

Corresponding author(s): Rosa Fernández

Last updated by author(s): 8th June 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

We did not use any software for collecting data.

Data analysis

For the data processing workflow we used fastp v.0.20.0 to clean raw RNA-seq data, Trinity v2.11 to assembly clean data, TransDecoder v.5.5.0 to predict candidate coding regions, BlobTools2 to filter contaminant sequences, DIAMOND to annotate protein sequences prior to decontamination, Iso-Seq v.3.1.2 to process PacBio raw reads, ANGEL to predict open reading frames for the PacBio data, CD-HIT v.4.8.1 to cluster highly identical sequences, BUSCO v4.1.4 to assess the completeness of the gene content for the processed genomes and transcriptomes as a quality check measure.

For the phylogenomic analysis workflow we used OrthoFinder v.2.5.1 to infer orthologous groups, PREQUAL to filter bad quality sequences prior to the alignment, MAFFT v. 7.407 to align amino acid sequences, FastTree v.2.1.1 to perform maximum-likelihood phylogenetic inferences, ModelFinder to estimate the best model to perform phylogenetic inferences, IQTREE v.1.6.12 to perform maximum-likelihood phylogenetic inferences, PhyloPyPruner v.1.2.4 to filter short sequences and prune paralogs in given gene trees, GeneSortR to quantify phylogenetic usefulness and rank candidate orthogroups to be used in the molecular dating approach, MCMCtree to infer time-calibrated trees, Tracer v.1.7.1 to supervise different molecular clocks and alternative runs implemented in MCMCtree, iTOL v.6 to visualize and render the phylogenetic trees, BadiRate v.1.35 to explore gene family evolutionary dynamics with a maximum-likelihood approach, eggNOG-mapper v.2.1.6 to annotate amino acid sequences through an orthology based approach, GOEnrichment v.2.0 to perform gene ontology (GO) enrichment analysis, REVIGO to visualize and summarize the GO enrichment results, GOSemSim v.3.16 to calculate pairwise semantic similarity measures on lists of GO terms generating a GO similarity matrix, apcluster v.1.4.10 to perform an affinity propagation clustering on the GO similarity matrix, ComplexHeatmap v.3.15 and ggplot2 to plot the GO clustering results, GOexploreR v.1.2.6 and rrvgo v.1.8.1 to process and summarize long lists of GO terms based on their hierarchy and semantic similarity. These last four R packages were implemented with Rstudio v.4.2.1.

We used some custom bash and python scripts to process raw data following the approach used in Fernandez et al. (2022):

"fetch_longest_iso.py" and "remove_isoforms_proteome.sh" to filter the longest isoform per gene for transcriptomes and proteomes respectively in order to obtain candidate gene sets for further analyses and "blobtools.sh" and "extract_phyla_for_blobtools.py" for the decontamination protocol. Moreover, we used the custom R script "get_badirate_gains_losses.R" to process information about gene family expansion and contraction from the raw output of BadiRate.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Raw reads have been deposited in the National Center for Biotechnology Information (NCBI; BioProject accession number PRJNA902350). All processed data and custom scripts have been deposited in github (https://github.com/MetazoaPhylogenomicsLab/Balart_Garcia_et_al_gene_repertoire_evolution_subterranean_coleoptera).

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

Population characteristics

Recruitment

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Data exclusions

Replication

Randomization

Blinding

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	<i>Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).</i>
Research sample	<i>State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.</i>
Sampling strategy	<i>Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.</i>
Data collection	<i>Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.</i>
Timing	<i>Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.</i>
Data exclusions	<i>If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.</i>
Non-participation	<i>State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.</i>
Randomization	<i>If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.</i>

