Selection on many loci drove the origin and spread of a key innovation

Authors: Sean Stankowski[1,2]*$, Zuzanna B. Zagrodzka[1], Martin Garlovsky[3], Arka Pal[2], Daria Shipilina[2,4], Diego Garcia Castillo[2], Alan Le Moan[5,6], Erica Leder[6], James Reeve[6], Kerstin Johannesson[6], Anja M. Westram[2,7], Roger K. Butlin[1,6].

Affiliations:

[1]1Ecology and Evolutionary Biology, School of Biosciences, University of Sheffield, Sheffield, UK

[2] Institute of Science and Technology Austria (IST Austria), Klosterneuburg, Austria

[3] Technische Universität Dresden, 01069 Dresden, Germany

[4]Department of Ecology and Genetics, Program of Evolutionary Biology, Uppsala University, Uppsala, Sweden

[5]CNRS & Sorbonne Université, Station Biologique de Roscoff, Roscoff, France

[6] Department of Marine Sciences, Tjärnö, University of Gothenburg, Strömstad

[7] Faculty of Biosciences and Aquaculture, Nord University, N-8049 Bodø, Norway

* Corresponding author: sean.stankowski@ista.ac.at

Abstract: Key innovations have played a central role in the origins of biodiversity, but their evolutionary origin and genetic architecture are usually unknown. A recent transition from egg-laying to live-birth in *Littorina* snails provides a rare opportunity to study the origin and genetic architecture of a young innovation. While recognized as one species, live-bearing individuals do not form a single clade in a genome-wide phylogenetic analysis, hinting at two independent origins. However, local genealogical analysis identified numerous genomic regions where samples group according to their reproductive mode. These regions are widespread across the genome, show clear evidence for live-bearer-specific positive selection, and are enriched for genes that are differentially expressed between egg-laying and live-bearing reproductive tissues. Thus, our results show that key innovations can have a polygenic basis, and that their historical origins can be obscured by a complex demographic history.

Main text: Evolution is a gradual process, but occasionally results in sudden changes in form and function that allow organisms to exploit new ecological opportunities (*1*, *2*). These game-changing adaptations, known as 'key innovations', are all around us: they have been crucial in driving major evolutionary transitions and catalyzing the diversification of many groups (*1*, *3*). Despite their significance, we know surprisingly little about the evolutionary origins and genetic architecture of most innovations (*1*). This is because most originated deep in the past, making it difficult to disentangle causal loci from the countless genetic changes that accumulated up to the present.

A recent transition from egg-laying to live-birth provides a rare opportunity to study the genetic architecture of a well-studied innovation whose origins and genetic basis are not well documented (*4*). We focus on a clade of intertidal gastropods (Genus *Littorina*), where the ancestral state is to lay a large egg-mass but one species gives birth to live young (Fig. 1, fig. S1) (*5*, *6*). Egg-layers have a gland that embeds egg-capsules into a protective jelly. In the live-bearer, *L. saxatilis*, this structure has evolved into a brood pouch where embryos develop inside the mother. Live birth is a recent innovation in the Littorinidae, considered key to the much broader geographic and ecological distribution of *L. saxatilis* compared to all egg-laying

Littorina (*6*). Egg-laying and live-bearing species have adapted in parallel to contrasting environments (*6*), partly decoupling reproductive mode from other axes of phenotypic divergence (Fig. 1B). There is also evidence for occasional hybridization between egg-layers and live-bearers (*9*). These features provide an opportunity to identify and study the genetic changes underlying the live-bearing innovation.

