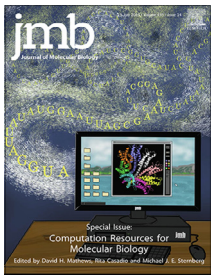

EvolClustDB: Exploring Eukaryotic Gene Clusters with Evolutionarily Conserved Genomic Neighbourhoods

Marina Marcet-Houben [1,2], Ismael Collado-Cala [1,2], Diego Fuentes-Palacios [1,2], Alicia D. Gómez [1,2], Manuel Molina [1,2], Andrés Garisoain-Zafra [1,2], Uciel Chorostecki [1,2] and Toni Gabaldón [1,2,3,4]*

*1 - **Institute for Research in Biomedicine (IRB Barcelona),** The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain*

*2 - **Barcelona Supercomputing Centre (BSC-CNS). Plaça Eusebi Güell,** 1-3, 08034 Barcelona, Spain*

*3 - **Catalan Institution for Research and Advanced Studies (ICREA),** Barcelona, Spain*

*4 - **Centro de Investigación Biomédica En Red de Enfermedades Infecciosas (CIBERINFEC),** Barcelona, Spain*

Correspondence to Toni Gabaldón: *Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain.*toni.gabaldon@bsc.es* *(T. Gabaldón)*@gabaldonlab*, *@Toni_Gabaldon* *(T. Gabaldón)*

https://doi.org/10.1016/j.jmb.2023.168013

Edited by David Mathews

Abstract

Conservation of gene neighbourhood over evolutionary distances is generally indicative of shared regulation or functional association among genes. This concept has been broadly exploited in prokaryotes but its use on eukaryotic genomes has been limited to specific functional classes, such as biosynthetic gene clusters. We here used an evolutionary-based gene cluster discovery algorithm (EvolClust) to pre-compute evolutionarily conserved gene neighbourhoods, which can be searched, browsed and downloaded in EvolClustDB. We inferred ~35,000 cluster families in 882 different species in genome comparisons of five taxonomically broad clades: Fungi, Plants, Metazoans, Insects and Protists. EvolClustDB allows browsing through the cluster families, as well as searching by protein, species, identifier or sequence. Visualization allows inspecting gene order per species in a phylogenetic context, so that relevant evolutionary events such as gain, loss or transfer, can be inferred. EvolClustDB is freely available, without registration, at http://evolclustdb.org/.

Introduction

Gene order tends to be poorly conserved across eukaryotic species.[1] Despite this general trend, some groups of genes tend to remain close to each other, even across long evolutionary distances. These genes are likely to be co-regulated and be related to each other either functionally or structurally.[2] Gene clustering has important biological implications as it affects co-regulation, recombination, or the sharing of epigenetic marks.[3–6] Illustrative examples of conserved gene clusters (GCs) are those encoding genes involved in secondary metabolite production, which are conserved across distantly related fungal species,[2] the galactose assimilation (GAL) cluster in budding yeasts,[7,8] or the major histocompatibility complex in mammals.[9] Several studies have performed comprehensive analyses of gene order conservation in eukaryotes.[10,11] However, few databases exist that compile groups of genes that share a conserved genomic neighbourhood across different species,

and most existing ones focus on a determined function (i.e. secondary metabolism GC databases such as **BiG-FAM**.[12]). Some existing databases are broad in scope, but provide information about operons found in prokaryotic genomes[13] or show gene order conservation surrounding a gene of interest,[14] and none provide pre-computed lists of evolutionarily conserved eukaryotic GCs.