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We explored the molecular underpinnings of adaptations to life in caves in Coleoptera using a genome-wide phylogenomic approach. To do so, we investigate and characterize evolutionary dynamics of the gene repertoire across surface-dwelling and subterranean species that represent multiple independent underground colonizations both in terrestrial and aquatic beetle lineages. Our results demonstrate that genomic exaptation prior to the colonization of subterranean habitats was key to facilitating these transitions in terrestrial and aquatic lineages, and that both parallel and convergent evolution paved the way to adapting to life in caves across the Coleoptera.
Research sample	We collected surface-dwelling and subterranean species of the tribes Leptodirini, that have a terrestrial ecology, and species of the tribes Bidessini and Hydroporini, that have an aquatic ecology. These three tribes correspond to two different subfamilies of beetles (Cholevinae and Hydroporinae) that diverged more than 300 Mya. Each tribe include several lineages that colonized independently terrestrial and aquatic subterranean habitats, representing ideal systems to explore parallel evolution and functional convergence in subterranean fauna from a macroevolutionary perspective. We selected 21 species representing at least six independent underground colonization events, including surface-dwelling and subterranean ecologies. Moreover we included 1 species of the tribe Catopini, representing a closely related lineage to Leptodirini.
Sampling strategy	We sampled between 15 and 2 specimens per species, depending on the availability and to ensure a efficient RNA isolation.
Data collection	We generated data for a total of 21 species, including 12 from the tribe Leptodirini, 1 from the tribe Catopini (i.e. a closely related tribe to Leptodirini), 2 from the tribe Hydroporini and 6 from the tribe Bidessini. Data collection for species of the tribe Leptodirini was in charge of Pau Balart-Garcia, Slavko Polak and Ignacio Ribera who participated in the sampling procedures such as selecting the species based on their phylogenetic relationships, planning the field trips and collecting the specimens. Wet lab procedures were done by Pau Balart-Garcia. Data generation for species of the tribes Bidessini and Hydroporini species was in charge of Steven J. B. Cooper and wet lab procedures were done by Tessa M. Bradford and Perry G. Beasley-Hall. Moreover, we collected publicly available transcriptomic and genomic raw data of other Coleoptera species and two outgroups (i.e. Strepsiptera and Neuroptera) in order to include a broader phylogenetic context to study gene repertoire evolution. This part was in charge of Pau Balart-Garcia, who selected and processed all the raw data.
Timing and spatial scale	Data was collected between 2018 and 2020 for several parallel projects. Once we collected all the data we proceeded with the processing analyses in order to follow the same strategy for all the data sets.
Data exclusions	During the collection of publicly available data we discarded several genomes and transcriptomes that did not pass the final quality check, which consist of estimating the completeness of the gene content based on the sets of benchmarking universal single-copy

orthologs with BUSCO. The data sets with less than 80 % of complete plus fragmented genes were excluded.

Reproducibility

Randomization

Blinding

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions

Location

Access & import/export

Disturbance

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a ☐ Involved in the study

☐ ☐ Antibodies

☐ ☐ Eukaryotic cell lines

☐ ☐ Palaeontology and archaeology

☐ ☒ Animals and other organisms

☐ ☐ Clinical data

☐ ☐ Dual use research of concern

Methods

n/a ☐ Involved in the study

☐ ☐ ChIP-seq

☐ ☐ Flow cytometry

☐ ☐ MRI-based neuroimaging

Antibodies

Antibodies used	Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.
Validation	Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.
Authentication	Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.
Mycoplasma contamination	Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

