Gephebase practicals

October 2022 (Arnaud Martin & Virginie Courtier-Orgogozo)
See detailed Documentation at the end of this document.

www.gephebase.org

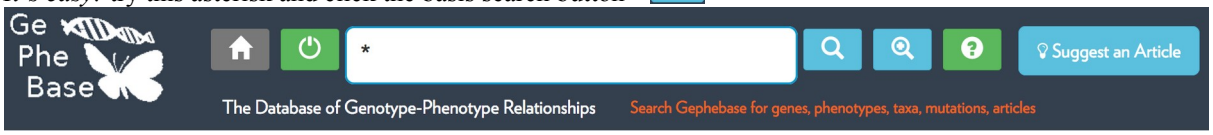
Gephebase compiles **genotype-phenotype relationships** under the form of associations between a genetic locus (a mutation, gene, or narrow genetic region) and a case of phenotypic variation that evolved naturally or that was domesticated. Gephebase consolidates data from the scientific literature about the genes and the mutations responsible for phenotypic variation in Eukaryotes (mostly animals, yeasts and plants). For now, genes responsible for human disease and for aberrant mutant phenotypes in laboratory model organisms are excluded and can be found in other databases (OMIM, OMIA, FlyBase, etc.). QTL mapping studies that did not identify single genes are not included in Gephebase.

Gephebase is accessible to all, no login is required. You can search online and view all the data contained in Gephebase. You can also import the data you want as a csv file.

Go with the flow and answer the questions in bold. You are more than welcome to explore with your own ideas and examples.

GETTING STARTED: Show me ALL THE DATA

It's easy: try this asterisk and click the basis search button 



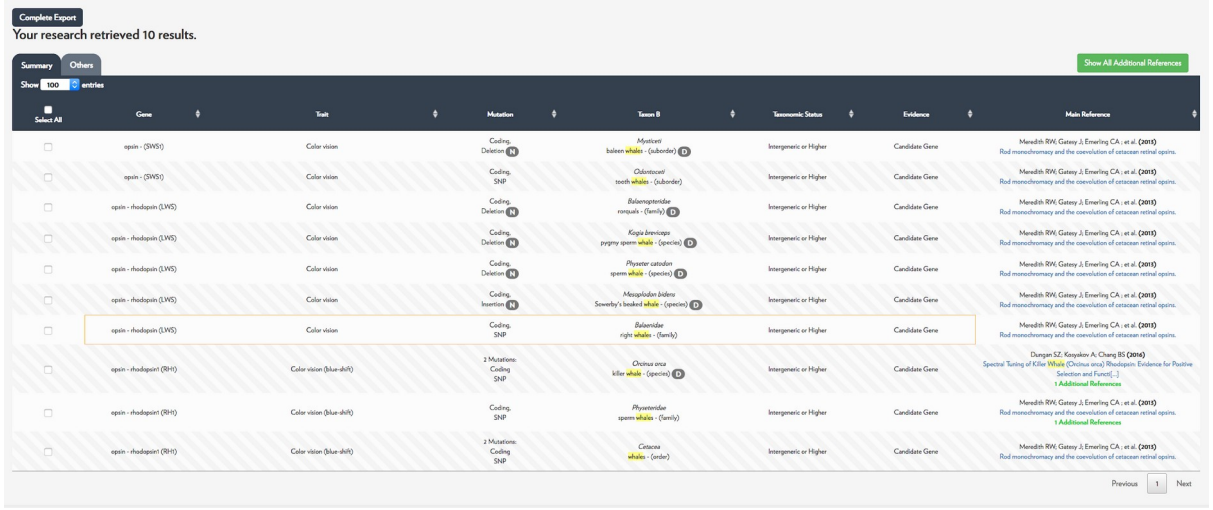
You can save the entire database as a csv file by clicking on "Complete Export".

1) How many entries are there in total now in Gephebase ?

