Computational and Structural Biotechnology Journal 20 (2022) 6570–6577

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Paralog Explorer: A resource for mining information about paralogs in common research organisms

Yanhui Hu [a,b,*,1], Ben Ewen-Campen [a,1], Aram Comjean [a,b], Jonathan Rodiger [a,b], Stephanie E. Mohr [a,b], Norbert Perrimon [a,b,c,*]

[a] Department of Genetics, Blavatnik Institute, Harvard Medical School, Harvard University, Boston, MA 02115, USA
[b] Drosophila RNAi Screening Center, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA
[c] Howard Hughes Medical Institute, Boston, MA 02138, USA

ARTICLE INFO

Article history:
Received 17 July 2022
Received in revised form 21 November 2022
Accepted 21 November 2022
Available online 24 November 2022

Keywords:
Evolution
Paralog
Drosophila
Model organisms
Bioinformatics resources

ABSTRACT

Paralogs are genes which arose via gene duplication, and when such paralogs retain overlapping or redundant function, this poses a challenge to functional genetics research. Recent technological advancements have made it possible to systematically probe gene function for redundant genes using dual or multiplex gene perturbation, and there is a need for a simple bioinformatic tool to identify putative paralogs of a gene(s) of interest. We have developed Paralog Explorer (https://www.flyrnai.org/tools/paralogs/), an online resource that allows researchers to quickly and accurately identify candidate paralogous genes in the genomes of the model organisms *D. melanogaster, C. elegans, D. rerio, M. musculus,* and *H. sapiens*. Paralog Explorer deploys an effective between-species ortholog prediction software, DIOPT, to analyze within-species paralogs. Paralog Explorer allows users to identify candidate paralogs, and to navigate relevant databases regarding gene co-expression, protein–protein and genetic interaction, as well as gene ontology and phenotype annotations. Altogether, this tool extends the value of current ortholog prediction resources by providing sophisticated features useful for identification and study of paralogous genes.

1. Introduction

Genes that arise as a result of gene duplication are known as paralogs. In cases where they retain overlapping function, this can represent a particular challenge to functional analysis [1]. Specifically, while loss-of-function experiments have been enormously successful to characterize individual gene function, this approach can fail when a target gene has a redundant or partially redundant paralog that can compensate in its absence. For example, large-scale studies in yeast provide evidence that, in aggregate, knocking out singleton genes (those without any paralog) tends to produce a stronger phenotypic effect than knocking out one member of a paralog pair, likely due at least partially to paralog-based redundancy [2]. Similarly, gene essentiality studies across numerous human cancer cell lines revealed that paralogs were far less likely than singleton genes to be essential [3,4].

Importantly, paralogs are not a minor curiosity in eukaryotic genomes. In fact, gene duplication and functional divergence have long been recognized as one of the most fundamental and widespread sources of evolutionary novelty [5], and paralogs are extraordinarily widespread in genomes. In the human genome, for instance, 70.5 % of all genes are estimated to have at least one paralog [6]. Similarly, 67 % of *Drosophila* genes have been estimated to have at least one paralog [1]. While many paralogs have diverged functionally and encode proteins of unrelated or non-overlapping roles, there is undoubtedly a large amount of functional diversity yet to be characterized which has thus far remained invisible to standard gene loss-of-function studies.

In recent years, it has become increasingly possible to perform double- or multiplex loss-of-function experiments using scalable techniques as double RNAi or CRISPR-based techniques. To date, massively parallel double-knock CRISPR approaches have been primarily applied to cell culture experiments, where it is possible to introduce dual-sgRNA libraries targeting tens of thousands of gene pairs [7–13]. However, in vivo CRISPR-based techniques for

* Corresponding authors at: Department of Genetics, Blavatnik Institute, Harvard Medical School, Harvard University, Boston, MA 02115, USA.
E-mail addresses: claire_hu@med.genetics.harvard.edu (Y. Hu), perrimon@receptor.med.harvard.edu (N. Perrimon).
[1] Contribute equally.

https://doi.org/10.1016/j.csbj.2022.11.041

dual- and multiplex loss-of-function experiments are rapidly being developed for model organisms [14–18].