Palaeontology and Archaeology

Specimen provenance	Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.
Specimen deposition	Indicate where the specimens have been deposited to permit free access by other researchers.
Dating methods	If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.
☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.	
Ethics oversight	Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals	The study did not include laboratory animals.
Wild animals	The surface-dwelling species of the tribe Leptodirini (<i>Bathysciola rugosa</i> , <i>Bathysciola zariquieyi</i> , <i>Bathysciola ovata</i>) and a species of the tribe Catopini (<i>Catops fuliginosus</i>) were collected in forest-litter using an entomological litter reducer and a metallic sieve of 3mm aperture. The subterranean species (<i>Breulia triangulum</i> , <i>Parvospeonomus canyellesi</i> , <i>Astagobius angustatus</i> , <i>Prospelaebates brelihi</i> , <i>Leptodirus hohenwartii</i> , <i>Troglocharinus ferrerii</i> , <i>Paranillochlamys urgellesi</i> , <i>Oryotus schmidtii</i> , <i>Bathysciomorpha slavkoi</i>) were captured manually in caves using an entomological aspirator. The surface-dwelling species of the tribes Bidessini (<i>Allodessus bistrigatus</i> and <i>Limbodessus amabilis</i>) and Hydroporini (<i>Paroster nigroadumbratus</i>) were captured using aquatic nets in shallow still water and temporary pools that tend to be formed in spring. Groundwater-monitoring boreholes were used to access the habitat of the subterranean species of Bidessini (<i>Neobidessodes gutteridgei</i> , <i>Limbodessus cueensis</i> , <i>Limbodessus hinkleri</i> , <i>Limbodessus palmuloides</i>) and Hydroporini (<i>Paroster macrosturtensis</i>) and specimens were captured with a mesh net attached to a fishing rod. When the localities were remote, we kept specimens in RNA-later and cold conditions to reduce risk of RNA degradation. Otherwise, specimens were kept alive in thermo boxes with moss to keep the humidity (terrestrial beetles) or tubes with the same water where they were found (aquatic beetles). Samples were transported in dark and constant temperature simulating the conditions of their habitat. Specimens were identified and killed flash-frozen with liquid nitrogen once we arrived to the laboratory. Specimens were killed in order to extract RNA and generate RNA-seq data.
Reporting on sex	Not relevant for this study.
Field-collected samples	Specimens were stored at -80°C immediately when arriving to the laboratory.
Ethics oversight	To our knowledge there is not official ethics protocol for wild insects.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	<i>Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.</i>
Study protocol	<i>Note where the full trial protocol can be accessed OR if not available, explain why.</i>
Data collection	<i>Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.</i>
Outcomes	<i>Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.</i>

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☒	☐ Public health
☒	☐ National security
☒	☐ Crops and/or livestock
☒	☐ Ecosystems
☒	☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
☒	☐ Demonstrate how to render a vaccine ineffective
☒	☐ Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐ Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐ Increase transmissibility of a pathogen
☒	☐ Alter the host range of a pathogen
☒	☐ Enable evasion of diagnostic/detection modalities
☒	☐ Enable the weaponization of a biological agent or toxin
☒	☐ Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links <i>May remain private before publication.</i>	<i>For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.</i>
Files in database submission	<i>Provide a list of all files available in the database submission.</i>
Genome browser session (e.g. UCSC)	<i>Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.</i>

Methodology

Replicates	<i>Describe the experimental replicates, specifying number, type and replicate agreement.</i>
------------	---

Sequencing depth	<i>Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.</i>
Antibodies	<i>Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.</i>
Peak calling parameters	<i>Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.</i>
Data quality	<i>Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.</i>
Software	<i>Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.</i>

Flow Cytometry

Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	<i>Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.</i>
Instrument	<i>Identify the instrument used for data collection, specifying make and model number.</i>
Software	<i>Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.</i>
Cell population abundance	<i>Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.</i>
Gating strategy	<i>Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.</i>
☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.	

Magnetic resonance imaging

Experimental design

Design type	<i>Indicate task or resting state; event-related or block design.</i>
Design specifications	<i>Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.</i>
Behavioral performance measures	<i>State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).</i>

Acquisition

Imaging type(s)	<i>Specify: functional, structural, diffusion, perfusion.</i>
Field strength	<i>Specify in Tesla</i>
Sequence & imaging parameters	<i>Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.</i>
Area of acquisition	<i>State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.</i>
Diffusion MRI	☐ Used ☐ Not used

Preprocessing

Preprocessing software	<i>Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).</i>
Normalization	<i>If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.</i>
Normalization template	<i>Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.</i>
Noise and artifact removal	<i>Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).</i>
Volume censoring	<i>Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.</i>

Statistical modeling & inference

Model type and settings	<i>Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).</i>
Effect(s) tested	<i>Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.</i>
Specify type of analysis:	☐ Whole brain ☐ ROI-based ☐ Both
Statistic type for inference (See Eklund et al. 2016)	<i>Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.</i>
Correction	<i>Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).</i>

Models & analysis

n/a	Involvement in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis
Functional and/or effective connectivity	<i>Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).</i>
Graph analysis	<i>Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).</i>
Multivariate modeling and predictive analysis	<i>Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.</i>