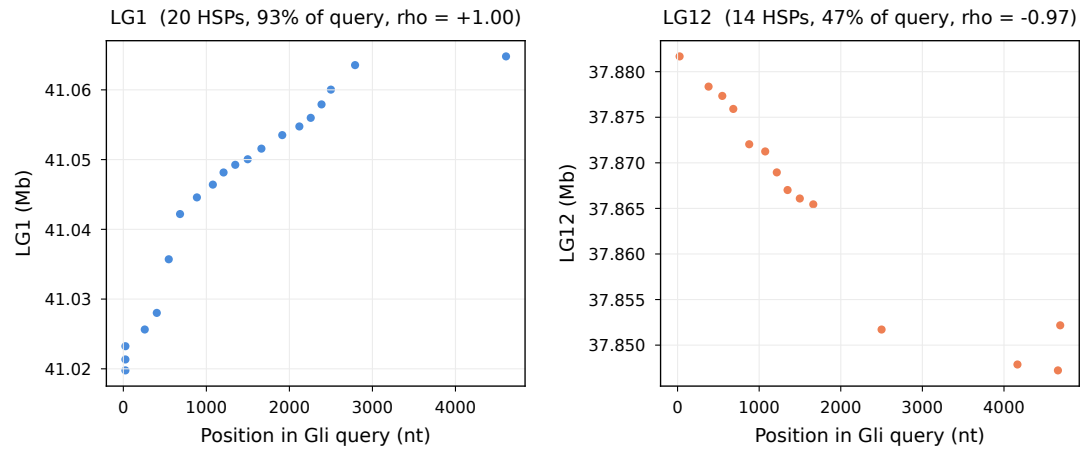


Figure 5. Two collinear *Gli* loci in the *Achatina immaculata* assembly. Hits of one *Gli* query (*Ariolimax columbianus*) are plotted against genomic position. LG1 is forward and covers 93% of the query; LG12 is inverted, covers 47%, and its high-scoring pairs are ordered along the query, so it is a partially degenerated copy rather than a repeat. All seven queries give Spearman ρ between +0.99 and +1.00 on LG1 and between -0.97 and -1.00 on LG12.



Figure 6. What happened to each component of the pathway. The four components are drawn in signalling order: Hh binds Ptch, Ptch inhibits Smo, and Smo allows Gli to act. Below each are the copies retained (blue, clade A; orange, clade B; grey, single copy; hollow, a partial second copy), the one-ratio ω , and the genomic context. The ω values are those of Figure 2C and the copy numbers those of 4.

326 One reading of this difference is that the two duplications had different mechanisms: a chromosome-
327 scale duplication—whether a whole-genome duplication or a large segmental event—copies the receptor
328 together with a block of flanking genes, which is the pattern we observe at *Patched*, whereas a duplication
329 of the ligand alone would leave the locus without conserved neighbourhoods, which is the pattern we
330 observe at *Hedgehog*.

331 We do not think the data compel that reading. Collinearity at the *Hedgehog* loci is reduced, not absent:
332 22 anchors still form a chain between the two chromosomes, and the loci themselves are joined by a direct
333 anchor. Faster local evolution, or a more complete loss of anchors around the *Hedgehog* locus after a
334 shared duplication, would produce the same quantitative difference. What the data establish is a contrast
335 in retained synteny; what produced it remains open.

336 **Constraint, dosage and the single-copy status of *Smoothened***

337 The gene-balance hypothesis predicts that components of stoichiometric complexes and dose-sensitive
338 signalling cascades are preferentially retained after duplication (Birchler and Veitia, 2012; Veitia et al.,
339 2008). Our data are consistent with a graded outcome rather than a binary one. *Smoothened*, the transducer
340 that is rate-limiting for pathway activity, is strictly single-copy across twelve genomes. *Patched*, the
341 receptor, retains two constrained copies. *Hedgehog*, the ligand, retains two copies of which one has
342 relaxed.

343 We emphasise that this is a pattern, not a mechanism. We have no expression data, no functional assays,
344 and no measure of protein abundance. The single-copy status of *Smoothened* is a robust observation,
345 replicated across genomes, but its cause—whether selection against duplicate retention, repeated loss, or
346 simply the absence of the duplication—cannot be distinguished with the data at hand.

347 A useful next step would be to determine whether *Smoothened* is single-copy outside the Stylom-
348 matophora as well. If single-copy status is general across molluscs, while *Patched* and *Hedgehog* are
349 repeatedly duplicated, the Stylommatophoran case becomes an instance of a general rule rather than an
350 idiosyncrasy.

