Nucleic Acids Research, 2019 **1**
doi: 10.1093/nar/gkz796

Gephebase, a database of genotype–phenotype relationships for natural and domesticated variation in Eukaryotes

Virginie Courtier-Orgogozo [1,*], Laurent Arnoult[1], Stéphane R. Prigent[1], Séverine Wiltgen[2] and Arnaud Martin[3,*]

[1]Institut Jacques Monod, CNRS, UMR 7592, Université de Paris, Paris, France, [2]Version Ops, Montpellier, France and [3]Department of Biological Sciences, The George Washington University, Washington, DC, USA

Received June 19, 2019; Revised August 21, 2019; Editorial Decision August 30, 2019; Accepted September 06, 2019

ABSTRACT

Gephebase is a manually-curated database compiling our accumulated knowledge of the genes and mutations that underlie natural, domesticated and experimental phenotypic variation in all Eukaryotes—mostly animals, plants and yeasts. Gephebase aims to compile studies where the genotype–phenotype association (based on linkage mapping, association mapping or a candidate gene approach) is relatively well supported. Human clinical traits and aberrant mutant phenotypes in laboratory organisms are not included and can be found in other databases (e.g. OMIM, OMIA, Monarch Initiative). Gephebase contains more than 1700 entries. Each entry corresponds to an allelic difference at a given gene and its associated phenotypic change(s) between two species or two individuals of the same species, and is enriched with molecular details, taxonomic information, and bibliographic information. Users can easily browse entries and perform searches at various levels using boolean operators (e.g. transposable elements, snakes, carotenoid content, Doebley). Data is exportable in spreadsheet format. This database allows to perform meta-analyses to extract global trends about the living world and the research fields. Gephebase should also help breeders, conservationists and others to identify promising target genes for crop improvement, parasite/pest control, bioconservation and genetic diagnostic. It is freely available at www.gephebase.org.

INTRODUCTION

Mutations form the raw bulk of heritable variation upon which traits evolve. Identifying the DNA sequence modifications that drive phenotypic changes is a primary goal of modern genetics, and could greatly improve our understanding of the mechanisms behind biodiversity and adaptation. However, this research program would be most successful if it reaches comparative capacity, for instance by allowing us to detect trends across the Tree of Life (1–3). Advances in genome sequencing and editing are accelerating the rate of discovery of the loci of evolution at a quick pace, making data integration increasingly challenging, and it is now crucial to develop a universal, single resource integrating this body of knowledge. As of today, compilations of genotype–phenotype relationships are available for a limited number of species in taxon-specific databases, for example OMIA for animals (4), OMIM for humans (5), TAIR for *Arabidopsis* (6), FlyBase for *Drosophila* (7), or the Monarch Initiative across the main laboratory animal model species (8,9). To date, there are no databases that consolidate genotype–phenotype relationships related to natural evolutionary cases across all Eukaryotes. For example, evolutionary changes in tigers, butterflies, monkeyflowers, or any non-traditional model organism are lacking from existing genotype–phenotype databases, preventing comparative insights on the diversity and similarities of sequence modifications that fuel the generation of observable differences in the living world.

To fill this gap, we developed Gephebase, a manually curated database that gathers published data about the genes and the mutations responsible for evolutionary changes in all Eukaryotes (mostly animals, yeasts and plants) into a single website. The content of Gephebase was developed over the past 10 years, with previous versions of the dataset published as supplementary spreadsheet files associated to

*To whom correspondence should be addressed. Tel: +33 1 57278043; Fax: +33 1 57278087; Email: virginie.courtier@ijm.fr
Correspondence may also be addressed to Arnaud Martin. Tel: +1 2029942384; Email: arnaud@gwu.edu
Present address: Stéphane R. Prigent, Institut de Systématique, Evolution, Biodiversité, ISYEB, Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, Paris, France.

© The Author(s) 2019. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz796/5572566 by arnaud@gwu.edu on 24 September 2019