To facilitate the functional studies of paralogs in model organisms, both in cell culture and in vivo, we have developed a simple but effective bioinformatic tool, Paralog Explorer (https://www.flyrnai.org/tools/paralogs/web/) to identify and explore paralogous genes. Paralog Explorer allows users to retrieve paralogs based on a single or multi-gene query, across a wide range of sequence similarity, and to provide relevant comparative information about the retrieved paralog pairs. Paralog Explorer is based on *D*rosophila RNAi Screening Center Integrative Ortholog Prediction Tool (DIOPT), which was developed to identify orthologous genes between species using an integrative approach [19]. By focusing the DIOPT algorithm within, as opposed to between, species, Paralog Explorer identifies paralogs within a given genome. Further, the resource retrieves associated public data and annotations such as chromosomal location, gene ontology annotation and protein–protein or genetic interactors as well as expression data from various tissues and cell lines for *Drosophila* and human. By providing paralog predictions alongside information such as expression profiling datasets, Paralog Explorer can help researchers predict which paralogous genes might act redundantly or otherwise in concert with one another, and thus to assist in designing targeted small- or large-scale experimental studies [4,11–13,20,21].

2. Materials and methods

2.1. Paralog information

Paralog information was obtained from DIOPT database release 8 [19]. DIOPT integrates 17 existing algorithms/resources and use a simple voting system for rapid identification of orthologs and paralogs among major model organisms. The organisms included in the Paralog Explorer resource are the nematode worm *C. elegans*, the fruit fly *D. melanogaster*, the mouse *M. musculus*, the zebrafish *D. rerio*, and human *H. sapiens.* Protein alignment information, including alignment length, percent similarity, and percent identity, were also imported from DIOPT. In addition, for genes in each paralog pair, the orthologs in more ancient species such as yeast orthologs for human, mouse, zebrafish, worm and *Drosophila* paralogous genes, and *Drosophila* orthologs for paralog genes in vertebrates are also analyzed. The common orthologs shared by both genes in a pair are identified and stored in database for display. Data files were exported from DIOPT in text format and were further processed using a local program. The output files are uploaded into a mySQL database.

2.2. Integration of omics datasets

For each gene in a paralog pair, we retrieved and integrated protein–protein interaction and genetic interaction data from MIST [22]. In addition, we also identified interactors in common for each paralog pair. Tissue- or cell line-specific expression datasets were also integrated. For each *Drosophila* paralog pair, modENCODE tissue-, developmental stage-, cell line-, and treatment-specific expression profiles provided by FlyBase were integrated [23–26]. For human paralog pairs, tissue-specific expression data from GTEx Portal (https://gtexportal.org/home/) as well as expression data for 490 ATCC cell lines from the Cancer Cell Line Encyclopedia (CCLE) (https://sites.broadinstitute.org/ccle/) were integrated [27,28]. In addition, Pearson correlation co-efficient scores were calculated for each dataset and synexpression analysis was done.

3. Integration of other annotation

The 'slim' versions of gene ontology (GO) annotations were retrieved from NCBI. For each gene pair, common GO slim terms were identified and stored. Genome coordinates were retrieved from NCBI EntrezGene. Phenotype annotations from FlyBase (r6.45) and gene group annotations from GLAD [29] were retrieved. This type of information is subject to update periodically in Paralog Explorer.

3.1. Web-based tool development

The Paralog Explorer web tool (https://www.flyrnai.org/tools/paralogs/) can be accessed directly or found at the 'Tools Overview' page at the DRSC/TRiP Functional Genomics Resources website (https://fgr.hms.harvard.edu/tools). The backend was written in PHP using the Symfony framework and the front-end HTML pages take advantage of the Twig template engine. The JQuery JavaScript library with the DataTables plugin is used for handling Ajax calls and displaying table views. The Bootstrap framework and some custom CSS are used on the user interface. A mySQL database is used to store the integrated information and analysis results (e.g., Pearson correlation co-efficient scores for synexpression). Both the website and databases are hosted on the O2 high-performance computing cluster, which is made available by the Research Computing group at Harvard Medical School.