351 **What can and cannot be said about the whole-genome duplication**

352 Our synteny results are consistent with the WGD hypothesis. In two *Cepaea* species, the *Patched*-
353 bearing chromosomes are chromosome-scale duplicates of one another, and the pattern is inverted and
354 near-complete, which is what a whole-chromosome duplication predicts.

355 We do not, however, claim to have demonstrated a WGD. Three limitations apply. First, only three
356 chromosome pairs in each *Cepaea* species met our coverage criterion, whereas a WGD would predict
357 approximately eleven; decay and rearrangement can account for this, but we cannot distinguish decay
358 from the possibility that only a subset of chromosomes was involved. Second, our attempt to date the
359 duplication by K_s was uninformative: the synonymous distance between duplicate pairs has a median
360 of 1.81, deep in the saturation regime, and no age estimate from our data can resolve a peak reliably.
361 Third, the WGD hypothesis itself is contested in the literature, on grounds that turn on the interpretation
362 of exactly the K_s signal we found uninformative (Nath et al., 2025).

363 We therefore describe the pattern as a chromosome-scale duplication, note its consistency with the
364 proposed Stylommatophoran WGD, and leave the question open.

365 **Annotation artefacts as a recurring obstacle**

366 A substantial part of this study consisted of detecting and correcting errors in public annotations. Two
367 *Hedgehog* models were truncated or split; a third species (*Bradybaena similaris*) had no *Hedgehog* in
368 its proteome despite carrying two complete genes in its genome; and *Gli*, reported as single-copy from
369 proteomes, has a second partially degenerated copy in *Lissachatina immaculata*.

370 These are not isolated curiosities. They share a common structure: a gene is absent or truncated in a
371 proteome, and a proteome-level inventory concludes that the gene is absent or degenerate. In each case
372 the conclusion is wrong, and in each case the error is detectable by returning to the genome.

373 We can offer no estimate of how common such errors are, and we do not attempt one. What we
374 can say is that in the four gene families examined here, across eleven genomes, annotation defects were
375 frequent enough to change conclusions about copy number in at least two cases. Gene-family studies
376 that rely on proteome inventories alone should verify their focal genes against genomic sequence before
377 drawing conclusions about copy number or pseudogenisation.

378 We also note two procedural lessons that materially affected our own results. First, counting by protein
379 rather than by gene inflates copy number when alternative transcripts are present; in one species four
380 Hedgehog domain hits corresponded to two genes. Second, reciprocal-best-hit filtering, the standard route
381 to syntenic anchors, retains only one partner per gene and therefore preferentially loses anciently diverged
382 duplicates; in one genome it retained 4% of the pairs found by a best-hit-per-gene scan. Studies of old
383 duplications that rely on reciprocal-best-hit anchors alone will systematically underestimate the number
384 of retained duplicates.

385 Limitations

386 Our conclusions rest on eleven to twelve genomes, unevenly distributed across Stylommatophora. *Ore-*
387 *ohelix idahoensis* was represented by a fragmented assembly and could not be included in the synteny
388 analysis at the gene level. *Candidula unifasciata* was excluded from all synteny analysis because its
389 assembly is too fragmented to support collinear blocks: 4,138 sequences, a median of five genes per
390 sequence, and a largest sequence containing only 105 genes.

391 We have no expression data. The published RNA-Seq for the focal species consists of a single
392 adult-foot sample, which is not informative about Hedgehog-pathway activity. We therefore cannot
393 distinguish relaxed selection from loss of function, although the absence of stop codons and the retention
394 of both domains in every copy argues against pseudogenisation.

395 Finally, our genome-level copy-number searches used BLASTN, which is insensitive to highly
396 diverged pseudogene remnants. A search at the translated level (tblastx) may recover additional copies,
397 and the results reported here should be treated as a lower bound on copy number.

398 CONCLUSIONS

399 Within a single signalling pathway in land snails and slugs, duplicated genes followed different trajectories.
400 *Patched* retains two constrained copies that lie on a chromosome-scale duplicate pair and keep conserved
401 neighbourhoods across most genomes examined. *Hedgehog* retains two copies whose collinearity is
402 much more eroded, and one of its clades evolves under consistently relaxed constraint, with the site-level
403 rate shift concentrated in the Hog domain. The two *Patched* copies show no such site-level divergence.
404 *Smoothened* remains single-copy across every genome examined. Annotation errors affect all of these
405 conclusions and must be corrected before copy number is interpreted.

406 USE OF AI-ASSISTED TOOLS

407 Claude Code (Anthropic), an AI coding assistant, was used in the preparation of this work. It assisted in
408 drafting and editing the manuscript text, in writing and debugging the analysis code, in generating the
409 figures, and in deriving and checking the quantitative results reported here. All content was reviewed by
410 the author, who takes full responsibility for the work and for its conclusions; the tool is not an author and
411 cannot take responsibility for it.