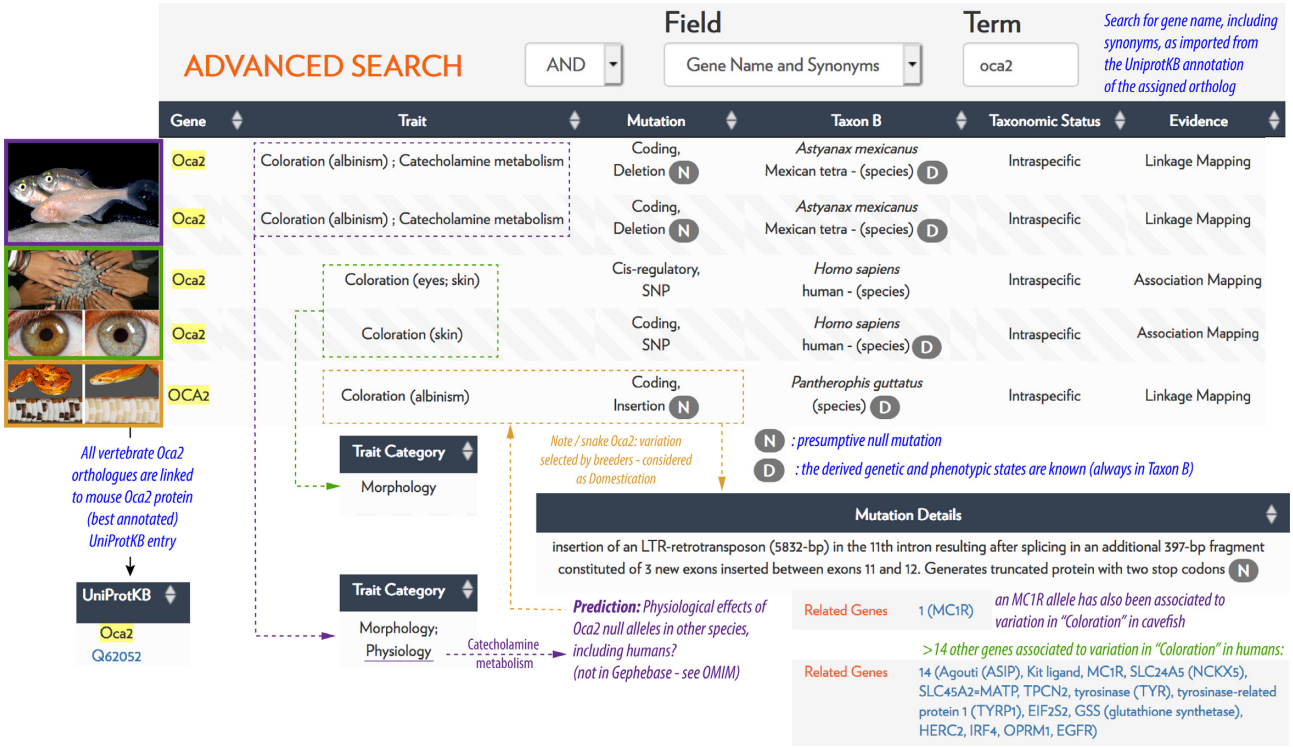


Figure 1. A snapshot view of the potential of Gephebase for comparative genetics. Part of the Summary Output Table of a gene name search for 'oca2' is provided here as an example, with a montage of Gephebase summary outputs, pictures, and annotations in color. Interestingly, the cavefish studies reveal that the *Oca2* null alleles also affect catecholamine metabolism in cavefish. Gephebase can be used as a hypothesis generator: by juxtaposition of these entries, one may expect similar effects in the corn snake *Oca2* mutants that remain to be tested. 'N' means that the mutation is *null*. 'D' means that Taxon B is inferred to bear the *derived* trait. Photo Credits, top to bottom: Reproduced with permission of Richard Borowsky from 'Beyond the zebrafish: diverse fish species for modeling human disease, Schartl M, Disease Models & Mechanisms 2014, 7: 181–192; doi: 10.1242/dmm.012245, https://creativecommons.org/licenses/by/4.0/. Reproduced from 'https://www.flickr.com/photos/neokainpakistan/324823058/', https://creativecommons.org/licenses/by/2.0/. Reproduced with permission from Aaron Pomerantz. Reproduced from 'Amelanism in the corn snake is associated with the insertion of an LTR-retrotransposon in the OCA2 gene, Saenko SV, Lamichhaney S, Martinez Barrio A, Rafati N, Andersson L, Milinkovitch MC, Sci Rep 2015,5: 17118; doi: 10.1038/srep17118, https://creativecommons.org/licenses/by/4.0/.

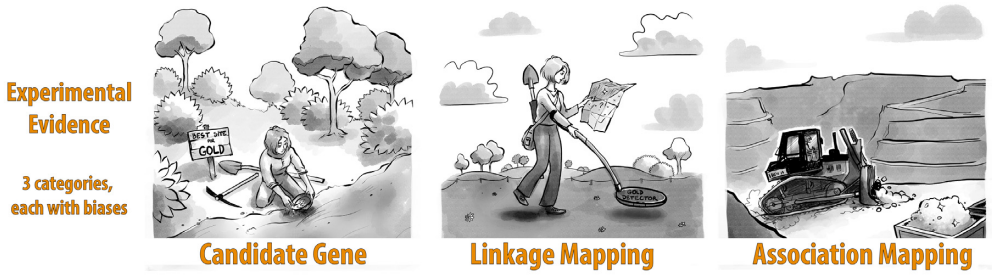
two review articles, which successively compiled 331 entries (1), and 1008 entries (2). These datasets have been used by various authors to highlight several trends regarding the genetic basis of natural variation. For example, based on these compilations it was found (a) that the mutations responsible for long-term evolution have distinct properties than the mutations responsible for short-term evolution (1,10), (b) that certain types of mutations are more likely to be fixed than others during the course of evolution (11), (c) that independent evolution of similar traits in distant lineages often involves mutations in the same orthologous gene (2), (d) that current data are biased towards a limited number of model organisms (12) and (e) that the cis-regulatory tinkering of signaling ligand genes is a recurring mode of morphological evolution (13).

We have now created an online version of the Gephebase database, accessible at www.gephebase.org, and we describe here its various features.