Is there any data on my favorite organism? Same process, but try a keyword. Use the "Esc" Key to ignore auto-completion suggestions, which could narrow your search more than you intend.

Pick an organism you can think of, and check what you find after performing a basic search.
(note: please avoid the most obvious like "human" and "dog"… think broadly, perhaps a mix of plants and animals?

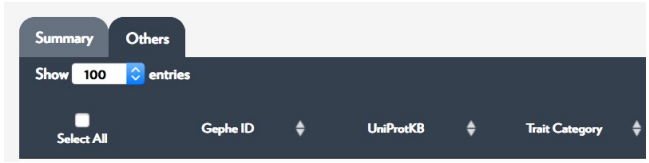
Check the results of the search for "whale":



2) What is the proportion of entries involving opsin genes?
3) What trait is associated to opsin genes?
4) Are they coding changes, cis-regulatory changes, or other types of changes?
5) How many entries involving opsin genes are in the database? (hint: less than 60)
6) And in what other animals besides cetaceans, birds and fishes? (cichlids are fishes)
7) Are they coding changes or cis-regulatory changes?

8) For all the opsin results, what does the "Evidence" column says? This column reflects how scientists were able to identify mutations in that genes that caused a given phenotypic difference in a trait.
9) Can you explain why the approach used as "Evidence" for Opsin genes is biased?

From a "killer whale" search, Click on the Tab "Others".



The UniprotKB hyperlink brings you to a unified resource for protein annotation. Click on P08100 to go to the UniprotKB entry. Gephebase curators use a well annotated gene from UniProtKB that is not necessarily from the species of interest.

10) Report the recommended protein name (full name and abbreviation).

11) Does this opsin mediate vision in bright light?

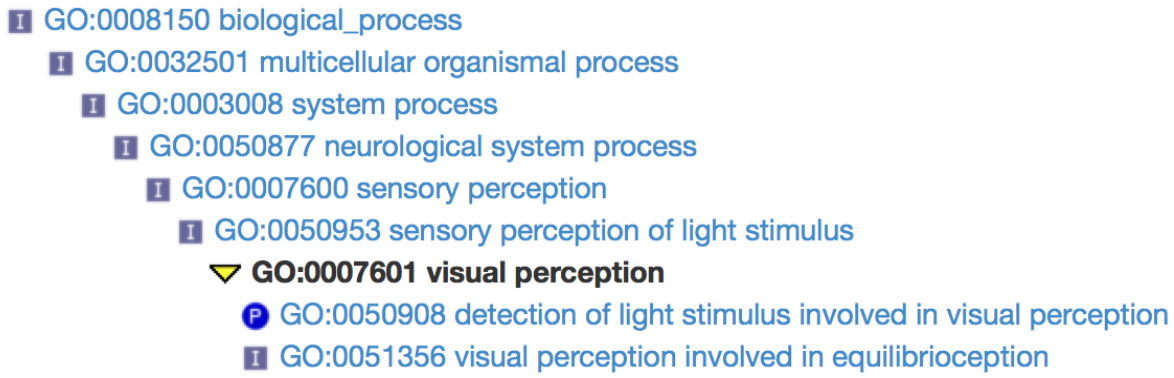
12) In the UniprotKB "Function" section, what does Keywords say?

This shall be a good summary of the Gene Ontology data for that gene, a very useful "ontology" with controlled vocabulary for analyzing big data in genetics and molecular biology.

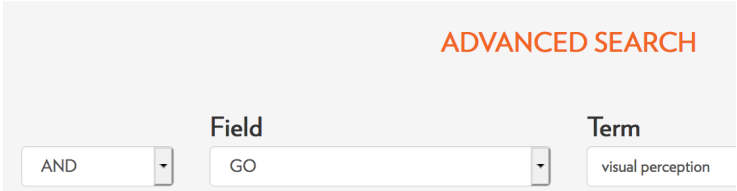
ADVANCED SEARCHES: click on the blue button with the plus sign, in the top panel. 

