LETTER

doi:10.1038/nature13685

Comparative population genomics in animals uncovers the determinants of genetic diversity

J. Romiguier[1,2], P. Gayral[1,3], M. Ballenghien[1], A. Bernard[1], V. Cahais[1], A. Chenuil[4], Y. Chiari[5], R. Dernat[1], L. Duret[6], N. Faivre[1], E. Loire[1], J. M. Lourenco[1], B. Nabholz[1], C. Roux[1,2], G. Tsagkogeorga[1,7], A. A.-T. Weber[4], L. A. Weinert[1,8], K. Belkhir[1], N. Bierne[1], S. Glémin[1] & N. Galtier[1]

Genetic diversity is the amount of variation observed between DNA sequences from distinct individuals of a given species. This pivotal concept of population genetics has implications for species health, domestication, management and conservation. Levels of genetic diversity seem to vary greatly in natural populations and species, but the determinants of this variation, and particularly the relative influences of species biology and ecology versus population history, are still largely mysterious[1,2]. Here we show that the diversity of a species is predictable, and is determined in the first place by its ecological strategy. We investigated the genome-wide diversity of 76 non-model animal species by sequencing the transcriptome of two to ten individuals in each species. The distribution of genetic diversity between species revealed no detectable influence of geographic range or invasive status but was accurately predicted by key species traits related to parental investment: long-lived or low-fecundity species with brooding ability were genetically less diverse than short-lived or highly fecund ones. Our analysis demonstrates the influence of long-term life-history strategies on species response to short-term environmental perturbations, a result with immediate implications for conservation policies.

Since the early studies of evolutionary genetics, there has been no understanding of how and why genetic diversity levels vary between species. This old puzzle, considered four decades ago as 'the central problem in population genetics'[1], is still essentially unsolved in the genomic era[2]. Meanwhile, there is increasing evidence that genetic diversity is central to many conservation challenges, such as species response to environmental changes, ecosystem recovery, and the viability of recently endangered populations[3–7]. In this context, our ability to understand and predict this key aspect of biodiversity seems critical. But is it possible to quantify the contributory ecological and genetic factors? How predictable is the level of genetic diversity of a given species?

Population genetic theory states that neutral genetic polymorphism (that is, diversity) increases with effective population size, N_e, which in a panmictic population is equal to the number of individuals contributing to reproduction. One would therefore expect the genetic diversity of a species to be linked to biological traits associated with abundance, such as body size or fecundity. However, this intuitive prediction has not yet been clearly confirmed by empirical data[2,8–10]. This is typically explained by invoking the many confounding factors potentially affecting genetic polymorphism, such as mutation rate, population structure, population bottlenecks, selective sweeps, and, more generally, ecological disturbances[11,12]. Whether demographic or adaptive, historical contingency is often considered to be the main driver of genetic diversity[11]. According to this viewpoint, polymorphism levels would be expected to fluctuate in time more or less randomly, irrespective of life-history traits.

In the absence of compelling empirical evidence, the relative importance of species biology and ecology (on the one hand) and historical, contingent factors (on the other) in shaping the genetic diversity of species is still highly contentious. Indeed, current knowledge on species genetic diversity is based on just a handful of model organisms, or small sets of molecular data[2,8,13]. Various animal taxa and lifestyles, particularly across the invertebrates, have yet to be explored. Here we fill this gap and present the first distribution of genome-wide polymorphism levels across the metazoan tree of life.

We focused on 31 families of animals spread across eight major animal phyla. In each family we produced high-coverage transcriptomic data (RNAseq) for about ten individuals of a particular species. In 25 of these families, we sampled one to three additional species of similar biology and ecology (two to seven individuals each), thus producing taxonomic replicates. The total data set consisted of 374 individual transcriptomes from 76 non-model species (Fig. 1, Extended Data Fig. 1 and Supplementary Tables 1 and 2), from which we predicted protein coding sequences[14] and identified diploid genotypes and single nucleotide polymorphisms[15,16] (Methods). Across species the number of analysed loci varied from 804 to 20,222 (median 5,347) and the number of polymorphic sites from 1,759 to ~230,000 (median 17,924).

Estimates of the synonymous nucleotide diversity (π_s) spanned two orders of magnitude across species, a range far wider than is usually observed in surveys restricted to a single taxonomic group. The extreme values of π_s were observed in two invertebrate species: 0.1% in the subterranean termite *Reticulitermes grassei*; 8.3% in the slipper shell *Bostrycapulus aculeatus*. Figure 1 illustrates the patchy distribution of low-diversity (green) and high-diversity (red) species across the metazoan phylogeny. It also shows that species π_s values tend to be similar within families, but distinct between families (analysis of variance; $P < 10^{-12}$). Such a strong taxonomic effect would be unexpected if stochastic disturbances and contingent effects were the main drivers of genetic diversity, because species from a given family are not particularly expected to share a common demographic history. Testing this hypothesis more thoroughly, we detected no strong relationship between π_s and any variable related to geography, such as the average distance between GPS records (regression test, $P = 0.19$, $r^2 = 0.02$), maximum distance between GPS records ($P = 0.02$, $r^2 = 0.07$), average distance to Equator ($P = 0.87$, $r^2 = 0.0003$), population structure (measured as F_{it}, $P = 0.22$, $r^2 = 0.02$), invasive status (Student's t-test, $P = 0.14$) and marine versus continental environment (Student's t-test, $P = 0.52$).

To test whether species biology can explain variations in π_s, we collected data for six life-history traits potentially related to N_e: adult size, body mass, maximum longevity, adult dispersion ability, fecundity and propagule size (Supplementary Table 3). In contrast to the geographic variables, all these traits were significantly correlated with the nucleotide diversity (Extended Data Fig. 2) and collectively explained 73% of the variance in π_s in a multiple linear regression test ($P < 10^{-10}$). Propagule size, here defined as the size of the stage that leaves its parents and disperses (egg or juvenile depending on species), is by far the most predictive

[1]UMR 5554, Institute of Evolutionary Sciences, University Montpellier 2, Centre national de la recherche scientifique, Place E. Bataillon, 34095 Montpellier, France. [2]Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland. [3]UMR 7261, Institut de Recherches sur la Biologie de l'Insecte, Centre national de la recherche scientifique, Université François-Rabelais, 37200 Tours, France. [4]Aix-Marseille Université, Institut Méditerranéen de Biodiversité et d'Écologie marine et continentale (IMBE) – CNRS – IRD – UAPV, 13007 Marseille, France. [5]Department of Biology, University of South Alabama, Mobile, Alabama 36688-0002, USA. [6]UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, CNRS, 69622 Lyon, France. [7]The School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK. [8]Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge CB3 0ES, UK.