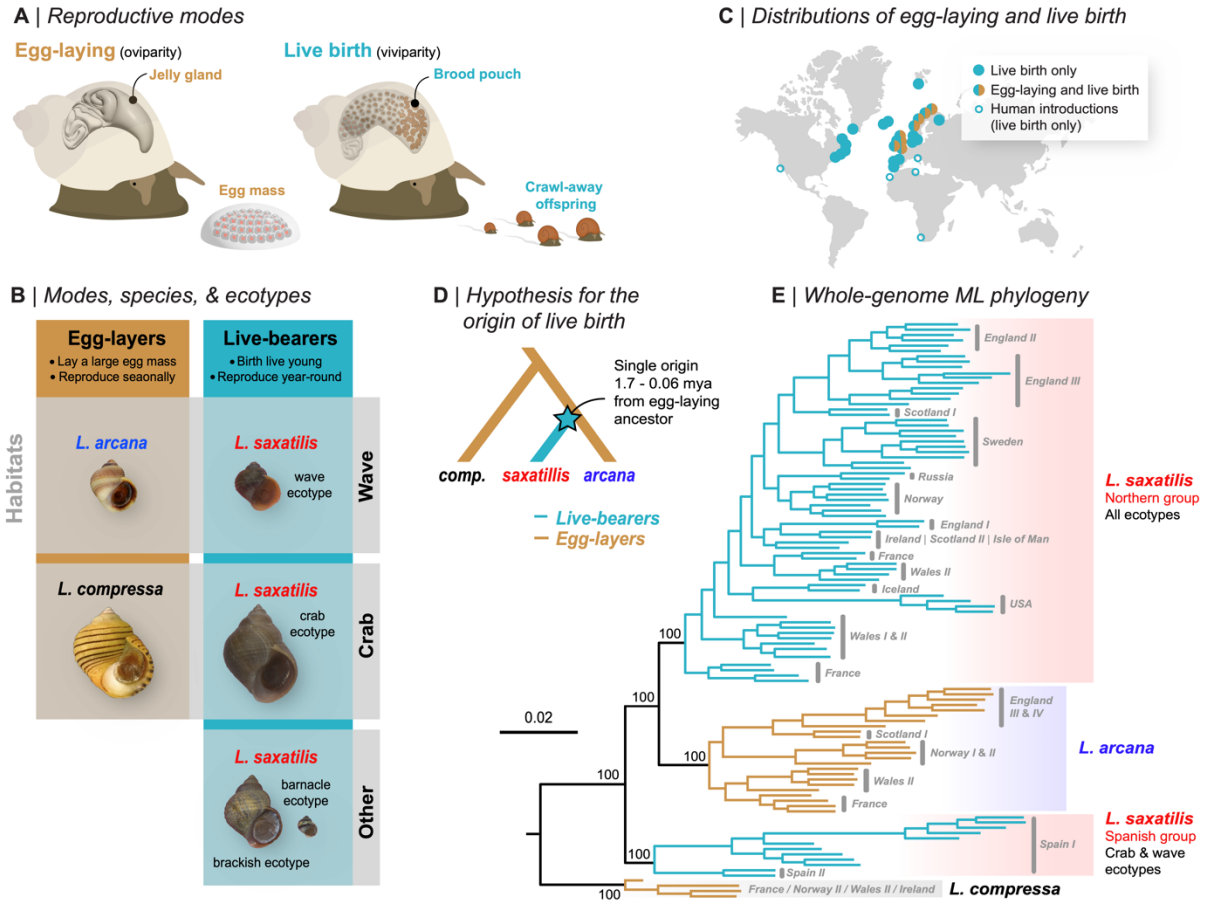

Figure 1. Variation in reproductive mode in *Littorina*. (A) Anatomical differences between modes (B) Egg-layers reproduce during a limited breeding season, while live bearers release offspring year-round. The two egg-layers share their habitats with ecotypes of the live-bearer, *L. saxatilis*. (C) Approximate distributions of the modes, highlighting the broader distribution of live birth. (D) Existing hypothesis for the origin of live birth. (E) Maximum-likelihood phylogenetic tree based on whole-genome sequences (108 individuals and 18.5 m variable sites). Bootstrap support for key nodes is shown.

We constructed a phylogeny from whole genome sequences to test the existing hypothesis of a single origin of live birth (1.7–0.06 Ma) (Fig. 1C) (*5*). Surprisingly, live-bearers formed two separate clades: one containing all *L. saxatilis* from Spain (hereafter 'Spanish *saxatilis*'), and another including all other *L. saxatilis* ('northern *saxatilis*') that was sister to egg-laying *L. arcana* (Fig. 1E). The discordance between the inferred relationships and reproductive mode (also seen in PCAs, fig. S6) has several possible explanations, including two genetically independent transitions between egg laying and live birth. However, given the close relationships of these species, a single origin could have been followed by sharing of causal alleles between lineages via gene exchange and selection (*11*). In this case, we would expect genealogies for loci

causing live birth to be strongly discordant from the genome-wide tree, with samples grouping by reproductive mode (*9*).

With this expectation in mind, we used topology weighting to identify genomic regions associated with reproductive mode. Specifically, we divided the genome into non-overlapping 100 SNP windows (mean size 5.8 kb, fig. S7), and calculated topology weights (*10*) for each window by iteratively sampling subtrees (Fig. 2A). Because we have four groups, each sampled subtree must fit one of three topologies (Fig. 2C, fig. S8): (i) the background topology, Tb, observed in our genome-wide analysis, (ii) the reproduction topology, Tr, where samples cluster by mode, and (iii) the control topology, Tc, which is of no specific interest except that it provides a control for distinguishing incomplete lineage sorting (ILS) from other processes that cause genealogical discordance. We took the novel approach of analyzing the joint distribution of topology weights in a ternary plot (Fig. 2A) and used simulations to understand how different factors shape the ternary distribution of weights (Fig. 2B; Supplementary text, fig. S8—S15; tables S3 & S4).