The EvolClust algorithm (https://github.com/Gabaldonlab/EvolClust)[15] searches for groups of neighboring genes that are located close together in two or more genomes and whose gene neighbourhood conservation is higher than the observed average of the compared genomes. These genes are considered to constitute a GC when they comprise more than four different gene families and the cluster genes are not separated by more than three intervening non-cluster genes. GCs are limited in size between 5 and 35 genes and can contain duplicated genes. The lower limit of 5 may omit relevant GCs such as the Penicillin GC that is formed only of three genes, but given the flexibility EvolClust allows in terms of number of non-cluster genes and duplications many meaningless GCs would be included if this was relaxed. Additionally, note that EvolClust does not filter by function and, hence, the detected clusters are not limited by functional association. When run on a set of proteomes, EvolClust first finds pairwise groups of genes with a conserved neighbourhood, calculates a conservation score and then compares it to the average conservation score of all possible conserved regions between the two species. All groups of genes that have a conservation score above the pre-calculated threshold are considered GCs. Detected GCs are then grouped into families according to conservation scores and form what we refer to as a cluster family (CF): a set of GCs across two or more species (see Figure 1). EvolClust was successfully applied on a group of 341 fungal species and a group of 145 insects and benchmarked against three other programs.[15] Additionally, the fungal dataset was benchmarked against a known set of secondary metabolism GCs.

Running of EvolClust is time-consuming and requires a local proteome database and computational resources that may not be available by researchers interested in discovering new CFs. To allow access to automatically predicted CFs we pre-computed CFs for five relevant eukaryotic groups: fungi, insects, metazoans, plants, and protists, which we made available in EvolClustDB (https://evolclustdb.org/).

EvolClustDB is a database designed to store groups of CFs inferred from sets of proteomes from a given taxonomic category inferred using the EvolClust algorithm.[11,15] EvolClustDB allows users to explore conserved groups of genes regardless of their functional annotation. Users can identify their gene of interest using protein or gene codes or using a sequence similarity search. This last type of search allows users to find out whether a single sequence can be found in a pre-calculated cluster or whether a group of genes are found in a cluster. Additionally, users can select a species of interest and find all its predicted clusters. Results are shown in a table or can be visualized in an image of the CFs next to a species tree. The image is interactive and allows users to access the available information regarding the gene description as stored in the database.

Results

Cluster families

We used EvolClust v.1.0 to detect conserved CFs, groups of genes with a conserved genomic neighbourhood across two or more species (see Figure 1). We used five sets of proteomes: fungi,

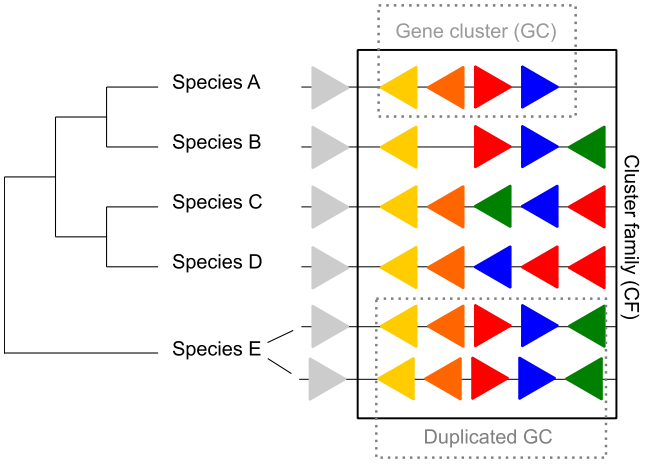

Figure 1. Graphical explanation of the terms used throughout the study. The image mimics images as shown by EvolClustDB and points out what is considered to be a cluster family as inferred through gene neighborhood, a gene cluster and a duplicated gene cluster.

insects, metazoa, plants and protists. In total 35,777 CFs were detected including 242,482 individual GCs and spanning 882 species (see Table 1). The largest number of CFs was found in the fungal dataset (12,120 CFs), which also comprises the largest number of species (341 species). On the other hand, the protists dataset contained the fewest number of CFs (3,971 CFs, 156 species) despite being of similar size as the insects dataset (8,778 CFs, 145 species). This difference is likely due to the broader evolutionary distances in the protists as compared to insects. If we consider the average number of CFs per species, plants have the largest set with 61 CFs per species, closely followed by insects with 60. As expected, protists have fewer CFs per species, containing only 25 on average. In terms of average number of GCs in each CF, the dataset of metazoans tends to have the largest CFs (11 GCs on average per CF), closely followed by fungi (10 GCs on average per CF).