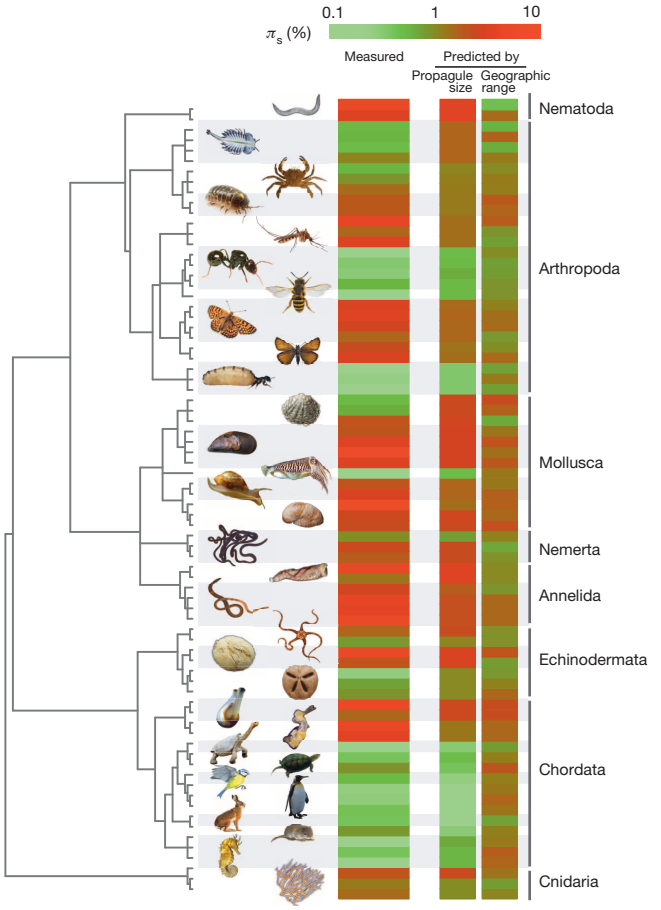


Figure 1 | Genome-wide genetic diversity across the metazoan tree of life. Each branch of the tree represents a species ($n = 76$). The leftmost vertical coloured bar is the estimated genome-wide genetic diversity (π_s), the central bar is the prediction of π_s based on a linear model with propagule size as the explanatory variable ($P < 10^{-14}$, $r^2 = 0.56$), and the rightmost bar is the prediction of π_s based on a linear model with average distance between GPS records, maximal distance between GPS records, average distance to Equator and invasive status as explanatory variables ($P = 0.16$). Each thumbnail corresponds to one metazoan family. Species are in the same order as in Supplementary Table 2 (from top to bottom).

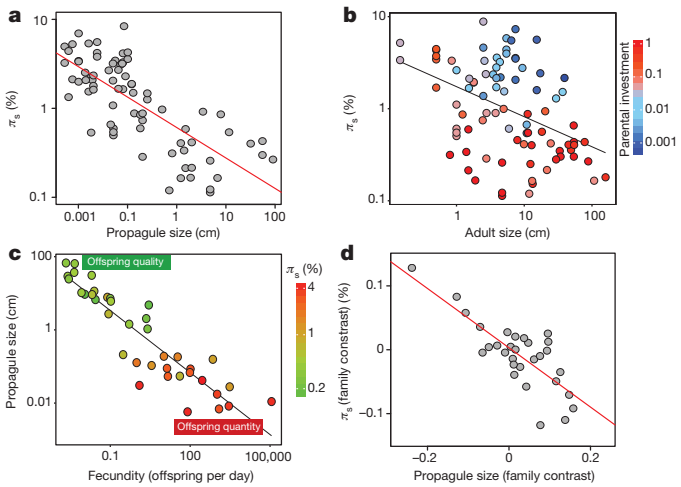


Figure 2 | Life-history traits and genetic diversity relationships. **a**, Relationship between propagule size and π_s ($P < 10^{-14}$, $r^2 = 0.56$, 76 species included; see Fig. 1). **b**, Relationship between adult size and π_s ($P < 0.05$, $r^2 = 0.07$, 76 species included). The colour scale represents the degree of parental investment, here defined as the ratio of propagule size to adult size. **c**, Effect of fecundity per day (x axis) and propagule size (y axis) on genetic diversity (colour scale; $P < 10^{-6}$, $r^2 = 0.69$, 29 family-averaged data points). **d**, Phylogenetic contrasts of family-averaged π_s versus family-averaged propagule size ($P < 10^{-6}$, $r^2 = 0.62$).

of these variables (linear regression test, $r^2 = 0.56$; Fig. 2a). This is illustrated in Fig. 1 by the good agreement between the observed distribution of π_s (leftmost coloured vertical bar) and the π_s value predicted from propagule size (central bar). The predicted π_s based on four demographic metrics is plotted alongside (rightmost bar) for visual comparison.

We explored in more detail the relative impact on π_s of the various life-history traits of interest here (Extended Data Fig. 2). Figure 2b plots the relationship between π_s and species adult size, a variable typically taken as a proxy for population size in some taxa[9]. Although significant, the correlation is not particularly strong ($P = 0.018$, $r^2 = 0.07$). In particular, species with low genetic diversity cover a large range of body sizes, from less than 1 cm to more than 1 m. Low-polymorphism species include amniotes (turtles, mammals and birds), but also brooding marine species (seahorses, brooding urchins, nemerteans and brittle-stars), eusocial insects (ants, bees and termites) and cuttlefish. These phylogenetically unrelated species have in common a large parental investment in their offspring, as represented in Fig. 2b by the ratio of propagule size to adult size (red). In contrast, species with minimal parental investment (blue) tend to carry high genetic diversity given their size. This is typically the case of highly fecund, broadcast spawning sessile species (such as mussels, non-brooding urchins, nemerteans and brittle-stars, sea squirts and gorgonians). The trade-off between offspring quantity (fecundity) and quality (propagule size) seems to be the most relevant factor explaining variations in polymorphism between species in the animal kingdom (Fig. 2c). We shall for simplicity hereafter categorize as *K*-strategists the species that tend to invest in the quality of their progeny, and as *r*-strategists those that favour quantity[17].

The correlation we report between life-history traits and π_s is not due to phylogenetic non-independence of the sampled species: taking family averages from Fig. 1 increased the correlation coefficients (from $r^2 = 0.56$ to $r^2 = 0.66$ with propagule size alone, from $r^2 = 0.73$ to $r^2 = 0.79$ with the six life-history traits). When we took into account the between-family phylogenetic tree using independent contrasts, this still resulted in highly significant correlations between π_s and life-history traits ($r^2 = 0.62$ for propagule size; Fig. 2d and Extended Data Fig. 3). These relationships were also unaffected by sampling strategy, sequencing depth, gene expression levels or contaminants (Methods, Supplementary Table 4 and Extended Data Figs 4–6). Finally, our conclusions were unchanged when we included 14 previously published species of mammals[10] or when we restricted the analysis to a subset of common orthologous genes (Supplementary Table 4).

The relationship between π_s and life-history traits, however strong, could in principle be mediated by causative variables that were not included in the analysis. One of these potential confounding factors is the mutation rate: a higher average per-generation mutation rate in *r*-strategists than in *K*-strategists could explain our results irrespective of the population size effect. However, theoretical models and empirical measurements actually support the opposite; that is, an increased per-generation mutation rate in large, long-lived organisms due to a larger number of germline cell divisions per generation and a reduced efficacy of natural selection on the fidelity of polymerases[18]. Therefore, as far as we can tell, across-species variations of mutation rate are likely to oppose, not strengthen, the main effect we are reporting here.