It is possible to use **Gene Ontology categories "GO"** to restrict a search to specific processes or molecular functions

Using AmiGO (http://amigo.geneontology.org/amigo), or by accessing a known Gephebase entry (*e.g.* one of the photoreceptors identified from a "whale" search), identify a GO category of interest (AmiGO output shown here)

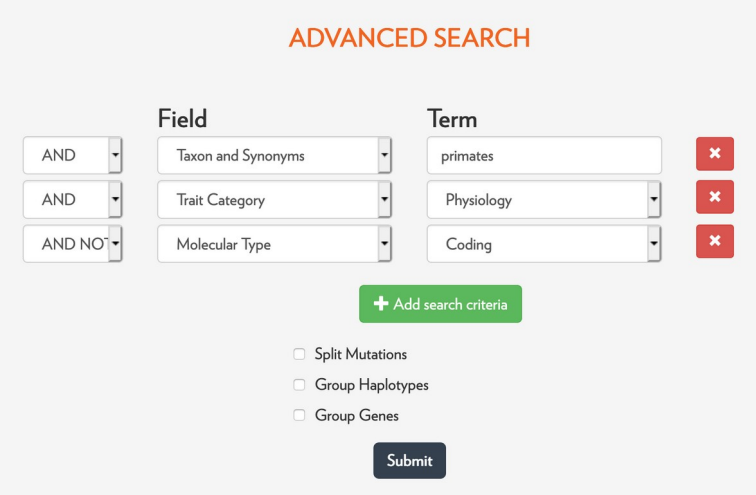


Accordingly, this search in Gephebase will return the entries involving a photoreceptor, as well as a few other entries:



Note that in the present version Gephebase only searches for the GO terms that have been directly attributed to the genes, and not for the GO terms of higher hierarchy.
Let's try something else…

Using the advanced search, you will look for the mutations that are **not coding** (cis-regulatory and others as well) and that have been associated with **physiological changes** in **primates**.



13) How many results do you get?

Click on the top of the Column "Trait" to sort the results alphabetically by trait.
Now let's explore two interesting cases of RECENT HUMAN EVOLUTION: the adaptation of human populations to the Neolithic diet, and then, the resistance to bad diseases like HIV and the malaria parasite.

14) A single gene is associated with lactose tolerance. Which gene is it?

15) How many mutations have independently evolved?

Go back to the not coding / physiology / primates results.
16) How many genes are associated with Body Fat traits?

These entries suggest that we are not genetically equal when it comes to the tendency to accumulate fat, although of course, the diet type and calorie intake are the most important variables. Understand that the phenotypic expression of these mutations is context-dependent. Here, they have been identified in a context where the individuals are most likely under a typical western diet (high in fat and high carbohydrates). The outcome may be different with different diets. Also mutations that favor fat deposition may have been selected by strong selection in the past in nomadic populations that had to travel long distance and undergo some episodes of starvation.We should absolutely celebrate the fact that the human gene pool is *diverse*, variable.

17) How many genes are associated to resistance to malaria in this search result?
18) What is the exact Trait name in Gephebase?
19) Why do you get more results if you do a basic search with this trait name?
20) Why do you get even more results if you do a basic search for "malaria"?