3.2. Curation of a human paralog test list

To generate a list of predicted human paralog pairs to test the reliability of Paralog Explorer, we downloaded a list of 3,132 non-redundant paralog pairs from literature [30]. This list is comprised of two published datasets: 1,436 gene pairs from recent small-scale duplication events, and the rest from ancient whole-genome duplication events. From this list, we excluded 15 gene pairs that we could not confidently map to NCBI Entrez gene IDs, resulting in a total of 3,117 gene pairs in our human paralog test list.

4. Results

4.1. Database content and user interface features

To build Paralog Explorer, we retrieved all paralog predictions from DIOPT for human, mouse, zebrafish, fly and worm (Fig. 1). The 'DIOPT score' is the number of algorithms (eg. 7 out of 16 for human and *Drosophila* ortholog mapping) that support a given prediction, which we previously showed provides a measure of confidence in each prediction [19]. Protein alignment information, including the alignment length, percent similarity, and percent identity, was also imported from DIOPT. We find that 34 %-69 % of paralog pairs in Paralog Explorer are supported by 4 or more algorithms and 15–39 % have score equal or>6 (Table 1). We also imported Gene Ontology (GO) terms [31,32], protein–protein and genetic interaction data from MIST [22], expression data from publicly-available databases such as modENCODE [24], GTEx [27] and CCLE [28], and phenotype data for *Drosophila* from FlyBase [33] (Fig. 1).

With the Paralog Explorer web-tool, users can query a specific gene of interest, a list of genes, or any one of several precomputed gene lists from GLAD [29]. In addition, users can establish a filter based on DIOPT score, and for *Drosophila* and human genes, can establish a cut-off of transcriptional expression level

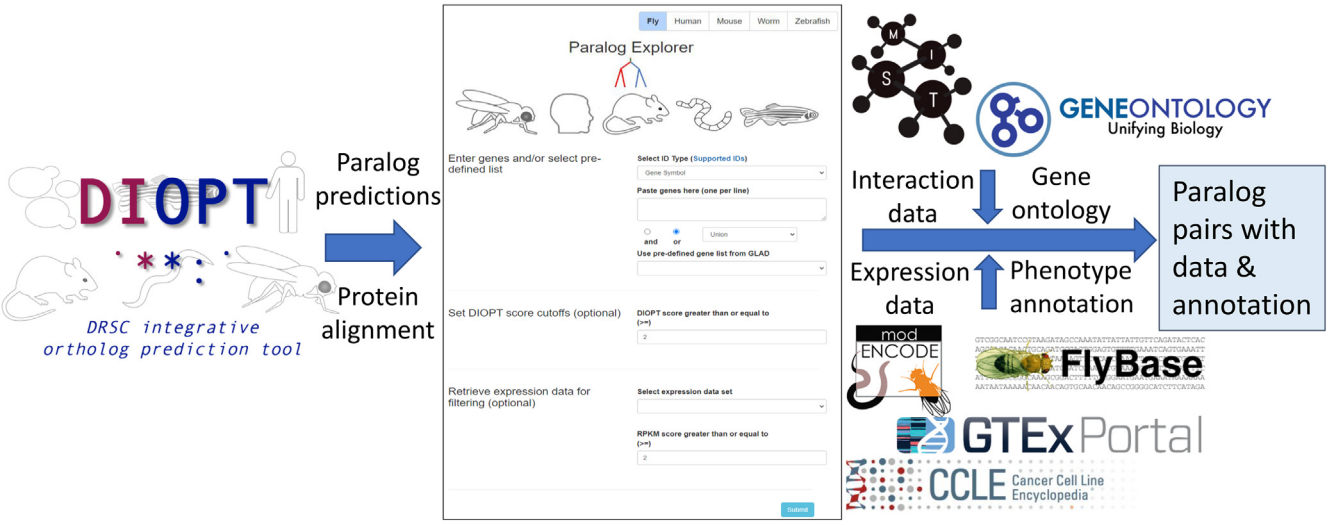

Fig. 1. Sources for information included in Paralog Explorer Paralog Explorer was built based on the integration of paralog predictions from DIOPT, PPI and genetic-interaction data from MIST, expression data from GTEx and modENCODE as well as gene ontology and phenotype annotation from GO consortium and FlyBase, respectively.