EvolClust allows for the detection of CF that have GCs that contain gene duplications, gene losses and changes in the internal order of genes. When considering gene duplications within a GC, we observed that GCs in protists show the largest number of gene duplications, with more than half of the GCs containing at least one duplicated gene (52%). This was followed by the metazoans dataset (41%), whereas fungi had the lowest fraction of GCs with duplicated genes (10%). On the other hand, whole GCs may also be duplicated within a species (see Figure 1). When considering this we found that plants display the largest number of duplicated GCs within a CF (17% of the CFs contain at least a duplicated GC) closely followed by protists (15%). On the other extreme, insects have the least duplicated GC within a CF (1.4%).

Due to intra-cluster genome rearrangements, the specific gene order within GCs in a CF is not necessarily well conserved. Variations in gene order make the comparative visualization of GCs from different species challenging. EvolClustDB offers a graphical representation of CFs next to a species tree, and this representation can show how GCs vary within the CF. One example of this is CF_003233, corresponding to the secondary metabolism GC responsible for the production of the fungal toxin Patulin.[16,17] Two species contain the complete GC: *Penicillium expansum* and *Aspergillus clavatus*, whereas the remaining seven species contain only subsets of the CF. As shown in the image of CF_003233 (https://evolclustdb.org/cluster/search/CF_003233) it is not possible to perfectly align homologous genes in the different GCs.

Table 1 Main statistics of the different datasets stored in EvolClustDB.

Dataset	Number of species	Number of clusters	Number of cluster families
Fungi	341	118,699	12,120
Insects	145	28,116	8,778
Metazoans	136	49,989	4,522
Plants	104	24,826	6,386
Protists	156	20,752	3,971

Front and backend implementation

EvoclustDB is currently hosted at the Barcelona Supercomputing Center (BSC-CNS). It runs on a Virtual Machine with Linux Ubuntu (release 20.04) where most tasks related to database and web interface operations are carried out. To avoid virtual machine memory overload, some of the more demanding operations are run asynchronously in the MareNostrum supercomputer, ie: similarity search. The web service APIs were deployed in a Gunicorn web server (https://gunicorn.org/) and the scripts used in data processing were written in Python (version 3.7.7). The database is stored and managed in MariaDB (version 10.2.22). The front-end uses the PHP Codeigniter framework (version 4.2.4 - https://codeigniter.com/), jQuery (version 3.6.0), Apache (version 2.4.25), PHP (version 7.1.25) and PhyD3 (version 3.5.17)[18] for the visualization of interactive trees. The EvolClustDB database is freely available at https://evolclustdb.org/.

EvolClustDB website

EvolClustDB has one main interface in which data on a given CF is displayed. To access it, the user has to perform one of the searches available in EvolClustDB (see below) resulting in a list of putative CFs. Clicking on the CF code will lead to the main CF page. Two tabs are then available, the first one contains general information about the CF and the second one contains specific information about the CF. The first element is a table that lists the species that encode this CF, followed by the list of genes included in each individual GC. Duplicated GC within the CF are numbered sequentially in the table. Each protein code found in the table is hyperlinked to the protein information page where information about the protein is shown. This includes the protein code, links to NCBI and UniProt databases, information about the species which encodes the protein and its taxonomic lineage and the description provided in the original NCBI files. Data from the table can be downloaded in different formats using the buttons at the top of the table.

The two other elements in the main CF page are a link to the CF image and a heatmap showing the conservation scores across GCs within the CF. The CF image contains a species tree (left) which is a dendrogram representing the evolutionary

relationships between all the species that contain the CF of interest. At the right of the tree the genes that are part of the CF are depicted. Each gene is represented by a triangle coloured based on the gene family. White triangles represent genes from gene families that only appear once, and therefore they are not homologous. Whenever possible, the GCs within the CFs are aligned to show whether gene order is conserved. The image is interactive. Scrolling the mouse over the genes will show the description associated with the gene. The user can click on the leaf in the tree or on any of the genes and this will show information specifically for the GC belonging to this species. The information is shown in a small window where the species name appears followed by the list of codes for the proteins found in the GC. This list does not only include the protein code, but also its description and a number indicating the gene family the gene belongs to. Note that gene family codes are not conserved across datasets. When the CF is found in many different species the tree may be very large. The PhyD3 library used to generate the images deals with large trees by reducing the number of shown leaves, therefore the user can use the mouse to scroll in and out of the tree for more species to appear. GCs that are duplicated inside a CF are not shown, instead one of the GCs is placed in the image and the information is marked on the tree with the symbol ***. Note that information about duplicated GCs is available in the table.