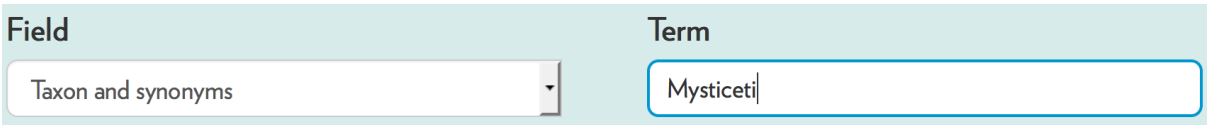
More Examples of Advanced Searches

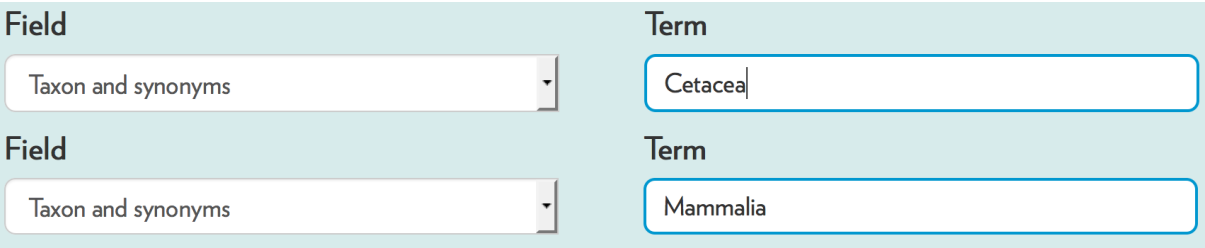
Let's imagine that while dolphins and orcas are cool, you are specifically interested in "toothless whales". How can you be sure to find them all?

If you access the Preview of a rhodopsin entry obtained with a Basic search "whale", notice the Taxon section:

Taxon B	
NCBI Taxonomy ID	30558
Name	*Balaenidae (right whales) - (Rank: family)*
Description	*Balaenidae (bowhead and right whale)*
Lineage	cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Laurasiatheria; Cetartiodactyla; Cetacea; Mysticeti

In Gephebase, we use the Controlled Vocabulary of **NCBI Taxonomy** to standardize taxon names as well as their taxonomic lineage. Using this information you can now perform an Advanced Search to look for all the Gephebase entries that include a given Taxonomic Lineage: Mysticeti (*ie*. toothless whales only) ; Cetacea (will include dolphins, sperm whales, orcas…) ; or even Mammals…





In this world of controlled vocabulary, you will have to find the identifiers that make the most sense for your taxonomic group of interest.

For instance, if you want to include or exclude all "green plants", use the NCBI Taxonomy term **"Viridiplantae"**

Use Wikipedia, NCBI Taxonomy (**https://www.ncbi.nlm.nih.gov/Taxonomy**), or OneZoom (**https://www.onezoom.org/** highly recommended !) to navigate the tree of life.

21) What taxon name would you use for making a search limited to butterflies and moths?
22) to flowering plants?
23) to New World Monkeys?

24) Using AND + ANDNOT in Advanced Search, report the number of results (you will have to find the right taxon name like above) for all vertebrates except mammals.
25) for all reptiles except birds.
26) for all plants except Monocots (which contain most cereals).

Let's imagine you are interested in Intraspecific differences

Cases of **Natural Evolution** represent Genotype-Phenotype relationships that have been identified between natural organisms. They can underlie a difference between individuals/populations, a difference between species, or a difference between deeper phylogenetic levels.. These cases are encoded as "Intraspecific", "Interspecific", or "Intergeneric or higher".

Gephebase also includes two additional types of evolutionary processes:

- **"Domesticated"**: in short, the phenotypic difference has been selected by human breeders *for an other purpose than the scientific study of genetics*. *eg.* Dog phenotypes, cereals, industrial yeast strains…
- **"Experimental Evolution": both the phenotypic difference and the underlying mutation(s) arose during an experimental evolution experiment** with a **controlled selective pressure** as well as a known, and **unique ancestral genotype** (if a given experiment starts from a pool of standing genetic variation, this variation must be considered "Intraspecific" or "Domesticated")

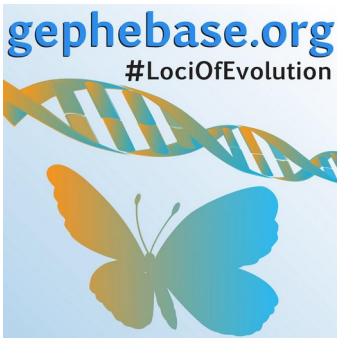
Use "Taxonomic Status" to limit your search to, or exclude cases of "Domestication" and "Experimental Evolution":

Using an Advanced Search, find all the entries related to
Trait = Coloration AND Taxonomic Status = Intraspecific AND Taxon and synonyms = Vertebrata
(notice you would get a lot of flower and insect coloration entries without that last term).