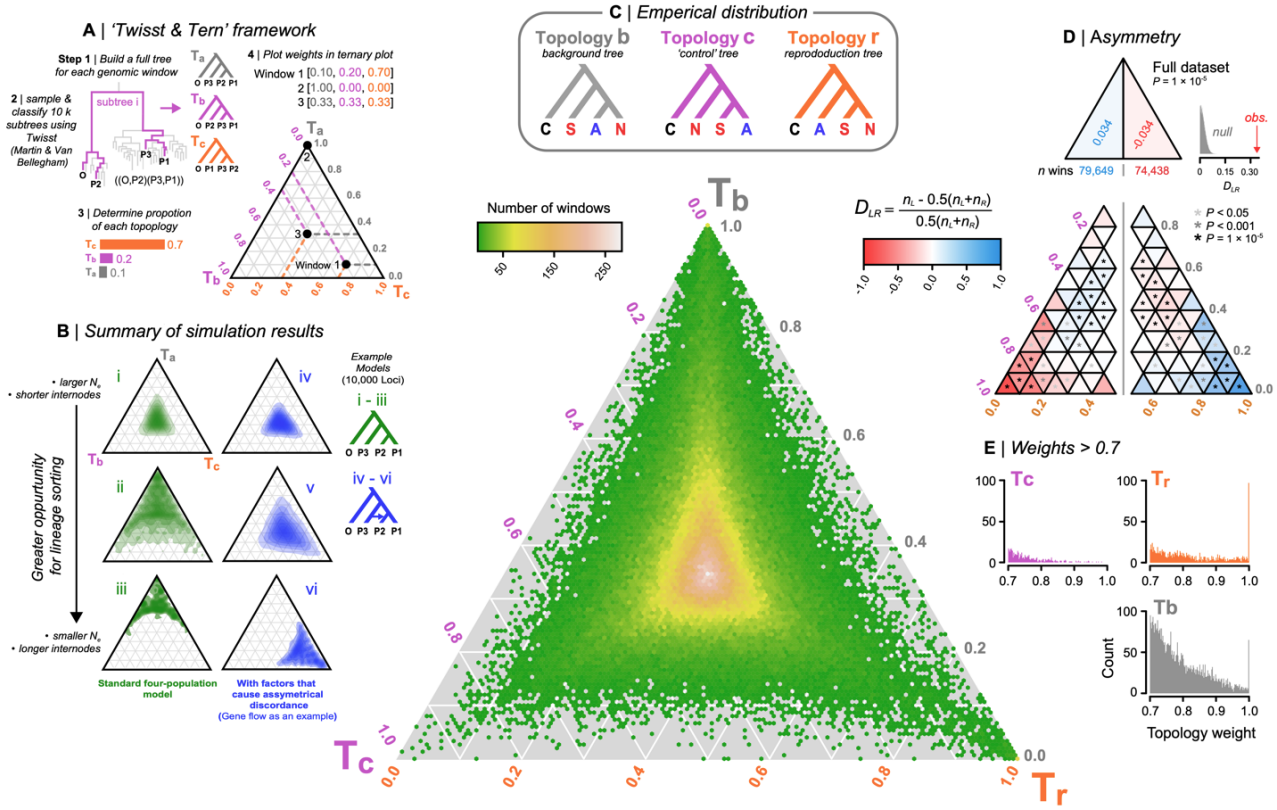

Figure 2. Topology weighting reveals genomic regions associated with reproductive mode. (A) For each window, we inferred a full tree including all haplotypes, and then sampled and classified 10k 'subtrees' by randomly picking one haplotype per group. Topology weights are the proportions of each topology among all subtrees. Windows were then plotted in a ternary plot based on their topology weights. (B) Simulated distributions of weights. A greater opportunity for lineage sorting (i - iii) biases the distribution toward the topology that matches the demographic history. Incomplete lineage sorting yields genealogies that are a better fit to one of the discordant trees, but the distribution is always symmetrical between the left and right half triangles. Additional factors, including gene flow, create a bias toward one of the discordant genealogies (panels iv - vi). (C) Possible topologies and the empirical distribution of weights for the 154,971 windows. Hexagonal bins are colored according to window count. (D) Counts of windows in the left and right half triangles, with the asymmetry quantified using D_{LR}. Further division into sub-triangles reveals left-right asymmetry throughout the distribution. Asterisks indicate significant asymmetry between corresponding left- and right-sided sub-triangles. (E) Distributions of weights > 0.7.

We expected the empirical distribution of weights to be biased toward Tb, because lineage sorting results in concordance between the demographic history and underlying gene trees (*11*) (Fig. 2B). However, the observed bias was only slight (Tb = 0.380, Tc = 0.310, Tr = 0.308), with just 62 of ~155,000 genomic regions perfectly fitting Tb (i.e., Tb = 1) (Fig. 2C). Instead, the bulk of the distribution fell close to the center of the triangle, revealing extensive ILS due to rapid diversification relative to the effective population size (*11*, *12*).

We found substantial left-right asymmetry in the distribution of weights (Fig. 2D). Such a bias is not expected to arise from ILS, because there is an equal chance that a given gene tree will more closely resemble either alternative topology (Fig. 2B) (*11*). We detected asymmetry using a new statistic, D_{LR} (Fig. 2D, fig. S16, table S5). A genome-wide test, performed by calculating D_{LR} between the two halves of the triangle, revealed a 3.4% excess of windows shifted toward the control topology (D_{LR} = 0.034, permutation test p = 1e-5). D_{LR} calculated between analogous left- and right-side sub-triangles, revealed that this asymmetry was driven by an excess of trees with a small bias toward Tc (table S5; fig. S17). Further exploration showed that this bias is due to several previously identified chromosomal inversions, where one arrangement is more common in Spanish *L. saxatilis* and *L. arcana*, and the other is more common in *L. compressa* and Northern *L. saxatilis* (D_{LR} for colinear regions = -0.007, p = 0.074) (Supplementary text; figs. S18—S20, table S6).

Much stronger asymmetry was observed between the extreme left and right sub-triangles, corresponding to windows that strongly fit one of the alternative topologies (Fig. 2D). However, the asymmetry was in the opposite direction to the genome-wide pattern, with a large excess of windows strongly biased toward the reproduction tree compared with the control tree (Tr > 0.7 = 1151 windows vs. 461 for Tc; D_{LR} = -0.43, p = 1e-5). A total of 88 windows perfectly fit the reproduction topology (i.e., Tr = 1), compared with 0 windows that perfectly fit the control topology (D_{LR} = 1.00, p = 1e-5; table S6).