SNAPSHOT SUMMARY

In short, Gephebase is a searchable, manually curated knowledge-base of the genetic loci of phenotypic variation. Each entry is a pair of alleles associated to a trait variation, be it naturally existing (inter- or intraspecific), selected by breeders (domestication), or occurring during a bout of experimental evolution in the lab. For instance, forward genetic studies have determined that independently derived null mutations of the *Oca2* gene have caused an amelanic phenotype in at least two subterranean populations of cavefish (14), and in a breed of corn snake that has been selected for the pet trade (15). A Gephebase search for the *Oca2* gene name reveals these findings, accessible in summary tables (Figure 1) or in a more detailed output (entry view, and CSV spreadsheet format). Gephebase also indicates that some *Oca2* allelic variants have been identified by Genome-Wide Association Studies of pigment variation. Importantly, the focus of Gephebase is always on genetic variations that emerge naturally - it never includes laboratory variants that were generated by random or directed mutagenesis. Thus the *Oca2* CRISPR knockout phenotypes that have been generated in frogs (16) do not have a dedicated Gephebase entry; the cavefish *Oca2* CRISPR/TALEN knockout phenotypes (17,18) do not have a dedicated entry either, but are used as Additional References to support the functionality of the two natural *Oca2* null alleles in Gephebase. This makes Gephebase complementary to the Monarch Initiative database, which compiles gene-to-phenotype relation-

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz796/5572566 by arnaud@gwu.edu on 24 September 2019



Experimental Evidence	Candidate Gene	Linkage Mapping	Association Mapping
Experimental Principle	**Reverse Genetics**: looking for sequence differences and trait effects based on previous studies of a given gene	**Forward Genetics**: trait mapping in hybrids obtained from laboratory crosses, using recombination over a few generations	**Forward Genetics**: statistical SNP/character state association in large cohorts, using recombination over many generations
Example	66 cases of color variation associated to *MC1R* coding mutations in vertebrates	F2 crosses between melanic and amelanic phenotypes in cavefish : identification of *MC1R* and *Oca2* alleles in distinct cave populations	GWAS of human pigmentation (skin, hair, eyes): identification and confirmation of causal variants at >15 genes including *Oca2* p.His615Arg in Eastern Asia
Ascertainment Bias on Locus Identification	**High**	**Low to Intermediate** (depending on resolution / cross size)	**Low**
Molecular Type Bias	Favors identification of **coding mutations**	Little molecular bias	Can miss structural variants (short read genotyping)
Trait Type Bias	Favors traits with small molecular targets, large-effect size	Amenable to dissection of complex traits with small-effect size (large crosses, multiparental families)	Most common approach for complex traits with small-effect size
Taxonomic Breadth	**Large**	**Narrow**, limited to interfertile lineages (populations or sister species)	**Very narrow**, limited to polymorphic or intermixing populations

Figure 2. Three kinds of experimental strategies for identifying gene-to-phenotype variations. Gephebase focuses on genes that have been mapped using a forward genetics approach, and supported as the causal agents by sufficient evidence. Candidate gene approaches are also included and cover broader phylogenetic distances (e.g. human/chimp), but tend to be biased towards the identification of coding changes for relatively simple traits. The search for the genes and mutations that drive phenotypic variation is somewhat analogous to searching gold: from left to right, targeted candidate gene approaches can identify variants of large effects at loci previously identified in other organisms; in a linkage mapping approach, the experimenter walks on chromosomes to narrow down the causal genetic interval, and can increase resolution and sensitivity with the analysis of more recombinants; association mapping (e.g. GWAS) takes advantage of statistical power across large datasets to extract genetic variants in linkage disequilibrium with the causal mutations. GWAS: Genome-Wide Association Studies, SNP: Single Nucleotide Polymorphism.

ships in humans, as well as in laboratory organisms and mutants generated by reverse genetics, but does not include non-model species such as cavefishes and corn snakes (8,9).

DATABASE CURATION AND STRUCTURE

Criteria for inclusion in Gephebase

Gephebase includes cases of domestication, experimental evolution and natural evolution but no human clinical phenotypes. Gene expression levels (eQTL) and DNA methylation patterns are not included. All kinds of traits above this level, whether morphological, physiological or behavioral, are included. For example, we include 'Recombination rate', 'Telomere length', 'Hematopoiesis', 'Hybrid incompatibility'.

Cases of genomic regions associated with a trait for which the underlying gene(s) is unclear are not included in Gephebase. Cases where the gene has been identified, but not the exact mutation, are included. Stringent inclusion criteria are used so that Gephebase compile only studies where a given genotype–phenotype association is well supported or understood. Association Mapping studies are included only if there is additional experimental support for the given gene. Candidate Gene studies require conclusive functional assays for inclusion in Gephebase. Overall, gene-to-phenotype links identified by Linkage Mapping with resolutions <500 kb have priority in the dataset. There are multiple types of experimental evidence that led to the discovery of a relationship between a genetic mutation and a phenotypic change. For sake of simplicity and efficiency, each gene-phenotype association is attributed only one type of Experimental Evidence among three possibilities: 'Association Mapping', 'Linkage Mapping', or 'Candidate Gene' (Figure 2). When several methods were used, the least biased one is chosen by the curator (Table 1). And when new evidences emerge, they are added to the entry.