To access the main CF page, EvolClustDB can be browsed in different ways. The simplest way is to search for a gene or protein name as found in NCBI. This leads to the main CF page explained above. Searches can also be done by species name, which will show all CFs detected in a given organism. The autocomplete feature incorporated in all searches helps identify whether the species of interest is present in the database or not. The search will lead to the main species page where information about the taxonomic lineage of the species is presented. As before there is a second tab called Cluster Families that leads to the list of all CFs found in the species of interest. Finally, there is also a direct cluster search in case a user wants to recover a CF they were previously working with. This search needs the internal CF code provided in the other two searches in EvolClustDB.

In addition to the keyword search a similarity search has been implemented. One or multiple sequences can be input in fasta format and a blast search will be performed against the complete sequence database of proteins found in at least one of the CFs. When one single sequence is provided, a blast search will be performed and a list of proteins with the best hits will be returned with the associated CFs. When multiple sequences are provided, a search is performed using a blast search using each of the proteins as a query. Then the results are joined to see whether there are multiple homologs within a single GC. The user is able to determine with the N-hits parameter how many homologs there should be in a GC for it to be a meaningful result. Only non-redundant CFs will be reported by this search.

The presence of EvolclustDB CFs in a proteome of interest that is not included in EvolClustDB, can be explored. For this, EvolClustDB provides for each CF a set of HMM profiles for each of its gene families. Using HMMs of the CF of interest, its presence on any proteome of interest can be identified using the HMMsearch as implemented in HMMER3,[19] which will provide a list of proteins identified with the given profile, and the information of gene order for the genome of interest (details on how to detect CFs in new proteomes can be found here: https://github.com/Gabaldonlab/EvolClust/wiki/7.-Profile-search).

A working example

The Frataxin gene encodes a protein targeted to the mitochondrion, where it enables several functions among which are ferroxidase activity and iron ion binding activity.[20] In addition it participates in several other processes, including negative regulation of release of cytochrome c from mitochondria and positive regulation of succinate dehydrogenase activity.

Interestingly, *C. elegans* has part of its genome (15%) in the form of polycistronic units or eukaryotic operons.[21–23] Among 881 *C. elegans* GCs detected in a previous study[21] one contains the Frataxin gene. While most of those operons include few genes (average of 2.3 genes per operon),[22] which would be undetectable for EvolClust, the GC containing Frataxin is formed by seven genes.[21] This is large enough for Evolcust to detect the GC as long as it is shared by at least one other species. As EvolClustDB contains *C. elegans* plus 11 additional nematode species, we set out to search whether the Frataxin operon found in *C. elegans* was also conserved in the remaining nematode species. First we used the keyword search to directly search for the Frataxin gene (CELE_F59G1.7). This search returned a CF (CM_002844) which was present in two additional nematode species (*Angiostrongylus costaricensis* and *Dictyocaulus viviparus*) which belong to a different taxonomic order (Figure 2). The GC in *C. elegans* as predicted by EvolClustDB is formed by six genes of which five are thought to be co-regulated.[21] The gene order in the other two nematode species is highly conserved.