We computed the non-synonymous nucleotide diversity, π_n, and this was also found to be correlated with species life-history traits (Extended Data Fig. 2). We found substantial variation in π_n/π_s across metazoan species, and significant correlations with life-history traits, the best predictor in this case being longevity (Extended Data Fig. 7). This positive correlation is predicted by the nearly neutral theory of molecular evolution[19]: in small populations (long-lived species), the enhanced genetic drift counteracts purifying selection and promotes the segregation of weakly

deleterious, non-synonymous mutations at high allele frequency. These results also confirm that the relationships we uncovered between life-history traits and diversity patterns are mediated in the first place by an effect of N_e, not of the mutation rate; synonymous and non-synonymous positions being physically interspersed, the π_n/π_s ratio is unaffected by the mutation rate.

Our analysis reveals that polymorphism levels are well predicted by species biology, whereas historical and contingent factors are only minor determinants of the genetic diversity of a species. This unexpected result opens new questions. How can life-history traits be so predictive of π_s in spite of the overwhelming evidence for the impact of ecological perturbations on patterns of genetic variation[11,12]? Why does the 'r/K gradient' affect genetic polymorphism so strongly?

In an attempt to resolve these paradoxes, we suggest that life-history strategies might influence the response of species to environmental perturbations. Because K-strategy species have been selected for survival and the optimization of offspring quality in complex, stable environments[17], we speculate that they might experience fewer occasional disturbances (or be less sensitive to them), thus ensuring the long-term viability of even small populations. In contrast, only species with a large population-carrying capacity could sustain the 'riskier' r-strategy in the long term, thus buffering the frequent bottlenecks experienced in the context of high environmental sensitivity (see Supplementary Equations for a model formalizing these arguments). According to this hypothesis, environmental perturbations would be a common factor affecting every species, but their demographic impact would depend on the life-history strategy of each species.

This study highlights the importance of species life-history strategy when it comes to turning genetic diversity measures into conservation policy. So far, conservation efforts have mainly been focused on large-sized vertebrates. Here we show that these popular animals represent only a subset of the existing low-diversity, K-strategists. Invertebrate species with strong parental investment are probably equally vulnerable to genetic risks. Our results also indicate that r-strategists will typically show elevated amounts of genetic diversity irrespective of their current demography, which suggests that species of this kind might face significant extinction risks[20] without any warning genetic signal.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 March; accepted 17 July 2014.

Published online 20 August; corrected online 12 November 2014 (see full-text HTML version for details).

1. Lewontin, R. *The Genetic Basis of Evolutionary Change* (Columbia Univ. Press, 1974).
2. Leffler, E. M. *et al.* Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* **10,** e1001388 (2012).
3. Keller, L. F. & Waller, D. M. Inbreeding effects in wild populations. *Trends Ecol. Evol.* **17,** 19–23 (2002).
4. Reusch, T. B. H., Ehlers, A., Hämmerli, A. & Worm, B. Ecosystem recovery after climatic extremes enhanced by genotypic diversity. *Proc. Natl Acad. Sci. USA* **102,** 2826–2831 (2005).
5. Hughes, A. R., Inouye, B. D., Johnson, M. T. J., Underwood, N. & Vellend, M. Ecological consequences of genetic diversity. *Ecol. Lett.* **11,** 609–623 (2008).
6. Johnson, W. E. *et al.* Genetic restoration of the Florida panther. *Science* **329,** 1641–1645 (2010).
7. Nair, P. Conservation genomics. *Proc. Natl Acad. Sci. USA* **111,** 569 (2014).
8. Bazin, E., Glémin, S. & Galtier, N. Population size does not influence mitochondrial genetic diversity in animals. *Science* **312,** 570–572 (2006).
9. Nabholz, B., Mauffrey, J.-F., Bazin, E., Galtier, N. & Glemin, S. Determination of mitochondrial genetic diversity in mammals. *Genetics* **178,** 351–361 (2008).
10. Perry, G. *et al.* Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res.* **22,** 602–610 (2012).
11. Banks, S. C. *et al.* How does ecological disturbance influence genetic diversity? *Trends Ecol. Evol.* **28,** 670–679 (2013).
12. Hewitt, G. The genetic legacy of the Quaternary ice ages. *Nature* **405,** 907–913 (2000).
13. Ellegren, H. Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* **29,** 51–63 (2014).
14. Cahais, V. *et al.* Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Mol. Ecol. Resources* **12,** 834–845 (2012).
15. Tsagkogeorga, G., Cahais, V. & Galtier, N. The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate *Ciona intestinalis*. *Genome Biol. Evol.* **4,** 740–749 (2012).
16. Gayral, P. *et al.* Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. *PLoS Genet.* **9,** e1003457 (2013).
17. MacArthur, R. H. & Wilson, E. O. *The Theory of Island Biogeography* (Princeton Univ. Press, 1967).
18. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26,** 345–352 (2010).
19. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246,** 96–98 (1973).
20. Pinsky, M. L. *et al.* Unexpected patterns of fisheries collapse in the world's oceans. *Proc. Natl Acad. Sci. USA* **108,** 8317–8322 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the following for providing samples: F. Delsuc, E. Douzery, M. Tilak, G. Dugas, S. Harispe, C. Benoist, D. Bouchon (woodlice), J. Bierne, M. Bierne, B. Houseaux, M. Strand, C. Lemaire, D. Lallias, Service Modèle Biologique Station Marine Roscoff (nemertines), X. Turon, S. Lopez-Legentil (Cystodytes), P. Jarne, P. David, R. Dillon, J. Auld, R. Relyea, C. Lively, J. Jokela, V. Poullain, T. Stewart (snails), S. Lapègue, V. Boulo, F. Batista, D. Lallias, L. Fast Jensen, M. Cantou (oysters), J. Do Nascimento, C. Daguin-Thiébaut, M. Cantou (crabs), L. Bonnaud (cuttlefish), D. Aurelle (gorgonians), F. Viard, Y. Pechenik, A. Cahill, R. Colins (slipper limpets), L. Dupont (earthworms), D. Jollivet (trumpet worms), M. A. Felix, I. Nuez (nematodes), N. Rodes, T. Lenormand, E. Flaven (brine shrimps), Rotterdam Zoo, Zurich Zoo, C. Libert, Montpellier Zoo, S. Martin, la Ferme aux Crocodiles, O. Verneau, C. Ayres, M. Carretero, M. Vanberger, K. Pobolsaj, M. Zuffi, C. Palacios, L. du Preez, B. Halpern, Budapest Zoo (turtles), P. Peret, C. Doutrelant, B. Halpern, B. Rosivall (tits), M. de Dinechin, B. Rey (penguins), Z. Melo-Fereira, P. Alves (hares), N. Brand, M. Chapuisat (bees), R. Blatrix, A. Lenoir, I. Nodet, A. Lugagne, S. Blanquart, L. Serres-Giardi, V. Roustang, N. François, G. Ballantyne, A. Carbonnel, Y. Samuel, G. James, G. Kalytta, F. Guerrini, S. Stenzel, J. Beekman, X. Cerda, S. Ikoen (ants), I. Hanski, S. Ikonen, J. Kullberg, Z. Kolev (fritillary butterflies), F. Viard, X. Turon, Di Jiang, D. Chourrout, B. Vercaemer, E. Newman-Smith, Ascidian Stock Center, Service Modèle Biologique Station Marine Roscoff (ciona), L. Excoffier, G. Heckel (voles), F. Dedeine (termites), C. Atyame, O. Duron, M. Weill (mosquitoes), M. Cantou, H. Violette, F. Batista, J. Hondeville (seahorses), C. Fraïsse, G. Pogson, N. Saarman, J. Normand (mussels), E. Poulin, C. Gonzalez-Weivar, and J. P. Feral (sea urchins). This work was supported by European Research Council advanced grant 232971 (PopPhyl).