Curation protocol

Searches for relevant papers to be included in the database are done manually by our team of curators. We screen major journals in evolutionary genetics, perform keyword searches using online search tools, and we pay particular attention to citations in primary research articles as well as in review papers. The 'Suggest an article' button in the top bar menu allows users to suggest articles to our curation team. Of note, our curations efforts have been maximal until 2013 and then relaxed due to our inability to support a full-time curator. Following our inclusion criteria, we estimate that the database is close to comprehensive for studies

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz796/5572566 by arnaud@gwu.edu on 24 September 2019

Table 1. List of the fields of a Gephebase entry (in order of appearance on the View-Entry page)

Field name	Description
Gephebase Gene	Generic gene name used in Gephebase across many species.
GepheID	Identifier of the gephe entry. One entry corresponds to a single mutation, or a group of linked mutations within a single gene, that has been associated with a phenotypic trait.
Main curator	Name of the curator who created the entry. The entry may have been modified later by another curator.
Trait Category	Only three possibilities: “Morphology”, “Physiology”, “Behavior” or a combination of them for ambiguous cases.
Trait	Controlled vocabulary describing the phenotypic trait at a broad level (eg. “Coloration”). Precisions can be indicated in parentheses.
Trait State in Taxon A	Free text (eg. “brown eyes”, “sensitive to tetrodotoxin”). If the direction of evolutionary change can be inferred, Taxon A is the taxon bearing the ancestral phenotypic state. If not, Taxon A is chosen arbitrarily as one of the two compared taxa.
Trait State in Taxon B	Free text (eg. “brown eyes”, “sensitive to tetrodotoxin”). If the direction of evolutionary change can be inferred, Taxon B is the taxon bearing the derived phenotypic state. If not, Taxon B is chosen arbitrarily as one of the two compared taxa.
Ancestral State	3 possibilities: “Taxon A” if Taxon A is inferred to bear the ancestral phenotypic state, “Unknown” if the direction of change cannot be inferred, “Data not curated”.
Taxonomic Status	5 possibilities: “Experimental Evolution”, “Domesticated”, “Intraspecific”, “Interspecific”, “Intergeneric or Higher”.
Taxon A	Name of the taxon inferred to bear the ancestral trait. If the direction of change is unknown, the two compared taxa are arbitrarily assigned to Taxon A and Taxon B. The fields “Latin Name”, “Common Name”, “Synonyms”, “Rank”, “Lineage” and “Parent” are directly fetched from NCBI using the taxon ID.
Taxon B	Name of the taxon inferred to bear the derived trait. If the direction of change is unknown, the two compared taxa are arbitrarily assigned to Taxon A and Taxon B. The related fields “Latin Name”, “Common Name”, “Synonyms”, “Rank”, “Lineage” and “Parent” are directly fetched from NCBI using the taxon ID.
Is Taxon A/B an Infraspecies?	“Yes” indicates that the phenotypic trait was observed in a differentiated gene pool that is associated to a name (*eg.*, a subspecies, a geographically restricted natural population ; a strain, breed or cultivar). As an exception, modern human populations are never encoded as infraspecies in Gephebase.
Taxon A/B Description	Additional information regarding Taxon A/B (eg. location, name of the subspecies or strain).
Generic Gene Name	Gene name as in UniprotKB. The related fields “Synonyms”, “String”, “Sequence similarities” and “GO” are fetched from UniProtKB using the UniProtKB ID.
UniProtKB	Well-annotated ortholog from a model organism (*eg. H. sapiens* or *M. musculus* for vertebrates, *D. melanogaster* for insects, *A. thaliana* for angiosperms), used to fetch gene ontologies from UniProtKB.
Gene Ontology Terms	Terms directly imported from UniProtKB using the UniProtKB ID.
GenebankID or UniProtKB	Genebank ID or UniProtKB ID of the gene in Taxon A or Taxon B.
Presumptive Null	4 possibilities: “Yes” when the mutation is supposed to abolish gene function (premature stop codon, complete loss of expression, etc.), “No” (a functional protein remains, subtle cis-regulatory changes), “Unknown”, “Not curated”.
Molecular Type	6 possibilities: “Unknown”, “Cis-regulatory” (includes methylation), “Coding” (including splicing mutations), “Gene Loss” (large deletion), “Gene Amplification” (includes copy number variations and gene duplications), “Other”.
Aberration Type	9 possibilities: “SNP”, “Insertion”, “Deletion”, “Indel” (when direction of change unknown), “Inversion”, “Translocation”, “Complex Change”, “Epigenetic Change”, “Unknown” (when multiple candidate mutations).
Aberration Size	For mutations that are not classified as ”SNP”. 8 possibilities: “Unknown”, “1-9 bp”, “10-99 bp”, “100- 999 bp”, “1-10 kb”, “10-100 kb”, “100-1000 kb”, “>1 Mb”.
SNP Coding Change	For mutations classified as ”SNP”. 5 possibilities: “Not applicable”, “Nonsynonymous”, “Nonsense”, “Synonymous”, “Unknown”.
Molecular Details of the Mutation	Free text describing the candidate mutation(s) at the genotypic level.
Experimental Evidence	3 possibilities: “Linkage Mapping”, “Association Mapping”, “Candidate Gene” - when several pieces of evidence, the best one is chosen (Linkage Mapping is prefered to Association Mapping, which is prefered to Candidate Gene).
Main reference	Main or first article supporting the relationship between the genetic locus and the phenotypic difference. The related fields “Authors” and “Abstract” are fetched from NCBI PubMed using the PubMed ID.
Additional references	Articles providing additional information regarding the relationship between the genetic locus and the phenotypic difference. The related fields “Authors” and “Abstract” are fetched from NCBI PubMed using the PubMed ID.
Related Genes	Gephebase entries corresponding to other genes associated with the same phenotypic trait in the same Taxon A and/or Taxon B.
Related Haplotypes	Gephebase entries corresponding to other mutations in the same gene and in the same Taxon A and/or Taxon B that have been identified in other individuals.
Comments	Free Text which provides additional information about the genotype-phenotype relationship. The following tags are used: @SexualTrait @Fitness @Pleiotropy @GxE @Parallelism @Splicing @TE @GeneDuplication @GeneConversion @ChimericGene @SuperGene @Inversion @Introgression @ILS @HGT @BalancingSelection @Epistasis @SeveralMutationsWithEffect @SeveralCandidateMutations @TwoNucleotideChangesInSameCodon @SuccessiveMutationsAtSameCodon @AllelicSeries.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz796/5572566 by arnaud@gwu.edu on 24 September 2019