Although neutral gene flow can generate strong asymmetry under some circumstances, we are unable to explain the observed Tr bias without invoking natural selection. We found strong additional evidence for live-bearer-specific positive selection in these regions (Fig. 3). First, window-based estimates of nucleotide diversity (π) in live-bearers decreased substantially with increasing Tr weight (Fig. 3A). We found no such relationship in egg-layers. Among perfectly associated regions, 95% (84 of 88) showed reduced π in live-bearers (mean $\pi_{\text{live-bearer}}$ = 0.0029 vs $\pi_{\text{Egg-layer}}$ = 0.0065; paired Wilcoxon test, p = 1.313e-15; Fig. 3B, fig. S22), consistent with selection having purged diversity from live-bearing haplotypes (*13*). Although this pattern could in principle result from a live-bearer-specific demographic bottleneck, we can rule this out because live-bearers and egg-layers have similar levels of genome-wide diversity (mean π live-bearer = 0.0065 vs. π Egg-layer = 0.0062; fig. S23). Further, relationships between π and the other weights (Ts and Tc) were weak, and similar for both groups, confirming that reduced π in live-bearers is specific to Tr rather than being a general feature of windows with extreme weights (fig. S24). The site-frequency spectra (SFS) and sample-size-corrected estimates of private alleles for perfectly associated regions provide further evidence for selection (Fig 3C & D; figs. S25—S28; table S9 & S10): the live-bearer SFS was strongly skewed toward rare variants (Tajima's D = -1.89, 95% CIs -1.77 – -2.01; figs. S25 & S26), the majority of which (80%) were private to the group. Both results are expected during the phase when diversity is recovered by mutation after a selective sweep (*14*). In contrast, the SFS for egg-layers was much closer to the neutral expectation (Tajima's D = -0.24, 95% CIs -0.037 – -0.437), with polymorphic sites being 2.14 times more abundant in egg-layers after accounting for the difference in sample size.