Author Contributions N.G. conceived the project. P.G., M.B., N.F., Y.C., L.A.W., G.T., A.C., A.W., J.R., N.G. and N.B. performed sampling and laboratory work. A.B., V.C., E.L., J.R., J.M.L., C.R., P.G., G.T., B.N., R.D., K.B., S.G. and N.G. developed the data analysis pipeline. J.R. collected life-history/geographic variables and produced figures. J.R., A.B., V.C., L.D., E.L. and N.G. analysed the data. S.G., N.B., B.N., J.R. and N.G. provided interpretations and models. J.R., N.B., S.G. and N.G. wrote the paper.

Author Information Data sets are freely available from the Sequence Read Archive (SRA) database (http://www.ncbi.nlm.nih.gov/sra) under project ID SRP042651 and from the Datasets section of the PopPhyl website (http://kimura.univ-montp2.fr/PopPhyl), in which predicted single nucleotide polymorphisms and genotypes are provided as .vcf files. Scripts and executable files are freely available from the Tools section of the PopPhyl website. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.G. (nicolas.galtier@univ-montp2.fr).

METHODS

Sampling and sequencing. The 76 analysed species were selected based on phylogenetic and ecological criteria with the goal of gathering a panel representative of the metazoan diversity. In each species, from two to eleven individuals were collected in various localities of their natural geographic range (Supplementary Table 1) or from zoos (*Chelonoidis nigra*). Depending on the species, the whole body, body parts, organs or tissues were dissected and preserved in RNA later, flash-frozen, or processed immediately (Supplementary Table 2). For each sample, total RNA was extracted using standard and improved protocols[21], and a non-normalized complementary DNA library was prepared. The libraries were sequenced on a Genome Analyzer II or Hiseq 2000 (Illumina) to produce 100-base-pair (bp) or 50-bp single-end fragments. In 12 species, an additional normalized random-primed cDNA library was prepared and sequenced for half a run using a 454 Genome Sequencer (GS) FLX Titanium Instrument (Roche Diagnostics). Illumina reads from two to four individuals in 14 mammalian species[10] were downloaded from the SRA database. Reads were trimmed of low-quality terminal portions with the SeqClean program (http://compbio.dfci.harvard.edu/tgi/). The fastQ program was applied to Illumina reads and revealed only a limited amount of motif enrichment: the number of motifs in significant excess varied between 0 and 17 across species, its median being 1.

Transcriptome assembly, read mapping, coding sequence prediction. *De novo* transcriptome assembly based on the 454 (when available) and Illumina reads was performed by following strategies B and D in ref. 14, using a combination of the programs Abyss and Cap3. Illumina reads were mapped to predicted cDNAs (contigs) with the BWA program. Contigs with a per-individual average coverage below $\times 2.5$ were discarded. Open reading frames (ORFs) were predicted with the Trinity package. Contigs carrying no ORF longer than 200 bp were discarded. In contigs including ORFs longer than 200 bp, 5′ and 3′ flanking non-coding sequences were deleted, thus producing predicted coding sequences that are hereafter referred to as loci.

Calling single nucleotide polymorphisms (SNPs) and genotypes. At each position of each locus and for each individual, diploid genotypes were called according to the method described in ref. 15 (model M1) and improved in ref. 16, using the reads2snps program. This method first estimates the sequencing error rate in the maximum-likelihood framework, calculates the posterior probability of each possible genotype, and retains genotypes supported at >95%; otherwise missing data are called. A minimum of ten reads per position and per individual were required to call a genotype. Then polymorphic positions were filtered for possible hidden paralogues (duplicated genes), using a likelihood ratio test based on explicit modelling of paralogy[16]. The across-species average percentage of SNPs discarded for suspicion of paralogy was 7.65%. Positions at which a genotype could not be called in a sufficient number of individuals were discarded. Calling k the number of sampled individuals for a given species ($2 \leq k \leq 11$), the minimum number of genotyped individuals required to retain a position was set to $k/2$ when $k > 5$, to $k - 1$ when $k = 4$ or $k = 5$, and to k when $k < 4$.

Control for contamination. Each locus of each species was translated to protein and compared with the non-redundant NR database using BlastP in search for possible contaminants. The percentage of loci for which no significant hit (e-value <0.001) was retrieved varied greatly between species, reflecting the taxonomic representation of sequences in NR. The percentage of no-hits was below 10% in mammals, but reached values above 20% in echinoderms and cnidarians. When one or several hits were found, the GenBank taxonomy of the first hit was recorded. Overall, 98.7% of first hits were assigned to Metazoa. The percentage of non-metazoan first hits was below 2% in 63 species out of 76, reaching its maximum (5.5%) in the trumpet worm *Pectinaria koreni*. Contamination by known microbes therefore seems negligible in our data set. Extended Data Figure 6 displays the taxonomic distribution of hits for four representative species—a mammal, an insect, a mollusc and an annelid. The results of our analyses were qualitatively unchanged when we removed loci that hit a non-metazoan and/or no-hit loci. In all cases, linear regression tests between π_s and propagule size yielded the same r^2 of 0.55 ($P < 10^{-13}$) as in our main analysis.

Life-history, ecological and geographical variables. Species life-history traits (adult size, juvenile size, body mass, longevity, fecundity and adult speed) were retrieved from the literature (Supplementary Table 3). Invasive/non-invasive status was obtained from the Global Invasive Status Database (http://www.issg.org/database/species/List.asp). The Global Biodiversity Information Facility database (http://www.gbif.org/) was queried to retrieve the GPS records corresponding to documented observations of individuals from the species of interest here. These data were merged with the GPS records of our own samples. For each species, the average and maximum distance between two distinct GPS records and the average distance to the Equator were computed (Supplementary Table 2).

Statistical analyses. Population genomic statistics π_s, π_n and F_{it} were calculated by using home-made programs that rely on the Bio++ libraries[22]. The F_{it} calculation was corrected for small sample sizes in accordance with ref. 23. Confidence intervals were obtained by bootstrapping loci. Regression analyses were conducted in R. Variables were log-transformed before linear regressions were performed. The linear model including π_s and propagule size alone met the required assumptions of normally distributed residual errors (Shapiro's test, $P = 0.19$) and homoscedasticity (Fligner–Killeen's test, $P = 0.48$). The same remark is valid for the multiple linear model including π_s and the six life-history traits (Shapiro's test, $P = 0.31$; Fligner–Killeen's test, $P = 0.49$). Family-level phylogenetic independent contrast analysis was performed with the APE package based on the tree shown in Extended Data Fig. 3, in which branch lengths are proportional to time. Divergence time estimates were retrieved from the TimeTree database (expert result, or average value if expert result was missing). When divergence time estimates were not available (Polycitoridae–Cionidae, Hesperiidae–Nymphalidae, Calyptraeoidae–Physidae, Mytilidae–Ostreidae), they were inferred on the basis of the divergence dates of neighbouring nodes.