published prior to 2013, and up to 30% complete for the 2014–2019 period.

Technical overview of the database and the web interface

Gephebase was developed using the Symfony framework (v2.8) and PHP (v5.6 compatible 7). MariaDB (v10) is used to store data. The database consists of 33 tables including users management and logs. The main table links genotypic change, phenotypic change, references and validation information. Most fields of other tables are automatically retrieved from NCBI databases. The import procedure uses the NCBI E-utility interface with XML to fill the corresponding tables. Gephebase entries of the main table can be imported and exported through a csv file. For convenience, fields retrieved automatically can be present in the csv file even though they are fetched and stored in other tables.

The project code was put under version control (git) from its inception. The code is available in the GitHub repository https://github.com/Biol4Ever/Gephebase-database under GPL (GNU General Public License) version 3.

DATABASE CONTENT AND WEB INTERFACE

Organization of the data into entries

This database currently comprises >1900 entries (Supplementary Table S1). One entry corresponds to a single mutation, or a group of linked mutations within a single gene, either between two closely related species or between two individuals of the same species, and its associated phenotypic change (Figure 3). For cases of repeated evolution (2), we use the following conventions. When several mutations are found within the same gene in a given individual, with each mutation affecting the trait of interest—i.e. several causative mutations within a haplotype, intralineage hotspot (2)—all are grouped into a single entry. In contrast, when independent mutations occur in the same gene in distinct individuals of the same species, leading to similar phenotypic changes (intraspecific parallel evolution, convergent evolution), we chose to create different entries for each lineage-specific haplotype. In cases where a genetic variant was invented once, and then spread into multiple branches of the gene pool, via Incomplete Lineage Sorting (ILS), secondary hybridization (introgression among organisms that are not completely reproductively isolated) or horizontal transfer, a single entry is created and multiple taxa with the derived trait are reported in the entry.

The various fields of a Gephebase entry

A Gephebase entry (Figure 3) comprises 29 manually curated fields regarding bibliographical information, molecular details and taxonomic information; some are free-text and others rely on controlled vocabulary (Table 1). In addition, for each entry, 20 fields are automatically fetched based on manually curated data, from NCBI Taxonomy using the Taxon ID (17), from UniProtKB using the UniProtKB ID (18), and from NCBI PubMed using the PubMed ID (19). Two fields are also automatically computed within Gephebase: 'Related Genes', which corresponds to the other genes in Gephebase associated with the same phenotypic trait in the same group of species, and 'Related Haplotypes', which displays the other mutations in Gephebase that are found in the same gene and that occurred in other lineage branches in the same group of species.

A single entry can include several traits if a mutation is pleiotropic. Taxon A represents the taxon(s) inferred to bear the ancestral phenotypic state and Taxon B the derived state. If the direction of change cannot be inferred, the field 'Ancestral State' is 'Unknown' and the two compared taxa are assigned arbitrarily to Taxon A and Taxon B. In most cases, Taxon A and Taxon B correspond to taxa at the species level. In cases of named breeds, cultivars, strains or geographically restricted populations, additional information about the Taxon A/B can be found in the field 'Taxon A/B Description'. The phenotypic states are described in 'Trait State in Taxon A/B'.

Exploration tools

Gephebase is designed for interactive exploration and analysis of the genotype–phenotype relationships across species and populations. First-time users can find help on the Frequently Asked Questions page, in tutorials available on the Documentation page and via 'contextual tips', small boxes providing information when the cursor hovers over an item. Data can be queried using boolean operators via the Search page, via SQL line or via custom tools after downloading the dataset of interest as a CSV file. The entire dataset can be downloaded as a CSV file by searching for the wild card * in the top bar panel, clicking on 'Select all' in the top left corner of the results table and then clicking on 'Complete Export'. A Browse page, accessible from the main menu, displays all the Trait names, Species names and also compiles the genes with the highest number of mutations reported in Gephebase.

Gephebase comprises two main views, the View-Entry page which displays a single entry (Figure 3) and the Search/Results page, which shows the results of a given search in a table format (Figure 1). Results of a search are displayed as a table and there are four view options. In the default view, one line of the table corresponds to one Gephebase entry. Under the option 'Split Mutations', one line corresponds to one mutation. Under the option 'Group Haplotypes', one line corresponds to all haplotypes of a given gene for a given pair of Taxon A and Taxon B. Under the option 'Group Genes', one line corresponds to all the genes associated with the same phenotypic trait in the same Taxon A and Taxon B. The number of lines in the Results Table is indicated above the table. Clicking on one line of the Results Table will display all the corresponding Gephebase entries if the line corresponds to several Gephebase entries and will lead to the corresponding View-Entry page if the line corresponds to one entry.