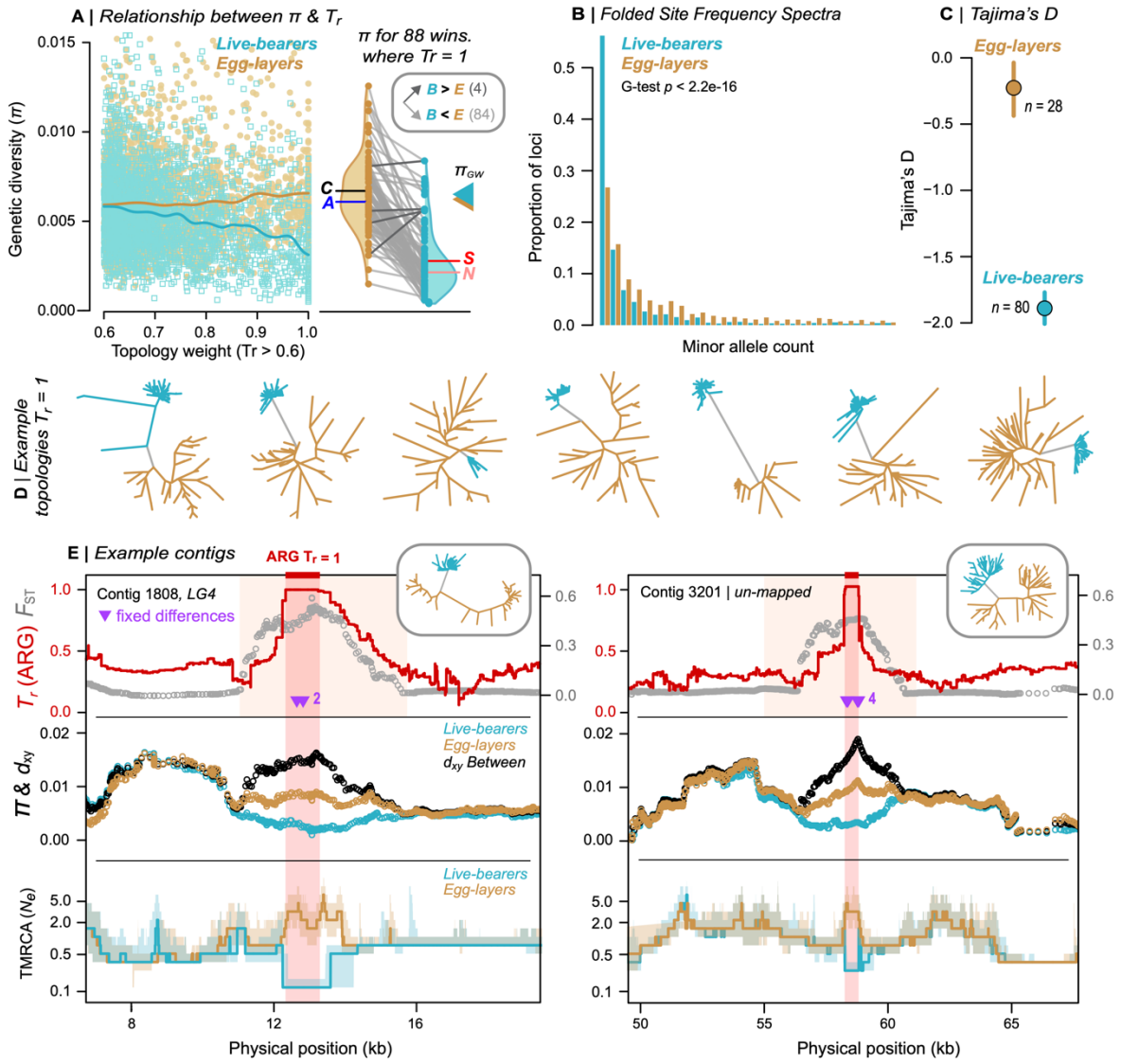

Figure 3. Evidence for positive selection on haplotypes associated with live birth. (A) Relationship between π and Tr for both reproductive modes. Triangles show genome-wide π. Violin plots show the distributions of π for windows where Tr = 1, with most showing lower diversity in live-bearers. Letters show mean values of π for egg-layers and live-bearers. (B) Folded SFS for each mode in perfectly associated regions, projected at the same sample size for comparison. (C) Estimates of Tajima's D with 95% CIs for perfectly associated regions. (D) Unrooted trees for example windows where Tr = 1. (E) Variation across two example contigs that contain a window where Tr = 1 (span of the orange box). The tree associated with each region is shown. Top panel: F_{ST} in 3kb sliding windows (30 bp step). TrARG shows the results of topology weighting applied to marginal trees obtained from inferred ancestral recombination graphs (ARGs). Purple arrows show fixed differences between modes. Middle panel: π and dxy in sliding windows. Bottom panel: traces of the time to the most recent common ancestor (TMRCA) obtained from ARG inference. Bold lines are the median estimates and envelopes the 95% CIs. The red box shows the inferred length of the core haplotype block associated with live birth.

We characterized footprints of selection within contigs to more accurately estimate the number and size of candidate regions (Fig. 3F). The 88 perfectly associated windows mapped to 50 contigs in our assembly (mean 1.7 ± sd 1.5 windows per contig; table S8). The regions were narrow, mostly spanning less than 20 kb (mean 12 kb ± sd 14.4 kb). Sliding-window analysis of each contig generally revealed clear peaks of allele frequency differentiation (F_{ST}) and sequence

divergence (d_{xy}) between the groups, as well as valleys of nucleotide diversity (π) in live-bearers (Fig. 3E; fig. S30). We also inferred ancestral recombination graphs (ARGs) for selected contigs to refine candidate regions (Fig. 3E). Unlike the trees for arbitrary windows, each marginal tree in an ARG corresponds to an inferred non-recombining segment of the genome (*15*). Thus, by applying topology weighting to the sequence of marginal trees, we were able to more precisely identify the segment of genome shared by all live bearing samples. In both cases, the core live-bearing haplotype spanned less than 2 kb. Live-bearers showed much shallower coalescence in these regions than egg-layers, as expected following a selective sweep (Fig. 3E).

The assignment of contigs to a genetic map revealed that associated windows are widespread across the genome, rather than co-localizing to one or a few genomic regions (Fig. 4a; table S11). As expected for a polygenic trait, the number of mode-associated windows on each LG was strongly predicted by LG size (Tr > 0.7, r = 0.79, p < 0.0001; Tr > 0.9, r = 0.71, p < 0.005). Associated windows were also widespread within linkage groups, in some cases with strong associations near opposite ends of the same LG (Fig. 4B).

Candidate regions also showed strong enrichment of genes that are differentially expressed between live-bearing and egg-laying reproductive tissues. To identify differentially expressed genes (DEGs), we collected reproductively mature samples of *L. arcana* and northern *L. saxatilis* from a single location to control for environmental effects, and compared transcriptomes from whole reproductive systems (brood pouch vs jelly gland) and a non-reproductive control tissue (foot tissues). We identified 1,598 DEGs, the majority of which showed differential expression between the reproductive tissues (1,297) (Fig. 4C). Of these, 66.1% (858) showed higher expression in the brood pouch of live bearers (Fig. 4D). To test for the enrichment of DEGs in regions associated with reproductive mode, we binned each DEG according to the Tr score of its associated genomic region (Fig. 4D). We found that the proportion of reproductive mode DEGs strongly increased with increasing Tr weight (Spearman's rho = 0.903, p = 9e-04).

Gene ontology analysis and functional annotation suggest that the transition to live-birth involved genes with diverse functions. Separate GO analyses conducted on a sequence-based gene set (574 genes in regions where Tr > 0.7) and expression-based gene set (1,450 reproductive mode DEGs) yielded 37 enriched gene ontology terms, including transmembrane transport, calcium ion binding, and ion channel activity (Fig. S35). We examined the putative functions of the 22 genes found in both sets in more detail (Table S13). These included genes putatively associated with antibacterial activity (LPS-like; higher expression in brood pouch), the synthesis of mucin-type oligosaccharides (GALNT10-like; higher expression in brood pouch), the formation of connective tissue (FBN3-like; lower expression in brood pouch) and vascular tissue (SEMA5A-like; lower expression in brood pouch), and two secretary genes that are involved in egg-mass production in another marine snail (both with lower expression in brood pouch).