SNP calling quality controls. The main analyses of this study were reproduced in three ways: first, with an increased minimum number of reads per position per individual of 30 instead of 10, second, removing five bases from each end of each read, and third, not using 454 data, thus controlling for a potential effect of insufficient sequencing depth, low-quality base calls near read ends and sequencing technology. In all three cases the results were highly similar to the main analysis (Supplementary Table 4; columns 'depth = 30X', 'clip_ends' and 'no_454', respectively), indicating that the analysis was robust to these technical caveats. No difference in π_s was detected between species showing versus not showing a significant excess of certain motifs by fastQC.

GC content. In each species, the correlation coefficient between contig GC content and contig π_s was calculated. It was significantly positive in 37 species, significantly negative in 18 species, and non-significantly different from zero in 21 species. The squared correlation coefficients (r^2) were relatively low (median r^2 0.007; maximum r^2 0.16 in *Physa acuta*), suggesting only a weak effect of GC content on π_s. For each species, the average contig GC content was calculated and correlated to the average π_s or propagule size. No significant relationship was detected, which indicates that the variation in GC content across genes and across species has no substantial impact on the results of this study.

Individual and locus sampling. No significant relationship was found between π_s and the number of sampled individuals per species, or between π_s and the number of sampled loci per species (Extended Data Fig. 4). The robustness to sampling of the relationship between propagule size and π_s was further assessed in two ways. First, for each species, loci were randomly subsampled. Extended Data Figure 5 displays the squared coefficient of correlation between propagule size and π_s as a function of the per species number of analysed loci. It shows that as few as 50 loci are enough to capture the relationship with a good probability. Second, for each species, only two individuals were randomly selected and the analyses were conducted again. Results were highly similar to the main analysis: the relationship between propagule size and π_s was unchanged and highly significant ($P < 10^{-15}$, $r^2 = 0.55$), thus indicating that population sample size is not an issue.

Orthologous genes. The coding sequences of 129 genes or gene fragments previously identified as orthologous across metazoans[24] (hereafter called 'core genes') were downloaded. In each of our species, contigs predicted to be orthologous to one of the core genes by reciprocal best BLAST hit were selected (expected e-value 0.0001, hits with a number of identical matches less than 80 and a bitscore of less than 1,200 were discarded). The number of such predicted core gene orthologues varied between 40 and 122 among species. We restricted the data set to the 39 species including at least 21 core gene orthologues, and reproduced the analysis. Colon 'orthologues' from Supplementary Table 4 shows that the correlation between π_s and life-history traits was still strong and significant when a subset of common genes was considered.

Expression level. In each species, the expression level of each locus was estimated as the average number of bases read per position. Correlating π_n/π_s to expression level across genes revealed no significant relationship in 33 species, and weak relationships ($r^2 < 0.27$) in 57 species. The relationship between π_n/π_s and expression level, when detected, was negative, as expected under the hypothesis or a stronger selective pressure acting on high-expressed genes. Then, for each species, loci were grouped into three equal-sized bins of genes with high, medium or low expression. Each of these categories, taken separately, provided a strong correlation between species propagule size and π_s ($r^2 = 0.57$, 0.62 and 0.62, respectively).

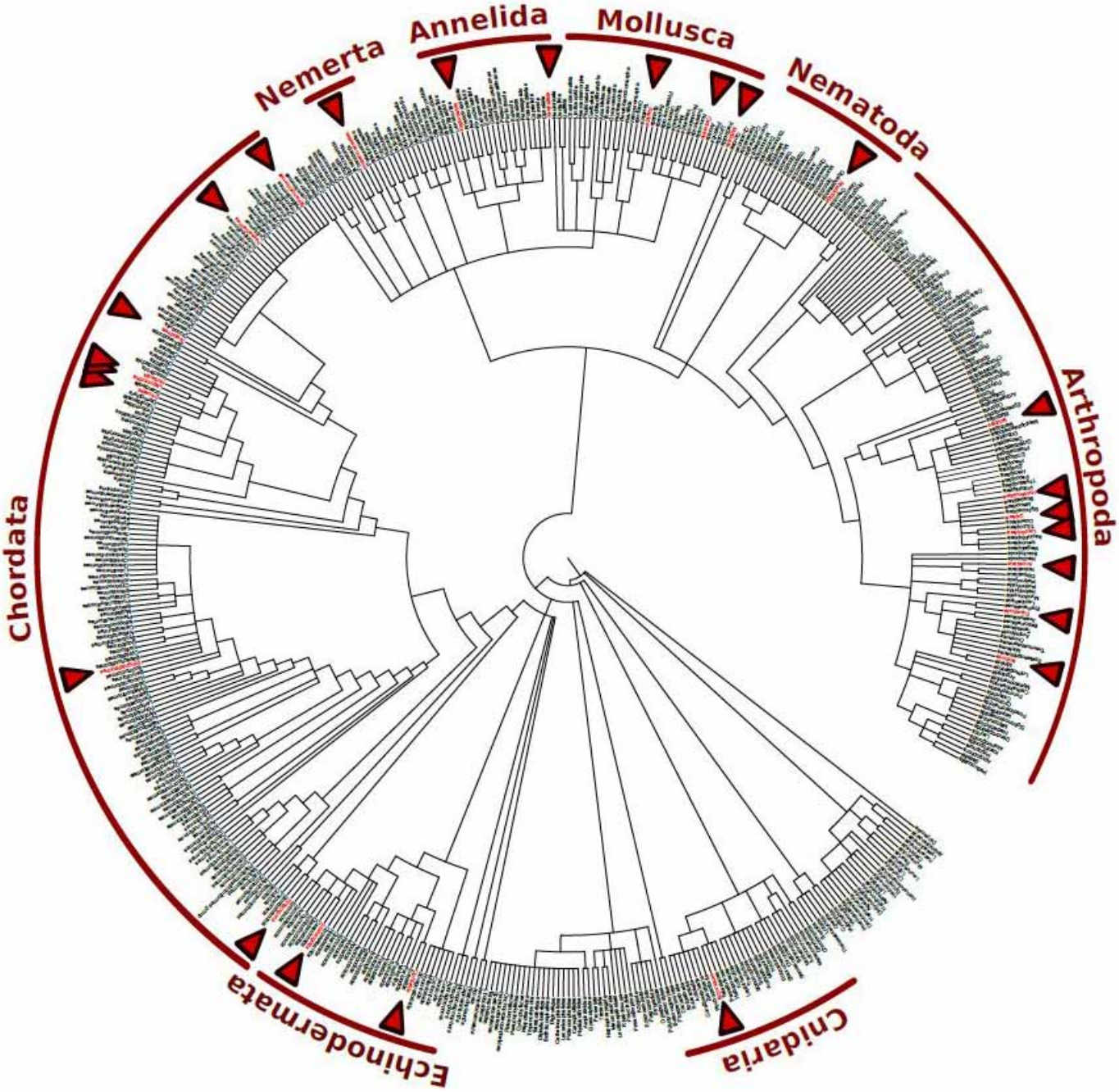
Linkage. The diversity of a neutral locus might be affected by selection at linked sites. This is particularly true of synonymous sites, which reside within coding sequences; that is, targets for natural selection. One would therefore predict a lower genetic diversity in species experiencing a low genomic average recombination rate. We lacked a recombination map in most of the analysed species; we therefore relied on generic taxonomic patterns to approach this issue. Eusocial hymenopterans are known to experience a recombination rate one order of magnitude higher than most animals[25]. In contrast, dipteran Culicidae (mosquitoes) experience relatively small amounts of recombination (median recombination rate in eusocial hymenoptera, 9.7 centimorgans per megabase (cM Mb^{-1}); median recombination rate in Culicidae, 0.3 cM Mb^{-1})[25]. The linkage effect would therefore predict a decreased π_s in Culicidae and an elevated π_s in eusocial hymenopterans. We observed the opposite: π_s varied from 0.0016 to 0.0058 in our five eusocial hymenopteran species, which is below the average metazoan