27) How many entries did you get?
28) Which gene is the most represented?
29) What is the experimental evidence for most of these cases?

Three options are possible for the Advanced Search: “Split Mutations”, “Group Haplotypes”, “Group Genes”.
30) What do these options do?

Detailed documentation



An entry in Gephebase contains the following fields:

- **Gephe ID** : identification number of the entry (specific to Gephebase, starts with the two letters "GP" and then contains 8 digits: GP00000001 is entry #1 in Gephebase)
- **Gene-Gephebase**: common general name of the gene (nomenclature specific to Gephebase)
- **Ortholog-Gephebase**: common general name of the gene family (classification specific to Gephebase)
- **UniProtKB ID**: UniProtKB identifier of the protein, for a given Ortholog group, and in a species with the optimal annotation level. This means that if the protein was poorly annotated in the Taxon A/B fields (see below), the ortholog version of a relevant, more "traditional" model organism was chosen in order to provide additional molecular annotations. Typically: *Homo* or *Mus* for mammals, *Gallus* or *Xenopus* for other tetrapods, *Danio* for non-tetrapod vertebrates, *Drosophila melanogaster* for insects, *Arabidopsis* for dicotyledons, *Zea* and *Oryza* for monocotyledons, *Saccharomyces cerevisiae* for yeasts. If several UniProtKB ID are available for a given protein, then only one of them is shown. In a few cases (mutation in a non coding RNA gene for example), this field is empty. This UniProtKB ID is used by Gephebase to capture gene names, synonyms, and GO terms automatically.
- **Gene Name**: common gene name based on Orthology (as in UniProtKB)
- **Synonyms**: all the gene synonyms (as in UniProtKB, plus sometimes a few other synonyms added)
- **GO – Molecular Function**: GO terms associated with the gene (https://www.ebi.ac.uk/QuickGO/LociOfEvolution Workshop Handbook.doc) *(as in UniProtKB)*
- **GO – Biological process**: GO terms associated with the gene (https://www.ebi.ac.uk/QuickGO/) *(as in UniProtKB)*
- **GO – Cellular component**: GO terms associated with the gene (https://www.ebi.ac.uk/QuickGO/) *(as in UniProtKB)*
- **Genbank ID** : Genbank identifier of the protein or nucleotide sequence in the species with the derived phenotypic trait. If several Genbank ID are available for a given sequence, then only one of them is shown. This Genbank ID is currently not used by Gephebase. In theory, it could be used to obtain information about transcript size, position along the chromosome, etc.
- **Trait**: small text explaining the phenotype. No controlled vocabulary. This data is entered manually by Gephebase curators.
- **Trait Category**: can be either Morphology, Physiology, or Behavior, or any combination of the three.
- **Ancestral State**: indicates the direction of the evolutionary change. Three exclusive possibilities: Data Not Curated, Unknown, Taxon A. "Taxon A" means that the ancestral trait is held by the taxon described in the field "Taxon A" and that the derived trait is held by "Taxon B".
- **Taxon A > NCBI Taxonomy ID**: NCBI Taxonomy ID for the taxon exhibiting the ancestral trait (when the direction of change is known) or one of the two alternative phenotypic traits (when the direction of change is unknown)
- **Taxon A > Name**: main taxon name, followed by common taxon names in parentheses, followed by the rank (order, species, etc.) in parentheses. The syntax is as follows: *Panthera tigris (tiger) - (Rank: species)* (data collected in NCBI Taxonomy using the NCBI Taxonomy ID)
- **Taxon A > Description**: free text describing the taxon and its phenotype of interest in more details. Ideally, the beginning of this text should contain the main taxon name. This data is entered manually by Gephebase curators.
- **Taxon A > Parent**: the smallest taxon of higher rank containing taxon A (data collected in NCBI Taxonomy using the NCBI Taxonomy ID)
- **Taxon A> Lineage**: list of all the taxons of higher rank containing taxon A (data collected in NCBI Taxonomy using the NCBI Taxonomy ID). For example: *cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Homo*
- **Taxon B > Name**: same as for Taxon A but for the taxon displaying the derived phenotypic trait
- **Taxon B > Description**: same as for Taxon A but for the taxon displaying the derived phenotypic trait.