Extensive links to external databases (UniProtKB, NCBI Taxonomy, NCBI PubMed) and to Gephebase itself allow in-depth analysis of curated data. Users can provide feedback using the Feedback section on each View-Entry Page (Figure 3) and can suggest new articles for curation in Gephebase using the 'Suggest an article' button in the top bar menu.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz796/5572566 by arnaud@gwu.edu on 24 September 2019

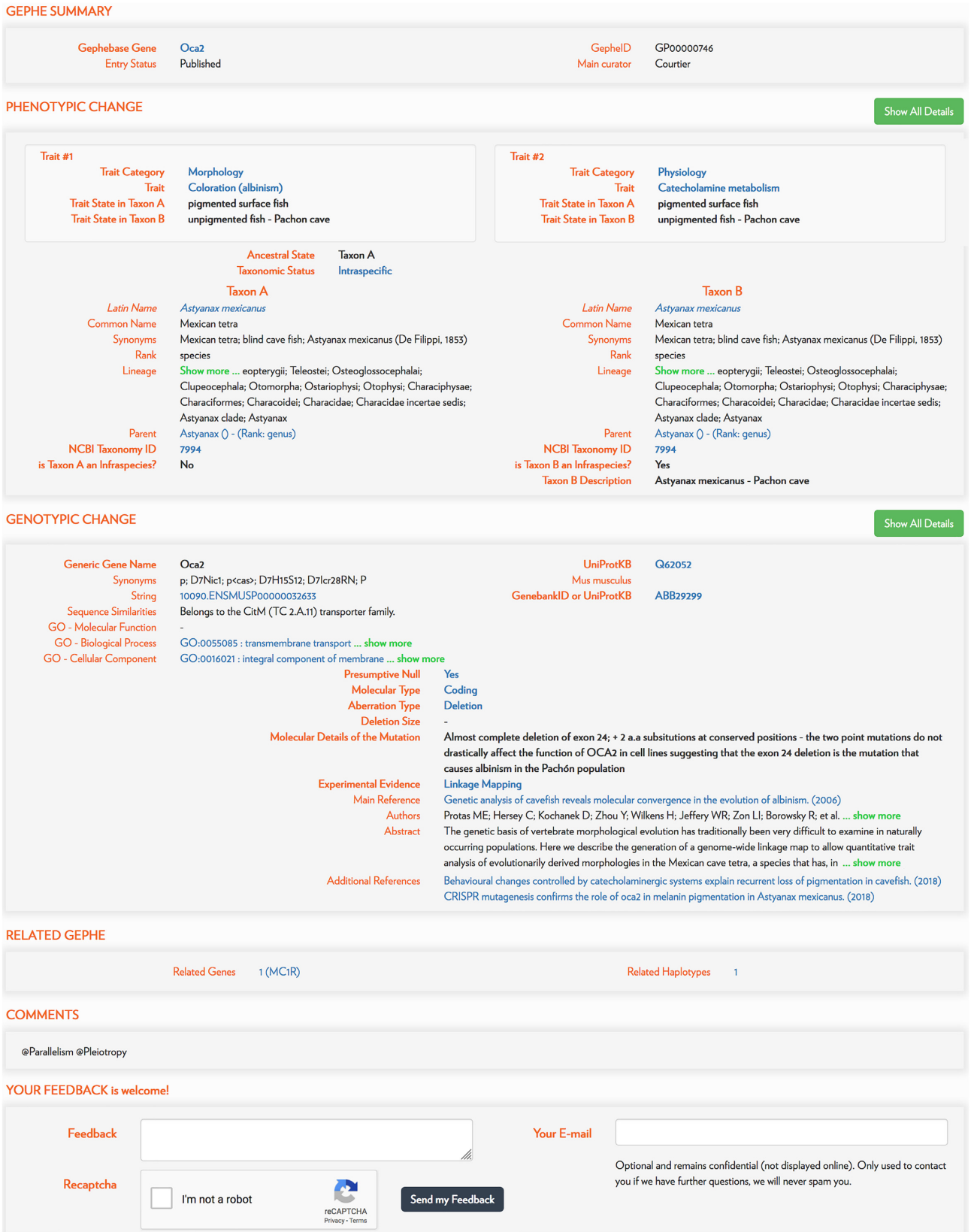


Figure 3. Example of a Gephebase entry. Gephe entries provide extensive information about phenotypic traits, taxon groups, genetic changes, as well as publications.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz796/5572566 by arnaud@gwu.edu on 24 September 2019