π_s (0.015), and one order of magnitude below the π_s of our three Culicidae species (0.016–0.041). This result, which is consistent with the propagule-size hypothesis, does not suggest that the between-species variation in genomic average recombination rate strongly influences our results.

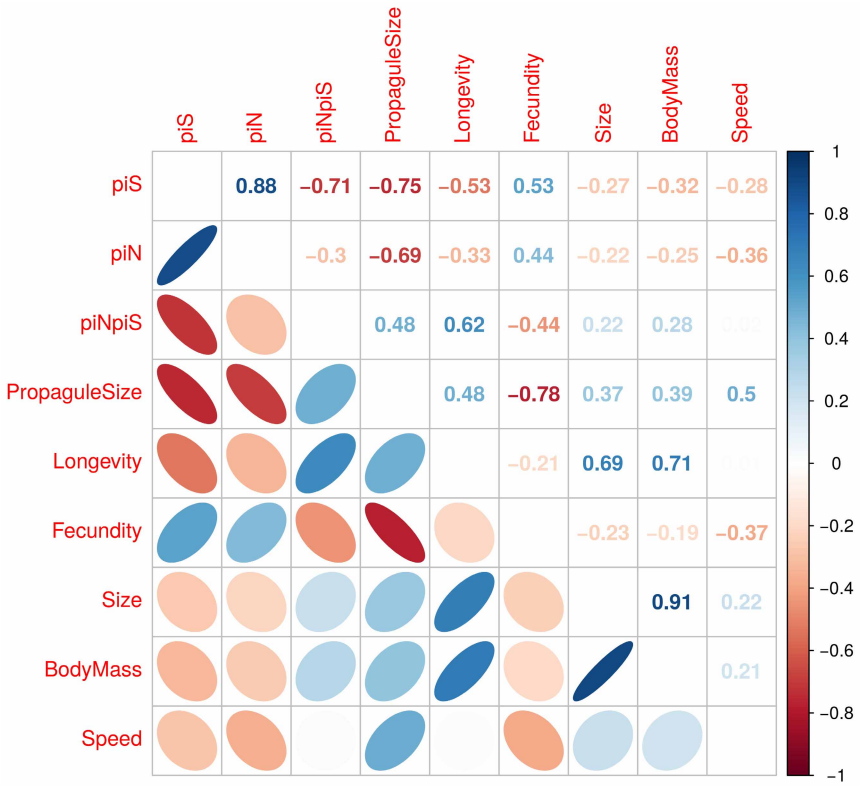
Population structure. The genetic distance between individuals was defined as $(H_b - H_w)/H_w$, where H_b is the probability of drawing two distinct alleles when sampling one copy from each of the two considered individuals, and H_w is the average heterozygosity of the two considered individuals. In species containing more than four individuals, the genetic distance was calculated for each pair of individuals and correlated to the geographic distance; the squared coefficient correlation, r^2, which measures genetic isolation by distance, ranged from 0.0008 to 0.73 (Supplementary Table 5). Consistent with the phylogeographic literature, it was high ($r^2 > 0.35$) and significant in, for example, *Ciona intestinalis* A, *Melitaea cinxia* and *Sepia officinalis*, and low ($r^2 < 0.02$, n.s.) in, for example, *Culex pipiens*, *Lepus granatensis* and *Crepidula fornicata*. The F_{it} statistic was significantly higher, on average, in terrestrial (median $F_{it} = 0.25$) than in marine (median $F_{it} = 0.02$) species (*t*-test, $P = 0.029$) when only species including at least five individuals were considered. No significant relationship was detected between π_s and absolute values of F_{it} ($P = 0.22$, $r^2 = 0.05$, only species with more than four individuals included), which does not suggest any confounding effect of population structure in our analysis.

Ethical statement. Living animals were manipulated according to the 'Charte Nationale Portant sur l'Ethique de l'Expérimentation Animale'. Sampling of protected species was performed under permits 53/2009 (Galicia, Spain, *Emys orbicularis*), 2009/11/12 (Aude, France, *Emys orbicularis*), 503/05/07/2006 (Pisa, Italy, *Emys orbicularis*), 35601-60/2005-4 (Slovenia, *Emys orbicularis*), and 009-01-1230/a34-455 (France, *Parus caeruleus*). *Aptenodytes patagonicus, Eudyptes moseleyi* and *Eudyptes filholi* individuals were sampled by Institut Polaire Français Paul Emile Victor, program IPEV 131. *Chelonoidis nigra* individuals were handled and sampled by the veterinarians and staff of the Zurich zoo (Switzerland), Rotterdam zoo (the Netherlands), and A Cupulatta zoo (France) in accordance with the Code of Practice and Code of Ethics established by the European Association of Zoos and Aquaria.