We then used the similarity search to see whether other nematode species had a different CF containing Frataxin. A second CF was found in the two closely related nematode species *Brugia malayi* and *Thelazia callipaeda* (CM_004270)

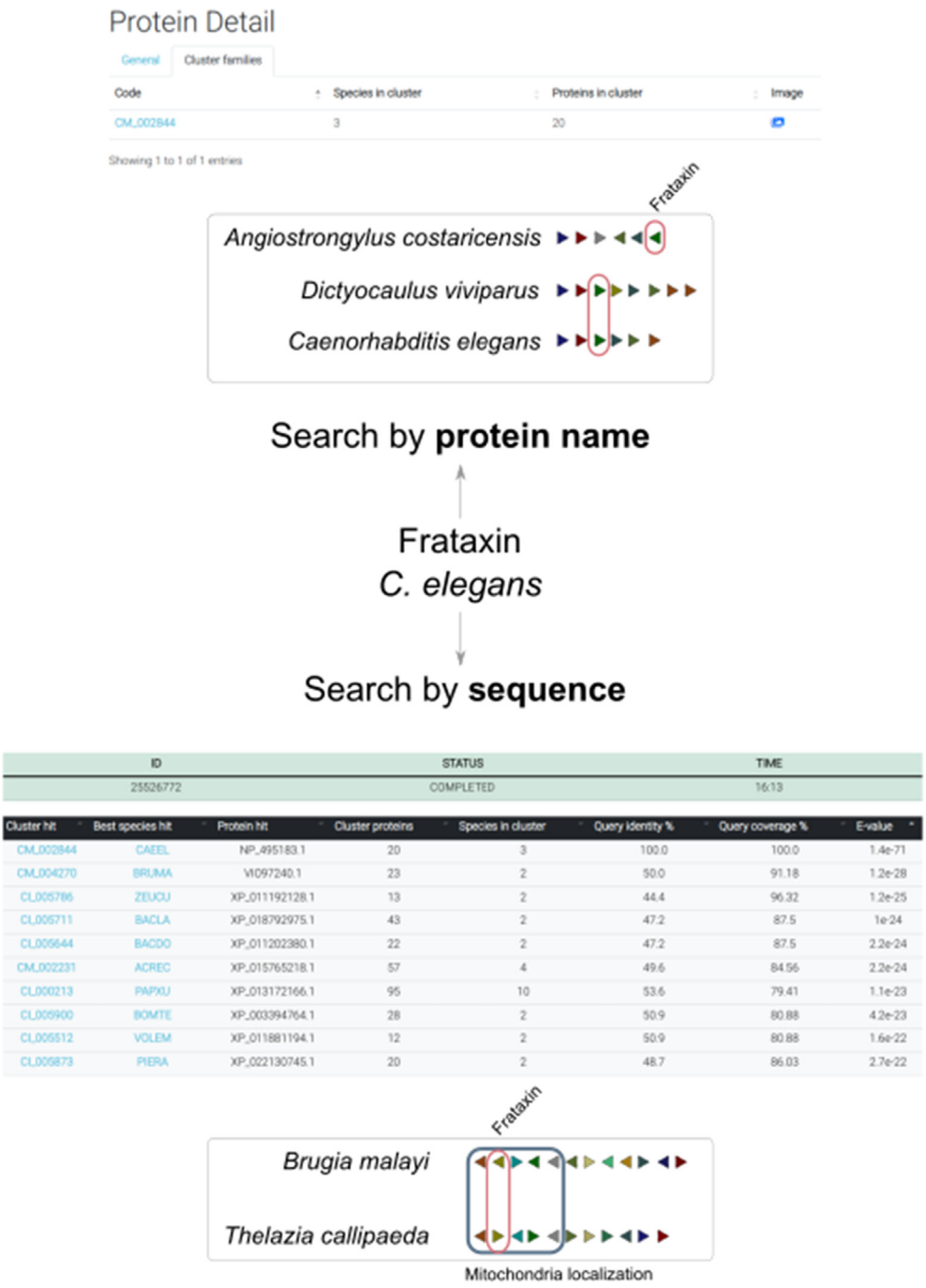

Figure 2. Example search results for the Frataxin gene in *C. elegans*. On the upper part of the figure are the results obtained with a search based on the protein code. On the lower part are the results obtained using a sequence based search.

(Figure 2). The CF had no gene families in common with the one found in *C. elegans* beyond the presence of the Frataxin gene. EvolClustDB allows not only the visualization of the CF associated with its species tree. It also allows the visualization of the annotated proteins. Using this feature, we explored the annotations of the genes found in the CF. The Frataxin gene in the operon found in *C. elegans* is associated with enzymes such as ceramide glucosyltransferase and tyrosine protein phosphatase, and a transcription enhancer factor. On the other hand the Frataxin gene found in the CF in *B. malayi* is surrounded by ribosomal proteins that localize to the mitochondria, and other proteins of unknown function.