- **Taxon B> Lineage**: same as for Taxon A but for the taxon displaying the derived phenotypic trait
- **Taxonomic status**: 5 exclusive possibilities: Domesticated, Experimental Evolution, Intergeneric or higher, Interspecific, Intraspecific. This choice is made by Gephebase curators.
- **Experimental Evidence**: 3 exclusive possibilities: **Association Mapping, Linkage Mapping, Candidate Gene**. This choice is made by Gephebase curators based on *the best evidence available* for a given genotype-phenotype relationship. Gene-to-phenotype identified by Linkage Mapping with resolutions below 500kb have priority in our dataset (see Supplementary materials in Martin and Orgogozo 2013). Association Mapping studies are included based on individual judgment, with a strong bias towards SNP-to-phenotype associations that have been confirmed in reverse genetics studies. In other words, Gephebase intends to be more stringent than a compilation of statistically-significant SNPs, and attempts to select studies where a given genotype-phenotype association is relatively well supported or understood.

- **Molecular details of the mutation**: free text entered manually by Gephebase curators
- **Mutation Type**: 6 exclusive possibilities: cis-regulatory, coding, structural (amplification), structural (loss), structural (rearrangement), unknown. This choice is made by Gephebase curators.
- **SNP coding change**: 5 exclusive possibilities: nonsynonymous, nonsense, synonymous, unknown, not curated. This choice is made by Gephebase curators.
- **Aberration Type**: 10 exclusive possibilities: SNP, insertion, deletion, indel, inversion, translocation, complex change, epigenetic change, unknown. This choice is made by Gephebase curators.
- **Main reference**: PubMed ID corresponding to the main publication describing the association between the genetic change and the phenotypic change.
- **Main reference > Title**: title of the publication (data collected in NCBI PubMed using the NCBI PubMed ID)
- **Main reference > Year**: year of the publication (data collected in NCBI PubMed using the NCBI PubMed ID)
- **Main reference > Type of publication**: type of the publication: journal article, review, etc. (data collected in NCBI PubMed using the NCBI PubMed ID). Only one value is imported in Gephebase if several values are present for the given reference in NCBI PubMed.
- **Main reference > Authors**: author names of the publication (data collected in NCBI PubMed using the NCBI PubMed ID).
- **Main reference > Journal name**: name of the journal in which the paper is published (data collected in NCBI PubMed using the NCBI PubMed ID).
- **Main reference > Journal abreviation**: abreviation of the journal in which the paper is published (data collected in NCBI PubMed using the NCBI PubMed ID).
- **Main reference > Volume**: volume number of the paper (data collected in NCBI PubMed using the NCBI PubMed ID).
- **Main reference > Pagination**: page numbers of the paper (data collected in NCBI PubMed using the NCBI PubMed ID).
- **Main reference > Abstract**: abstract of the paper (data collected in NCBI PubMed using the NCBI PubMed ID).
- **Main reference > DOI**: doi of the paper (data collected in NCBI PubMed using the NCBI PubMed ID).
- **Main reference > URL**: relevant website link for the paper (this field is filled up only when the reference is not present in NCBI PubMed and when the reference data is imported via a .ris file into Gephebase).
- **Additional References**: PubMed ID corresponding to the other publications describing the association between the genetic change and the phenotypic change. Gephebase automatically imports all the relevant fields of these other references as for the main reference.
- **Comments**: free text entered by Gephebase curators.