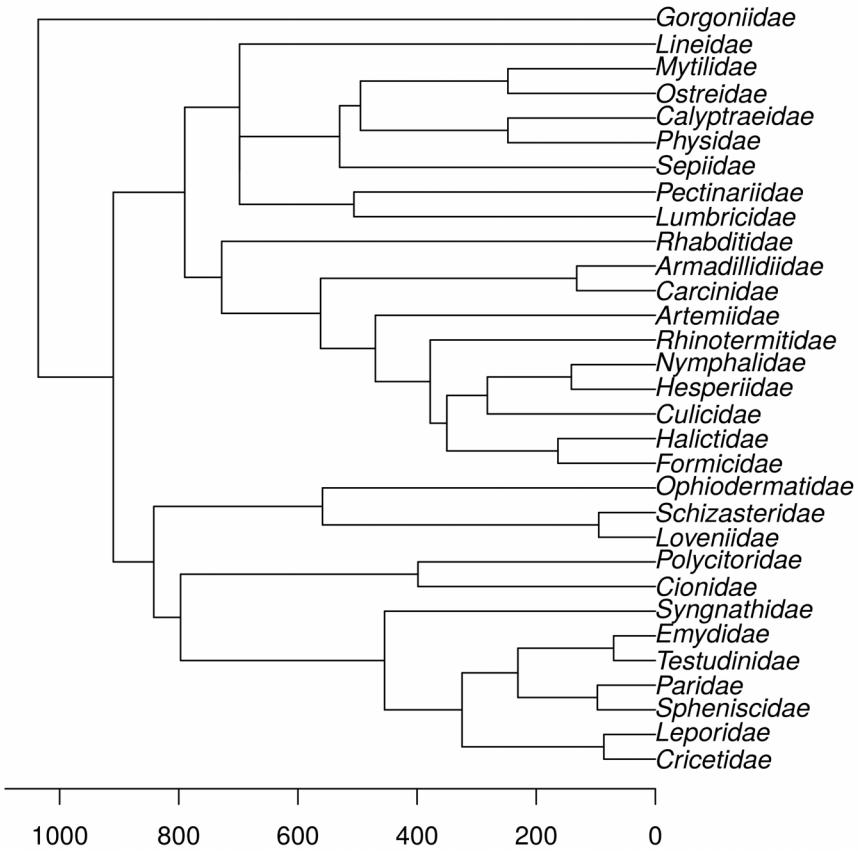
21. Gayral, P. *et al.* Next-generation sequencing of transcriptomes: a guide to RNA isolation in non-model animals. *Mol. Ecol. Resources* **11,** 650–661 (2011).
22. Guéguen, L. *et al.* Bio++: efficient, extensible libraries and tools for molecular evolution. *Mol. Biol. Evol.* **30,** 1745–1750 (2013).
23. Weir, B. S. & Cockerham, C. C. Estimating *F*-statistics for the analysis of population structure. *Evolution* **38,** 1358–1370 (1984).
24. Philippe, H. *et al.* Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.* **19,** 706–712 (2009).
25. Wilfert, L., Gadau, J. & Schmid-Hempel, P. Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity* **98,** 189–197 (2007).



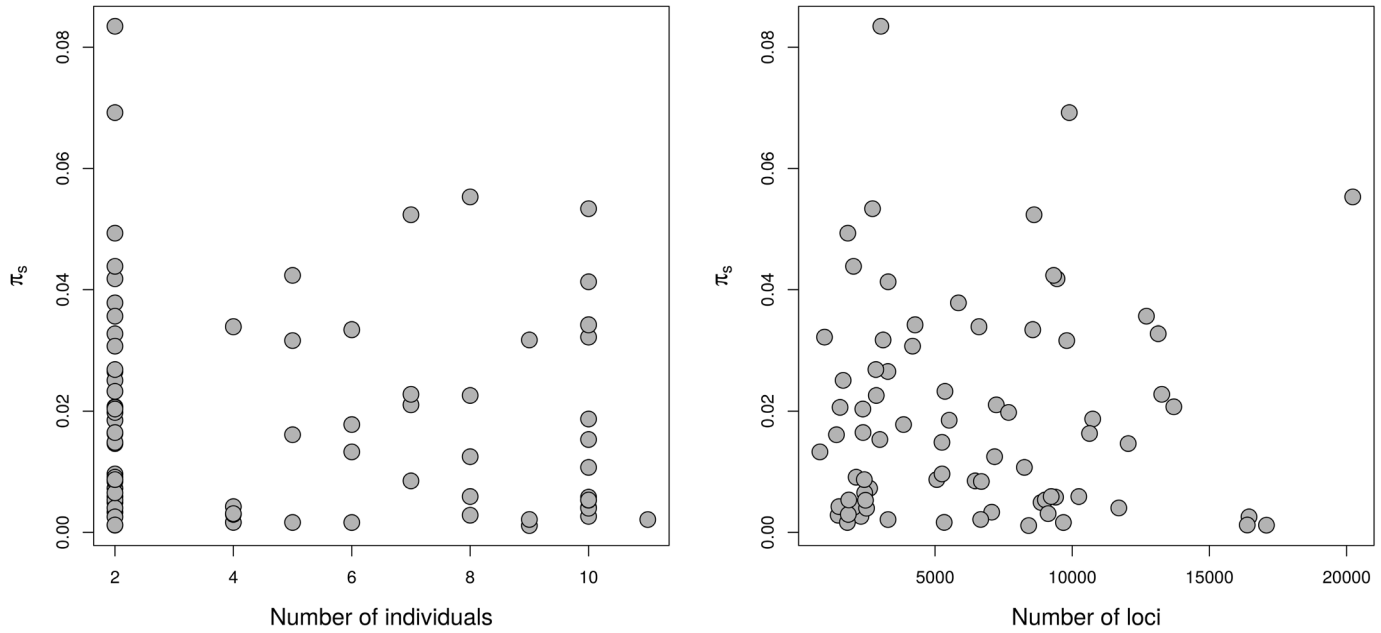
Extended Data Figure 1 | Phylogenetic tree of metazoan orders and the position of the taxa analysed in this study. The tree topology is consistent with the NCBI taxonomy. Red arrows identify 25 orders that were sampled. Five gastropod species from two distinct families (Calyptraeidae (*Crepidula fornicata*, *C. plana* and *Bostrycapulus aculeatus*) and Physidae (*Physa acuta* and *P. gyrina*)) are not represented because they lacked any assignment to an order in current taxonomy.



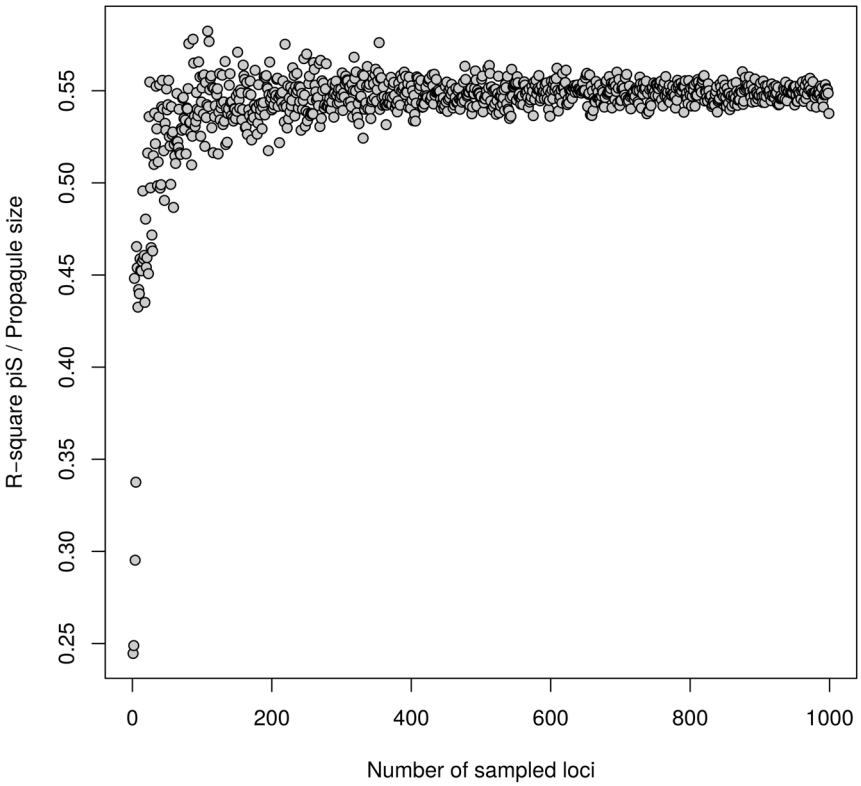
Extended Data Figure 2 | Correlations between genetic diversity and life history variables. Blue indicates a positive relationship, red a negative one; colour intensity is proportional to Pearson's correlation coefficient.