The search also revealed that there were multiple CF found in the insect dataset containing Frataxin homologs (CI_005786, CI_005711, CI_005644, CI_000213, among others) and in the fungal

dataset (CF_012296 and CF_009163). This shows that the Frataxin genomic context tends to be conserved, but that the specific neighboring genes are clade-specific. The functional relationship between the Frataxin-surrounding genes and Frataxin in the different clades is unknown and deserves further investigation.

Materials and methods

Data and cluster definition

The datasets for fungi and insects were previously described.[15] For the three new datasets we downloaded 136 Metazoan, 104 plants and 156 protist proteomes from NCBI (Oct 2019). Proteomes were parsed so that they had the fasta headers needed to run EvolClust (see https://github.com/Gabaldonlab/EvolClust/wiki/4.–Input-files for details). An all against all BlastP[24] search was then performed for each of the datasets and based on these results a MCL clustering was performed (I = 1.5) in order to define the gene families. Once the gene families were defined for each dataset, EvolClust v1.0 was run using the default parameters and CFs were inferred.

Species trees

Species trees were obtained from NCBI using ETE3, as such they may contain multifurcations. Only the fungal species tree was obtained from the analysis performed in the previous study.[11] The species trees were pruned to each of the CFs using ETE3 so that they only contained species that had the CF in their genome, as this is the information shown in EvolClustDB.

Similarity searches performed in EvolClustDB

The similarity searches are performed using Diamond (v2.0.9.147).[25] Results are then filtered to keep only hits against proteins that are present in one of the CFs. Results are required to have a minimum e-value of 0.0001 and a minimum sequence overlap of 50%. If multiple hits point to proteins of the same CF only the best hit is shown to avoid repetitions. At most the top 150 hits per query are selected.

When a multiple fasta search is performed, Diamond results are grouped into genes that are close together in the genome based on the internal codes used in EvolClust and EvolClustDB. A minimum number of proteins homologous to the input needs to be found for the GC to be identified. This minimum is set by the user but it is recommended that it represents at least half of the GC for the results to be meaningful. Note that small GCs (less than 5 proteins) are unlikely to provide any results due to the limitations set by the EvolClust algorithm.

CRediT authorship contribution statement

Marina Marcet-Houben: Conceptualization, Data curation, Methodology, Writing – original draft, Writing – review & editing. **Ismael Collado-Cala:** Software, Visualization. **Diego Fuentes-Palacios:** Software, Visualization, Writing – original draft, Writing – review & editing. **Alicia D. Gómez:** Software, Visualization. **Manuel Molina:** Software, Visualization. **Andrés Garisoain-Zafra:** Data curation. **Uciel Chorostecki:** Conceptualization, Software, Visualization, Writing – original draft, Writing – review & editing. **Toni Gabaldón:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Acknowledgements

We want to acknowledge Joel Moro for his help in the analysis of the protist dataset and Daniel Majer for his help with the web interface. TG group acknowledges support from the Spanish Ministry of Science and Innovation for grant PGC2018-099921-B-I00, cofounded by European Regional Development Fund (ERDF); from the Catalan Research Agency (AGAUR) SGR423; from the European Union's Horizon 2020 research and innovation programme (ERC-2016-724173); from the Gordon and Betty Moore Foundation (Grant GBMF9742); from the "La Caixa" foundation (Grant LCF/PR/HR21/00737), and from the Instituto de Salud Carlos III (IMPACT Grant IMP/00019 and CIBERINFEC CB21/13/00061-ISCIII-SGEFI/ERDF). UC was funded in part through H2020 Marie Skłodowska-Curie Actions (H2020-MSCA-IF-2017-793699) and MICINN (IJC2019-039402-I).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received 4 November 2022;
Accepted 11 February 2023;
Available online 16 February 2023

Keywords:
gene order;
comparative genomics;
web server;
eukaryotes;
gene neighbourhood