Taken together, our results show that the adaptive origin of live birth in *Littorina* is underpinned by a complex polygenic architecture, as in the only comparable analysis in *Zootaca* lizards (*16*). All of our live-bearing individuals carry the same set of core haplotypes across many independent genomic regions. Thus, while our genome wide analysis hinted at two independent origins of live birth, live-bearing alleles at each locus clearly have a single, recent origin, and then spread across space and genetic background. Rather than alleles arising in many different locations followed by the buildup of range-wide LD, we hypothesize that live bearing initially arose in a single location. Levels of nucleotide diversity (Fig. 3A, Fig. S24), private

alleles (Fig. S30), and estimates of Tajima's D (Fig. S28), are consistent with the greater recovery of post-sweep variation in Spanish live-bearers, suggesting an origin near the southern extent of the current natural range where egg-layers are currently absent. Due to the complexity of the trait and number of associated loci, it is unlikely that live-birth arose in a single mutational step, as suggested by models of saltational evolution (*17*). Rather, live birth probably evolved gradually as a by-product of selection on related reproductive traits, such as embryo retention time (*4*). Live- bearing snails then eventually spread north, bringing them into contact with egg-layers. Gene flow, was then sufficient to obscure this history, while selection was sufficient to maintain sets of haplotypes necessary for contrasting reproductive modes. Regardless of the precise details, our results show that key innovations can have a polygenic basis, and that their historical origins can be obscured by a complex demographic history.

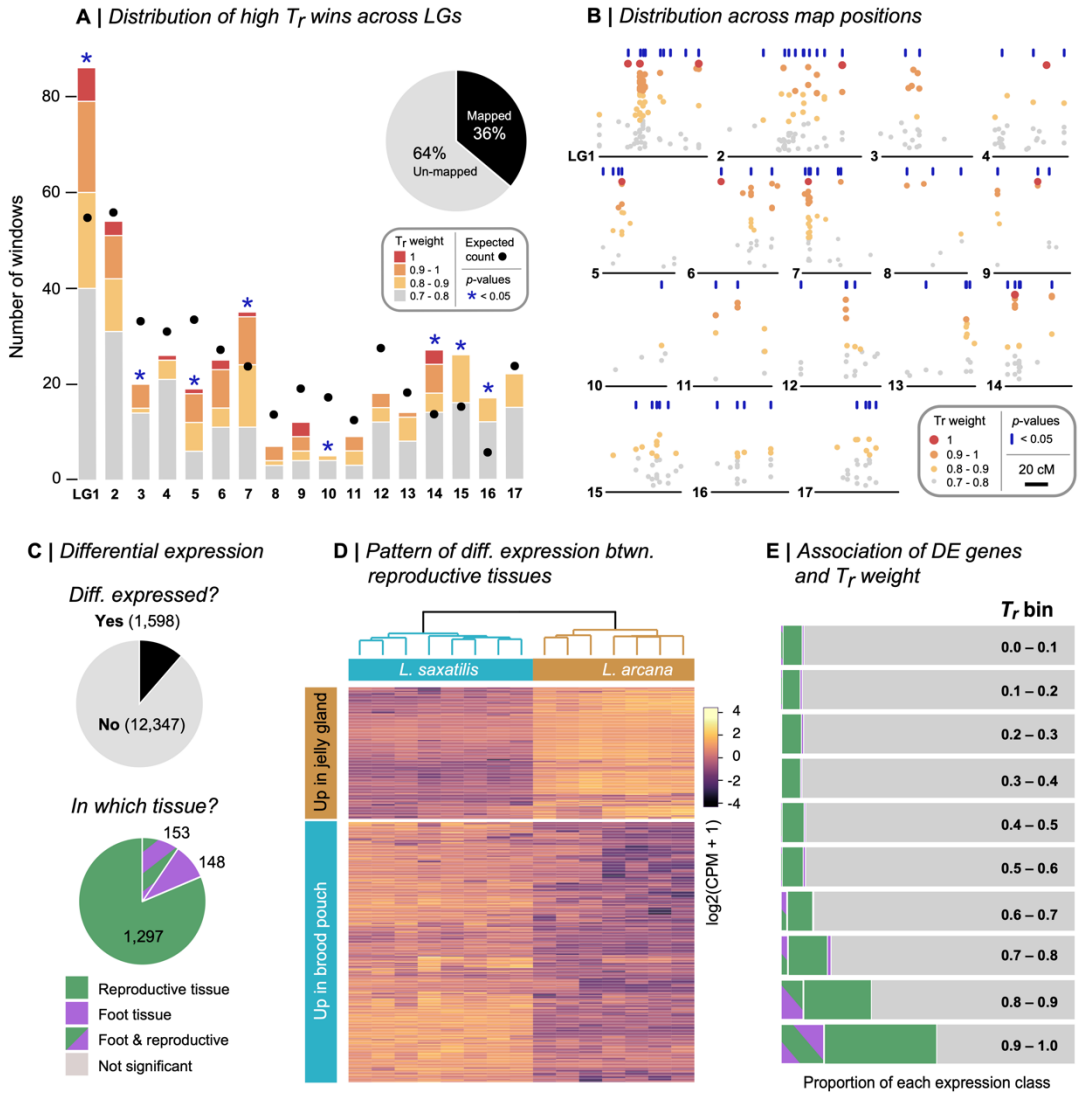

Figure 4. Candidate regions are widespread across the genome and enriched for genes that are differentially expressed between reproductive systems. (A) The number of high Tr windows (Tr > 0.7) assigned to each of the 17 *L. saxatilis* LGs. The circles show the expected number given the total assigned of windows to each LG. Asterisks indicate when the observed number is unlikely to be recovered by chance ($p < 0.05$). (B) Distribution of high Tr

windows across LGs. Vertical blue lines indicate map positions that are enriched for high Tr windows. (C) Number of genes that showed differential expression (DE) and the number of DE genes in each expression class. (D) Clustering of reproductive tissue libraries based on patterns of expression. (E) The proportion of genes in each DE class after binning each gene according to the Tr weight.