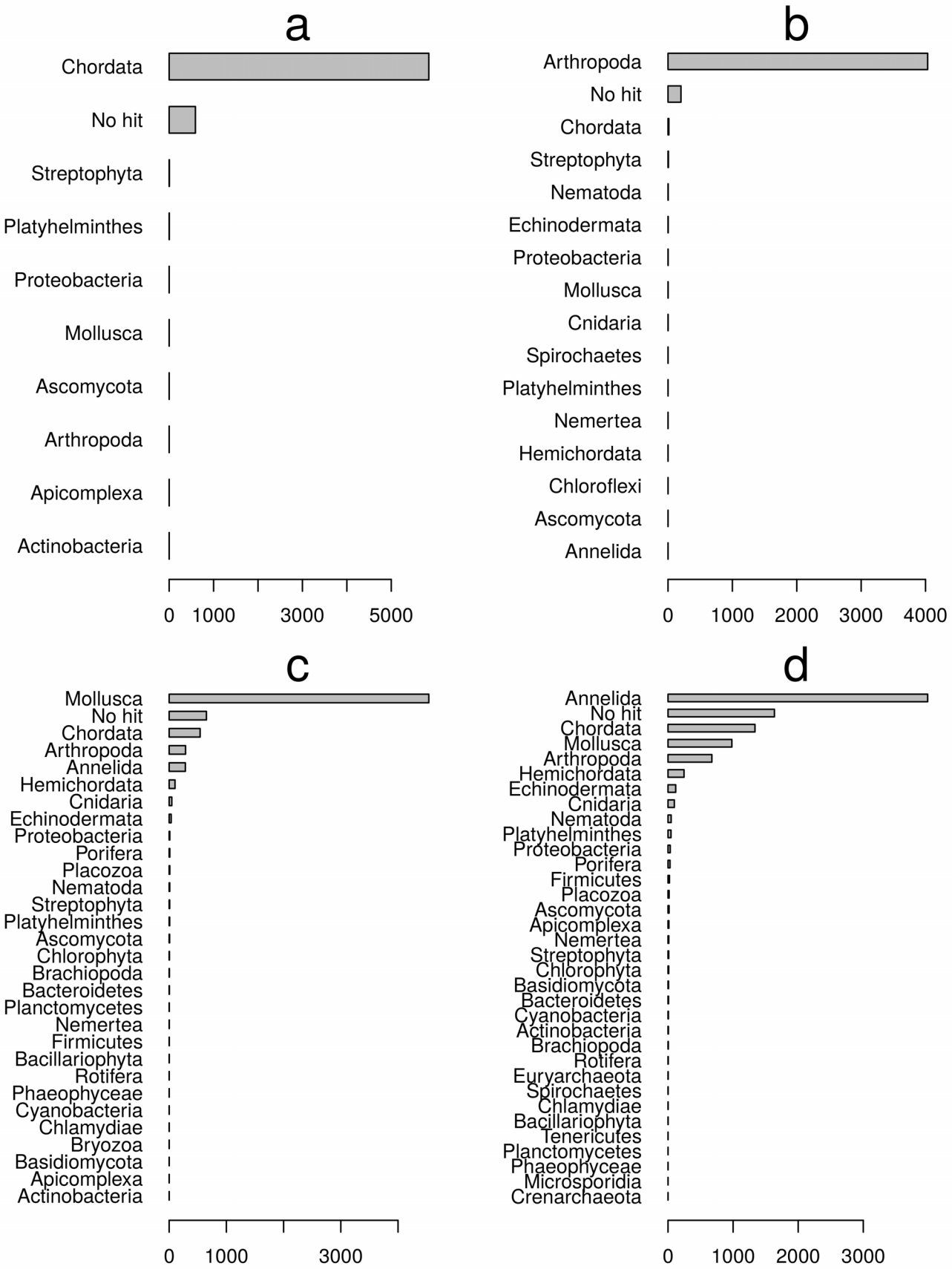
Extended Data Figure 3 | Family-level phylogenetic tree (31 families included). The scale is in million years of divergence.



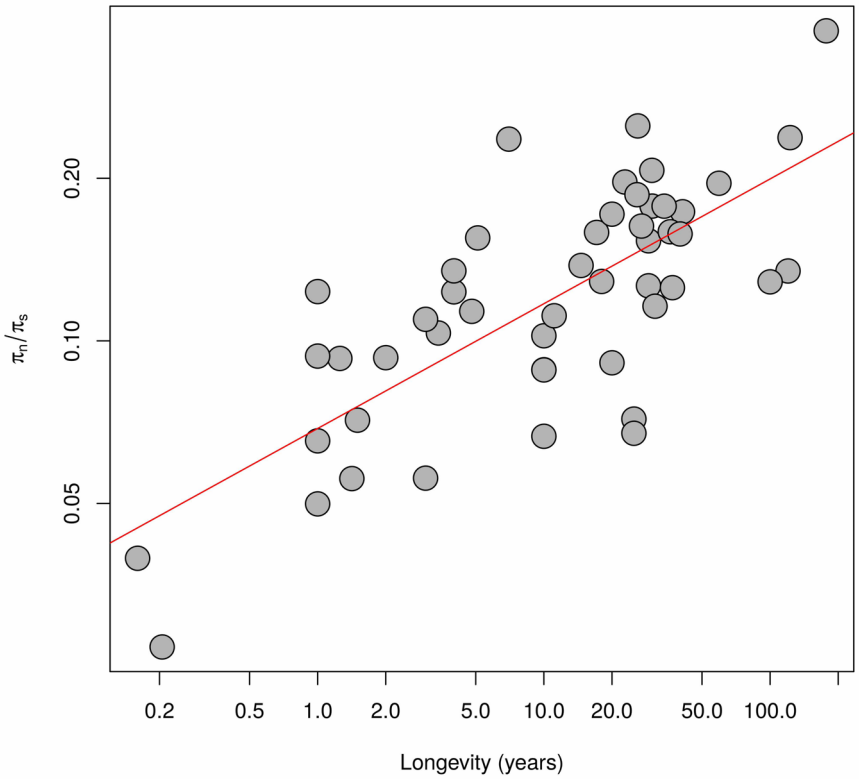
Extended Data Figure 4 | Absence of significant correlation between species genetic diversity with individual sampling size ($P = 0.47$, $r^2 = 0.007$) and locus sampling size ($P = 0.78$, $r^2 = 0.001$).



Extended Data Figure 5 | Relationship between the π_s/propagule-size r^2 and the number of sampled loci.



Extended Data Figure 6 | Phylum distribution of the first BLAST hit in four representative species. **a**, Common vole (*Microtus arvalis*). **b**, Glanville fritillary butterfly (*Melitaea cinxia*). **c**, Blue mussel (*Mytilus edulis*). **d**, Earthworm (*Allolobophora chlorotica*).



Extended Data Figure 7 | Correlation between π_n/π_s and maximum longevity ($P < 10^{-8}$, $r^2 = 0.54$). Only species with at least four individuals are included.