Abbreviations:
CF, Cluster family; GC, Gene cluster

References

1. Hurst, L.D., Pál, C., Lercher, M.J., (2004). The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5**, 299–310.
2. Sproul, D., Gilbert, N., Bickmore, W.A., (2005). The role of chromatin structure in regulating the expression of clustered genes. *Nat. Rev. Genet.* **6**, 775–781.
3. Trowsdale, J., (2002). The gentle art of gene arrangement: the meaning of gene clusters. *Genome Biol.* **3** COMMENT2002.
4. Noonan, J.P., Grimwood, J., Schmutz, J., Dickson, M., Myers, R.M., (2004). Gene conversion and the evolution of protocadherin gene cluster diversity. *Genome Res.* **14**, 354–366.
5. Razin, S.V., Ioudinkova, E.S., Kantidze, O.L., Iarovaia, O. V., (2021). Co-Regulated Genes and Gene Clusters. *Genes* **12** https://doi.org/10.3390/genes12060907.
6. Pfannenstiel, B.T., Keller, N.P., (2019). On top of biosynthetic gene clusters: How epigenetic machinery influences secondary metabolism in fungi. *Biotechnol. Adv.* **37**, 107345
7. Slot, J.C., Rokas, A., (2010). Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. *PNAS* **107**, 10136–10141.
8. Hittinger, C.T., Rokas, A., Carroll, S.B., (2004). Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *PNAS* **101**, 14144–14149.
9. Horton, R., Wilming, L., Rand, V., Lovering, R.C., Bruford, E.A., Khodiyar, V.K., Lush, M.J., Povey, S., et al., (2004). Gene map of the extended human MHC. *Nat. Rev. Genet.* **5**, 889–899.
10. Lee, J.M., Sonnhammer, E.L.L., (2003). Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* **13**, 875–882.
11. Marcet-Houben, M., Gabaldón, T., (2019). Evolutionary and functional patterns of shared gene neighbourhood in fungi. *Nat. Microbiol.* **4**, 2383–2392.
12. Kautsar, S.A., Blin, K., Shaw, S., Weber, T., Medema, M. H., (2021). BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res.* **49**, D490–D497.
13. Pertea, M., Ayanbule, K., Smedinghoff, M., Salzberg, S.L., (2009). OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res.* **37**, D479–D482.
14. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., et al., (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613.
15. Marcet-Houben, M., Gabaldón, T., (2020). EvolClust: automated inference of evolutionary conserved gene clusters in eukaryotes. *Bioinformatics* **36**, 1265–1266.
16. Lopez-Diaz, T.M., Flannigan, B., (1997). Production of patulin and cytochalasin E by Aspergillus clavatus during malting of barley and wheat. *Int. J. Food Microbiol.* **35**, 129–136.
17. Morales, H., Marín, S., Rovira, A., Ramos, A.J., Sanchis, V., (2007). Patulin accumulation in apples by Penicillium expansum during postharvest stages. *Lett. Appl. Microbiol.* **44**, 30–35.
18. Kreft, L., Botzki, A., Coppens, F., Vandepoele, K., Van Bel, M., (2017). PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics* **33**, 2946–2947.
19. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., Punta, M., (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121.
20. Bencze, K.Z., Kondapalli, K.C., Cook, J.D., McMahon, S., Millán-Pacheco, C., Pastor, N., Stemmler, T.L., (2006). The structure and function of frataxin. *Crit. Rev. Biochem. Mol. Biol.* **41**, 269–291.
21. Blumenthal, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W.L., et al., (2002). A global analysis of Caenorhabditis elegans operons. *Nature* **417**, 851–854.
22. Wang, F., Huang, S., Ma, L., (2010). Caenorhabditis elegans operons contain a higher proportion of genes with multiple transcripts and use 3' splice sites differentially. *PLoS One* **5**, e12456.
23. Blumenthal, T., Davis, P., Garrido-Lecca, A., (2018). Operon and non-operon gene clusters in the *C. elegans* genome. *WormBook*.
24. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
25. Buchfink, B., Reuter, K., Drost, H.-G., (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368.