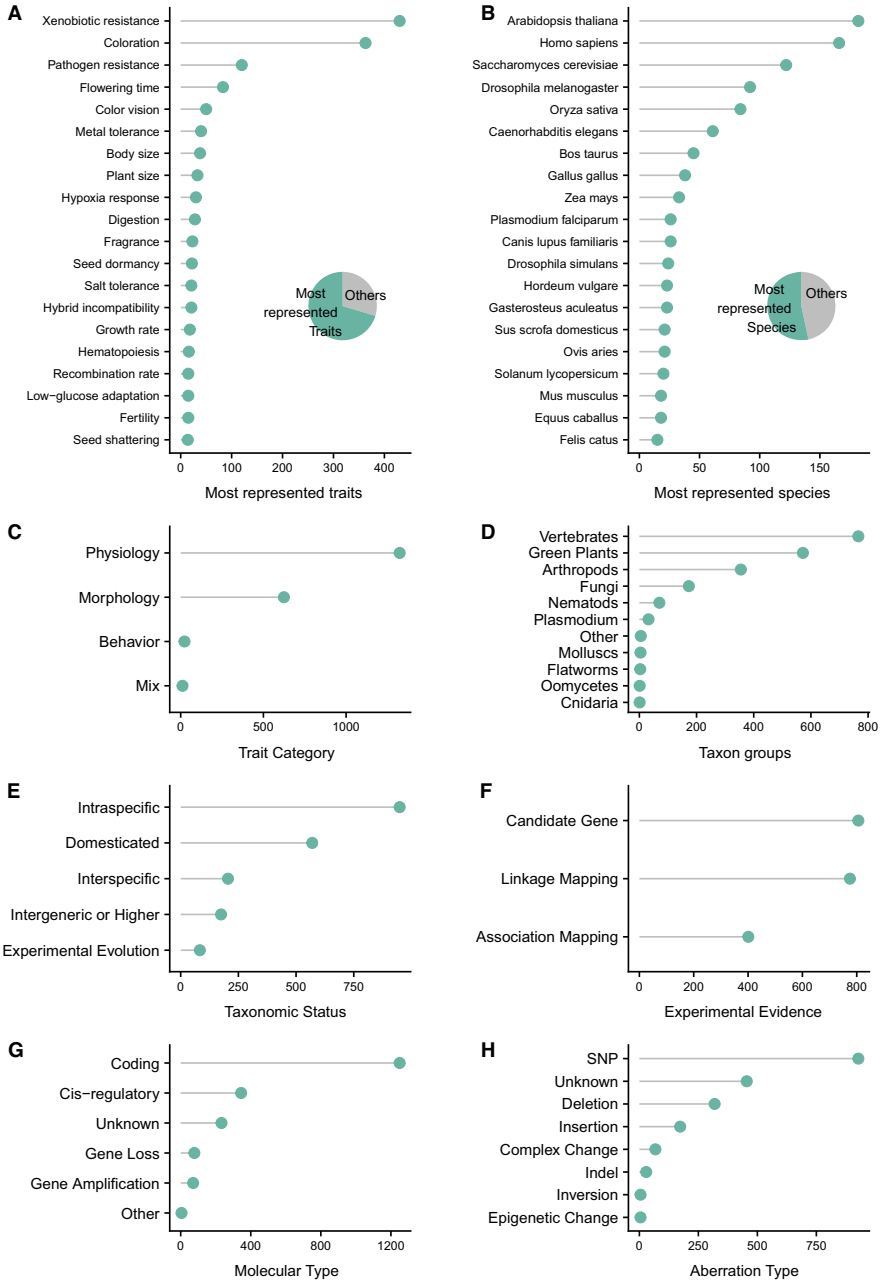


Figure 4. Summary statistics for current Gephebase data. Y-axes indicate the number of entries for each category. (**A**) Distribution of the twenty most represented Traits. (**B**) Distribution of the twenty most represented Taxon B species. Pie charts in (A, B) show the distribution of all entries between the twenty most represented Traits (A)/Species (B) and the remaining others. (**C**) Distribution of Trait Categories (Morphology, Physiology, Behavior, Combination of at least two trait categories). (**D**) Distribution of Taxon Groups. (**E**) Distribution of Taxonomic Status. (**F**) Distribution of Empirical Evidences. (**G**) Distribution of Molecular Types of genetic changes. 'Other' corresponds to mutations creating chimeric genes or to super gene loci. (**H**) Distribution of Aberration Types. 'Indel' corresponds to cases where the direction of change is unknown. When several mutations are reported within one entry, only the first curated mutation of the entry was used for statistical analysis (see Data S1 for the R script).

DISCUSSION AND CONCLUSION

Gephebase contains data for more than 450 eukaryote species and more than 900 distinct genes (Supplementary Table S1, Data S1). In Gephebase, physiological traits represent 67% of the entries, morphological traits 31%, behavioral traits 1% and mixed morphological/physiological/behavioral traits 1% (Figure 4A,C). The most represented traits are Xenobiotic Resistance (21% of the entries) and Coloration (18%). The most represented taxa are Vertebrates (40% of the entries), Green Plants (30%) and Arthropods (17%) (Figure 4D), and a large contribution from traditional model organisms (12). Most data correspond to intraspecific changes (48% of the entries) and domesticated cases (30%) whereas interspecific cases correspond to 10% of the entries (Figure 4E). The three categories of Experimental Evidence are relatively

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz796/5572566 by arnaud@gwu.edu on 24 September 2019

well-distributed among entries (Figure 4F). Gephebase contains a higher number of coding mutations (63% of the entries) compared to cis-regulatory changes (18% of the entries, Figure 4G). While a significant fraction of Gephebase correspond to cases where the exact mutation has not been identified (23% of entries with Aberration Type 'Unknown', Figure 4H), most mapped mutations are single nucleotide changes (47%) and indels (26%).