412 FUNDING

413 This work received no specific funding.

414 COMPETING INTERESTS

415 The author declares no competing interests.

416 AUTHOR CONTRIBUTIONS

417 Zhi-Jie Xu conceived the study, performed all analyses and wrote the manuscript.

418 DATA AVAILABILITY

419 All scripts, intermediate results and figure-generation code are archived at Zenodo, <https://doi.org/10.5281/zenodo.23471279>. Genome assemblies and annotations are available from En-
420 sembl Rapid Release and NCBI.
421

REFERENCES

- Birchler, J. A. and Veitia, R. A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences*, 109(37):14746–14753.
- Briscoe, J. and Théron, P. P. (2013). The mechanisms and functions of Hedgehog signalling across the metazoa. *Nature Reviews Genetics*, 14(6):418–429.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421.
- Chen, Y., Xu, X., Pan, Y., Wang, S., Zhao, W., Zhou, B., Zhou, J., Zheng, Y., Zhou, Z., and Gu, X. (2025). DIVERGE v4: a platform for large-scale analysis of functional divergence across multi-gene families. *Molecular Biology and Evolution*, 42(11):msaf277.
- Chen, Z., Doğan, Ö., Guiglielmoni, N., Guichard, A., and Schrödl, M. (2022). Pulmonate slug evolution is reflected in the de novo genome of *Arion vulgaris* Moquin-Tandon, 1855. *Scientific Reports*, 12:14226.
- Eddy, S. R. (2011). Accelerated Profile HMM searches. *PLoS Computational Biology*, 7(10):e1002195.
- Gu, X. (1999). Statistical methods for testing functional divergence after gene duplication. *Molecular Biology and Evolution*, 16(12):1664–1674.
- Gu, X. and Vander Velden, K. (2002). DIVERGE: phylogeny-based analysis for functional–structural divergence of a protein family. *Bioinformatics*, 18(3):500–501.
- Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize implements and enhances circular visualization in R. *Bioinformatics*, 30(19):2811–2812.
- Ingham, P. W. and McMahon, A. P. (2001). Hedgehog signaling in animal development: paradigms and principles. *Genes & Development*, 15(23):3059–3087.
- Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.
- Kielbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, 21(3):487–493.
- Li, H. (2023). Protein-to-genome alignment with miniprot. *Bioinformatics*, 39(1):btad014.
- Liu, C., Ren, Y., Li, Z., Hu, Q., Yin, L., Wang, H., Qiao, X., Zhang, Y., Xing, L., Xi, Y., Jiang, F., Wang, S., Huang, C., Liu, B., Liu, H., Wan, F., Qian, W., and Fan, W. (2021). Giant African snail genomes provide insights into molluscan whole-genome duplication and aquatic–terrestrial transition. *Molecular Ecology Resources*, 21(2):478–494.
- McHale, F., Mulhair, P. O., and Holland, P. W. H. (2025). Evolution of duplicated Hox gene clusters in land snails and slugs. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 344(6):363–368.
- Nath, R., Panda, B., Rakesh, S., and Krishnan, A. (2025). Lineage-specific Class-A GPCR dynamics reflect diverse chemosensory adaptations in Lophotrochozoa. *Molecular Biology and Evolution*, 42(3):msaf042.
- Nüsslein-Volhard, C. and Wieschaus, E. (1980). Mutations affecting segment number and polarity in *Drosophila*. *Nature*, 287(5785):795–801.
- Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science*, 320(5875):486–488.
- Veitia, R. A., Bottani, S., and Birchler, J. A. (2008). Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends in Genetics*, 24(8):390–397.
- Winingear, S., Outerbridge, M. E., Garcia, G., Stone, A. C., Wilson, M. A., and Polly, P. D. (2026). Genome report: de novo genome assembly of the greater Bermuda land snail, *Poecilozonites bermudensis* (Mollusca: Gastropoda), confirms ancestral genome duplication. *G3: Genes, Genomes, Genetics*, 16(2):jkaf291.
- Wong, T. K. F., Ly-Trong, N., Ren, H., Demotte, P., Baños, H., Roger, A. J., Susko, E., Bielow, C., De Maio, N., Goldman, N., Hahn, M. W., Dos Reis, M., Vinh, L. S., Huttley, G., Lanfear, R., and Minh, B. Q. (2026). IQ-TREE 3: phylogenomic inference software using complex evolutionary models. *Molecular Biology and Evolution*, 43(5):msag117.
- Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.