Acknowledgments: Juan Galindo, Mauricio Montaño-Rendón, Natalia Mikhailova, April Blakeslee, Einar Arnason and Petri Kemppainen provided samples. Richard Turney, Graciela Sotelo, Jenny Larsson and Stéphane Loisel helped with sample collection and processing. Mark Dunning helped develop bioinformatic pipelines. Hernan Morales, Vitor Sousa, and Nick Barton provided advice and discussion.

Funding:
National Environment Research Council grant NE/P001610/1 (RKB)
European Research Council grant ERC-2015-AdG693030-BARRIERS (RKB)

Author contributions:
Conceptualization: SS, AMW, KJ, RKB
Analysis: SS, ZBZ, MG, AP, DS, DGC, EL, JR, ALM
Visualization: SS, MG
Writing – original draft: SS, RKB
Writing – review & editing: SS, ZBZ, MG, AP, DS, DGC, TB, ALM, EL, JR, KJ, AMW, RKB

Competing interests:
Authors declare that they have no competing interests.

Data and materials availability:
Sequence data are available on the short-read archive (SRA). All other data and analysis scripts are available on Github at https://github.com/seanstankowski/Littorina_reproductive_mode.

Supplementary materials
Materials and Methods
Supplementary Text
Figs. S1 to S32
Tables S1 to S12

References and notes

1. A. Wagner, *The Origins of Evolutionary Innovations: A Theory of Transformative Change in Living Systems* (OUP Oxford, 2011).
2. A. H. Miller, J. T. Stroud, J. B. Losos, The ecology and evolution of key innovations. *Trends Ecol. Evol.* (2022), doi:10.1016/j.tree.2022.09.005.
3. A. Wagner, The molecular origins of evolutionary innovations. *Trends Genet.* **27**, 397–410 (2011).
4. C. M. Whittington, J. U. Van Dyke, S. Q. T. Liang, S. V. Edwards, R. Shine, M. B. Thompson, C. E. Grueber, Understanding the evolution of viviparity using intraspecific variation in reproductive mode and transitional forms of pregnancy. *Biol. Rev. Camb. Philos. Soc.* **97**, 1179–1192 (2022).

5. D. G. Reid, P. Dyal, S. T. Williams, A global molecular phylogeny of 147 periwinkle species (Gastropoda, Littorininae). *Zool. Scr.* **41**, 125–136 (2012).
6. D. G. Reid, Systematics and evolution of Littorina. The Ray Society, London, 463 p (1996).
7. S. Stankowski, A. M. Westram, Z. B. Zagrodzka, I. Eyres, T. Broquet, K. Johannesson, R. K. Butlin, The evolution of strong reproductive isolation between sympatric intertidal snails. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190545 (2020).
8. T. Warwick, A. J. Knight, R. D. Ward, Hybridisation in the Littorina saxatilis species complex (Prosobranchia : Mollusca). *Hydrobiologia*. **193**, 109–116 (1990).
9. M. W. Hahn, L. Nakhleh, Irrational exuberance for resolved species trees. *Evolution*. **70**, 7–17 (2016).
10. S. H. Martin, S. M. Van Belleghem, Exploring Evolutionary Relationships Across the Genome Using Topology Weighting. *Genetics*. **206**, 429–438 (2017).
11. W. P. Maddison, Gene Trees in Species Trees. *Syst. Biol.* **46**, 523–536 (1997).
12. R. R. Hudson, Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*. **7**, 44 (1990).
13. J. M. Maynard Smith, J. Haigh, The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
14. J. M. Braverman, R. R. Hudson, N. L. Kaplan, C. H. Langley, W. Stephan, The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*. **140**, 783–796 (1995).
15. D. Shipilina, A. Pal, S. Stankowski, Y. F. Chan, N. H. Barton, On the origin and structure of haplotype blocks. *Mol. Ecol.* (2022), doi:10.1111/mec.16793.
16. H. Recknagel, M. Carruthers, A. A. Yurchenko, M. Nokhbatolfoghahai, N. A. Kamenos, M. M. Bain, K. R. Elmer, The functional genetic architecture of egg-laying and live-bearing reproduction in common lizards. *Nat Ecol Evol*. **5**, 1546–1556 (2021).
17. G. Theißen, Saltational evolution: hopeful monsters are here to stay. *Theory Biosci.* **128**, 43–51 (2009).