Gephebase stands out compared to the other current databases of genotype–phenotype relationships in that it compiles genotype–phenotype data across all Eukaryotes. We consider our dataset to be highly complementary to other available databases, which are more species-specific and which usually include more detailed information about genotype–phenotype relationships. Gephebase can be used in various ways: as a powerful bibliographic tool, as a place to formulate hypotheses (Figure 1), as a list of potential targets for breeders interested in transferring traits of interest to new species, as an extensive compilation for broad meta-analyses on the genetic loci of evolution, and also as a resource for epistemologists interested in biases and sociological aspects in the field of genetic evolution. Moving forward, we invite the community of scientists interested in comparative genetics and genotype–phenotype associations to join us in our efforts to curate and synthesize accumulating data.

DATA AVAILABILITY

Gephebase is freely available at gephebase.org. The code is available on GitHub. The entire dataset is freely available for download by searching for '*' and clicking on 'Complete Export'.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the twenty participants of the 'Loci Of Evolution Workshop' (Paris, September 2016) for their enthusiasm and encouragements. We are indebted to the team of AtoutLibre (France) and especially Kyle Ratteree for developing the software and website behind Gephebase. Nathalie Vessilier drew the illustrations featured in Figure 2.

FUNDING

John Templeton Foundation in 2014–2017 [JTF 43903 to A.M. and V.C.]; European Research Council Starting grant ROBUST [FP7/2007–2013 337579 to V.C.-O.]. Funding for open access charge: NSF grant IOS-1923147 to A.M.
Conflict of interest statement. None declared.

REFERENCES

1. Stern,D.L. and Orgogozo,V. (2008) The loci of evolution: how predictable is genetic evolution? *Evolution*, **62**, 2155–2177.
2. Martin,A. and Orgogozo,V. (2013) The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution*, **67**, 1235–1250.
3. Rockman,M.V. (2012) The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution*, **66**, 1–17.
4. Nicholas,F.W. (2003) Online Mendelian Inheritance in Animals (OMIA): a comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals. *Nucleic Acids Res.*, **31**, 275–277.
5. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.
6. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G. and Montoya,M. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
7. Gramates,L.S., Marygold,S.J., dos Santos,G., Urbano,J.-M., Antonazzo,G., Matthews,B.B., Rey,A.J., Tabone,C.J., Crosby,M.A. and Emmert,D.B. (2016) FlyBase at 25: looking to the future. *Nucleic Acids Res.*, **45**, D663–D671.
8. Mungall,C.J., McMurry,J.A., Köhler,S., Balhoff,J.P., Borromeo,C., Brush,M., Carbon,S., Conlin,T., Dunn,N. and Engelstad,M. (2016) The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, **45**, D712–D722.
9. McMurry,J.A., Köhler,S., Washington,N.L., Balhoff,J.P., Borromeo,C., Brush,M., Carbon,S., Conlin,T., Dunn,N. and Engelstad,M. (2016) Navigating the phenotype frontier: the monarch initiative. *Genetics*, **203**, 1491–1495.
10. Stern,D.L. and Orgogozo,V. (2009) Is genetic evolution predictable? *Science*, **323**, 746–751.
11. Streisfeld,M.A. and Rausher,M.D. (2011) Population genetics, pleiotropy, and the preferential fixation of mutations during adaptive evolution. *Evolution*, **65**, 629–642.
12. Arnoult,L.A. (2014) La marche génétique de l'évolution. *Biol. Aujourdhui*, **208**, 237–249.
13. Martin,A. and Courtier-Orgogozo,V. (2017) Morphological evolution repeatedly caused by mutations in signaling ligand genes. In: *Diversity and Evolution of Butterfly Wing Patterns - An Integrative Approach*. Springer, Singapore.
14. Protas,M.E., Hersey,C., Kochanek,D., Zhou,Y., Wilkens,H., Jeffery,W.R., Zon,L.I., Borowsky,R. and Tabin,C.J. (2006) Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nat. Genet.*, **38**, 107.
15. Saenko,S.V., Lamichhaney,S., Barrio,A.M., Rafati,N., Andersson,L. and Milinkovitch,M.C. (2015) Amelanism in the corn snake is associated with the insertion of an LTR-retrotransposon in the OCA2 gene. *Sci. Rep.*, **5**, 17118.
16. Sakane,Y., Iida,M., Hasebe,T., Fujii,S., Buchholz,D.R., Ishizuya-Oka,A., Yamamoto,T. and Ken-ichi,T.S. (2018) Functional analysis of thyroid hormone receptor beta in Xenopus tropicalis founders using CRISPR-Cas. *Biol. Open.*, **7**, bio030338.
17. Federhen,S. (2011) The NCBI taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
18. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M., Bansal,P., Bridge,A.J., Poux,S., Bougueleret,L. and Xenarios,I. (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. In: *Plant Bioinformatics*. Springer, pp. 23–54.
19. Geer,L.Y., Marchler-Bauer,A., Geer,R.C., Han,L., He,J., He,S., Liu,C., Shi,W. and Bryant,S.H. (2009) The NCBI biosystems database. *Nucleic Acids Res.*, **38**, D492–D496.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkz796/5572566 by arnaud@gwu.edu on 24 September 2019