The following refrences are cited in the supplementary materials

18. K. Johannesson, Z. Zagrodzka, R. Faria, A. Marie Westram, R. K. Butlin, Is embryo abortion a post-zygotic barrier to gene flow between Littorina ecotypes? *J. Evol. Biol.* **33**, 342–351 (2020).
19. R. K. Butlin, M. Saura, G. Charrier, B. Jackson, C. André, A. Caballero, J. A. Coyne, J. Galindo, J. W. Grahame, J. Hollander, P. Kemppainen, M. Martínez-Fernández, M. Panova, H. Quesada, K. Johannesson, E. Rolán-Alvarez, Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. *Evolution*. **68**, 935–949 (2014).
20. H. E. Morales, R. Faria, K. Johannesson, T. Larsson, M. Panova, A. M. Westram, R. K. Butlin, Genomic architecture of parallel ecological divergence: Beyond a single environmental contrast. *Sci Adv*. **5**, eaav9963 (2019).
21. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).
22. A. M. Westram, M. Rafajlović, P. Chaube, R. Faria, T. Larsson, M. Panova, M. Ravinet, A. Blomberg, B. Mehlig, K. Johannesson, R. Butlin, Clines on the seashore: The genomic architecture underlying rapid divergence in the face of gene flow. *Evol Lett*. **2**, 297–309 (2018).

23. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013), (available at http://arxiv.org/abs/1303.3997).
24. G. Tischler, *biobambam2: Tools for early stage alignment file processing* (Github; https://github.com/gt1/biobambam2).
25. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
26. P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, H. Li, Twelve years of SAMtools and BCFtools. *Gigascience*. **10** (2021), doi:10.1093/gigascience/giab008.
27. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **30**, 1312–1313 (2014).
28. E. M. Ortiz, *vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis* (2019; https://zenodo.org/record/2540861).
29. A. Rambaut, FigTree. Tree Figure Drawing Tool. *http://tree.bio.ed.ac.uk/software/figtree/* (2009) (available at https://ci.nii.ac.jp/naid/10029324212/).
30. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. **4**, 7 (2015).
31. B. L. Browning, S. R. Browning, Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*. **194**, 459–471 (2013).
32. S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
33. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, Ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
34. S. H. Martin, K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, F. Simpson, M. Blaxter, A. Manica, J. Mallet, C. D. Jiggins, Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.* **23**, 1817–1828 (2013).
35. F. Baumdicker, G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, G. Tsambos, S. Zhu, B. Eldon, E. C. Ellerman, J. G. Galloway, A. L. Gladstein, G. Gorjanc, B. Guo, B. Jeffery, W. W. Kretzschumar, K. Lohse, M. Matschiner, D. Nelson, N. S. Pope, C. D. Quinto-Cortés, M. F. Rodrigues, K. Saunack, T. Sellinger, K. Thornton, H. van Kemenade, A. W. Wohns, Y. Wong, S. Gravel, A. D. Kern, J. Koskela, P. L. Ralph, J. Kelleher, Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*. **220** (2022), doi:10.1093/genetics/iyab229.
36. J. Kelleher, K. Lohse, Coalescent Simulation with msprime. *Methods Mol. Biol.* **2090**, 191–230 (2020).
37. S. H. Martin, J. W. Davey, C. D. Jiggins, Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. *Mol. Biol. Evol.* **32**, 244–257 (2015).
38. R. E. Green, J. Krause, S. E. Ptak, A. W. Briggs, M. T. Ronan, J. F. Simons, L. Du, M. Egholm, J. M. Rothberg, M. Paunovic, S. Pääbo, Analysis of one million base pairs of Neanderthal DNA. *Nature*. **444**, 330–336 (2006).

39. N. E. Hamilton, M. Ferry, ggtern: Ternary diagrams using ggplot2. *J. Stat. Softw.* **87**, 1–17 (2018).
40. R. Faria, P. Chaube, H. E. Morales, T. Larsson, A. R. Lemmon, E. M. Lemmon, M. Rafajlović, M. Panova, M. Ravinet, K. Johannesson, A. M. Westram, R. K. Butlin, Multiple chromosomal rearrangements in a hybrid zone between Littorina saxatilis ecotypes. *Mol. Ecol.* **28**, 1375–1393 (2019).
41. A. M. Westram, R. Faria, K. Johannesson, R. Butlin, Using replicate hybrid zones to understand the genomic basis of adaptive divergence. *Mol. Ecol.* **30**, 3797–3814 (2021).
42. T. Jombart, adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*. **24**, 1403–1405 (2008).
43. P. Kemppainen, C. G. Knight, D. K. Sarma, T. Hlaing, A. Prakash, Y. N. Maung Maung, P. Somboon, J. Mahanta, C. Walton, Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. *Mol. Ecol. Resour.* **15**, 1031–1045 (2015).
44. F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. **123**, 585–595 (1989).
45. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics*. **27**, 2156–2158 (2011).
46. M. D. Rasmussen, M. J. Hubisz, I. Gronau, A. Siepel, Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* **10**, e1004342 (2014).
47. S. Andrews, Others, FastQC: a quality control tool for high throughput sequence data (2010).
48. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
49. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).
50. M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, S. L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
51. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. **26**, 139–140 (2010).
52. G. Pertea, M. Pertea, GFF Utilities: GffRead and GffCompare. *F1000Res.* **9** (2020), doi:10.12688/f1000research.23297.2.
53. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).
54. S. X. Ge, D. Jung, R. Yao, ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*. **36**, 2628–2629 (2020).