Table 1
Summary of predicted paralogs in Paralog Explorer.

NCBI TaxID (organism)	Maximum DIOPT score	Gene count (DIOPT score>=2)	Gene count (DIOPT score>=4)	Gene count (DIOPT score>=6)
6239 (worm)	10	15,333	5206 (34 %)	2245 (15 %)
7227 (fly)	10	10,092	4776 (47 %)	2486 (25 %)
7955 (zebrafish)	9	19,788	13,604 (69 %)	7685 (39 %)
9606 (human)	10	18,114	11,153 (62 %)	6559 (36 %)

in transcriptomic datasets from various tissues and cell lines for both genes in a pair. Altogether, the user interface is designed to allow users to address a variety of questions. These include very straightforward questions such as, does my gene of interest have one or more paralogs? Or, which of the genes in a list (e.g. hits from a genetic screen) have paralogs? Paralog Explorer also supports more complex queries such as, what are all the paralogous genes expressed in a given tissue or cell line? What are all the paralogous genes encoding transporters that are expressed at high levels in the adult digestive system?

For each query, the Paralog Explorer web-tool reports the total number of paralogs identified within a genome, each of which is shown on a separate line. For each paralog pair, Paralog Explorer displays information including the DIOPT score of the paralog, the genomic location of each gene, as well as various measures of protein alignment and Gene Ontology (GO) annotation for each member of the paralog pair, as well those GO terms common to both paralogs.

For each gene in a paralog pair, Paralog Explorer also reports the top-scoring ortholog(s) from the distantly-related outgroup yeast (*S. cerevisiae*), if such orthologs exist. This allows users to assess whether both paralogs in an animal model correspond to a single ortholog in yeast, which may assist in generating functional hypothesis. Similarly, for all vertebrate organisms, Paralog Explorer returns the closest fly ortholog for each member of a paralog pair. For example, PTPN11 and PTPN6, a paralogous gene-pair in humans, are both orthologous to *csw* in the *Drosophila* genome. This information can help to clarify whether a given paralog pair is the result of a lineage-specific gene duplication, or whether the duplication predated the divergence of these lineages [34].

The tool also integrates several -omics datasets of protein–protein and genetic interaction, to identify genetic and physical interactors of each gene in the paralog pair. Previous research has shown that protein interactions can be conserved after gene duplication [35], and in some cases paralogous genes which share common protein interactors may be more likely to be functionally related. This information may therefore be useful when prioritizing paralogs for further study or designing functional experiments.

For two paralogs to have redundant or partially redundant function in the cell, they must be expressed in the same cells and at the same time. Thus, when generating such hypotheses, it can be very helpful to compare expression patterns between paralogs. To facilitate this, we integrated tissue-specific and cell line-specific RNA-seq data from publicly available resources such as the GTEx and CCLE portals for human genes, as well as various modENCODE RNAseq datasets for *Drosophila*. Pearson correlation co-efficient scores for co-expression patterns are calculated based on each dataset respectively and are retrieved for users. For example, users can assess the co-expression of each human and *Drosophila* gene pairs based on either tissue-specific or cell line specific dataset. For each paralog pair, users also have the option to view the expression levels of each gene in the paralog gene pair from various datasets side-by-side as a bar graph (Fig. 2).

Users have the option to view a list of interacting partners for each gene or a list of interacting partners common to both genes (Fig. 2). The choice of columns to be displayed can be customized by the user and a results table can be exported as an Excel or tab-delimited text file so that the list by a parameter of choice can be easily filtered.

5. Application

Paralogs exist across a very broad range of evolutionary scenarios. In the conceptually simplest cases, a gene may have a single, evolutionarily recent paralog that is highly conserved at the sequence level, and perhaps located at an adjacent location in

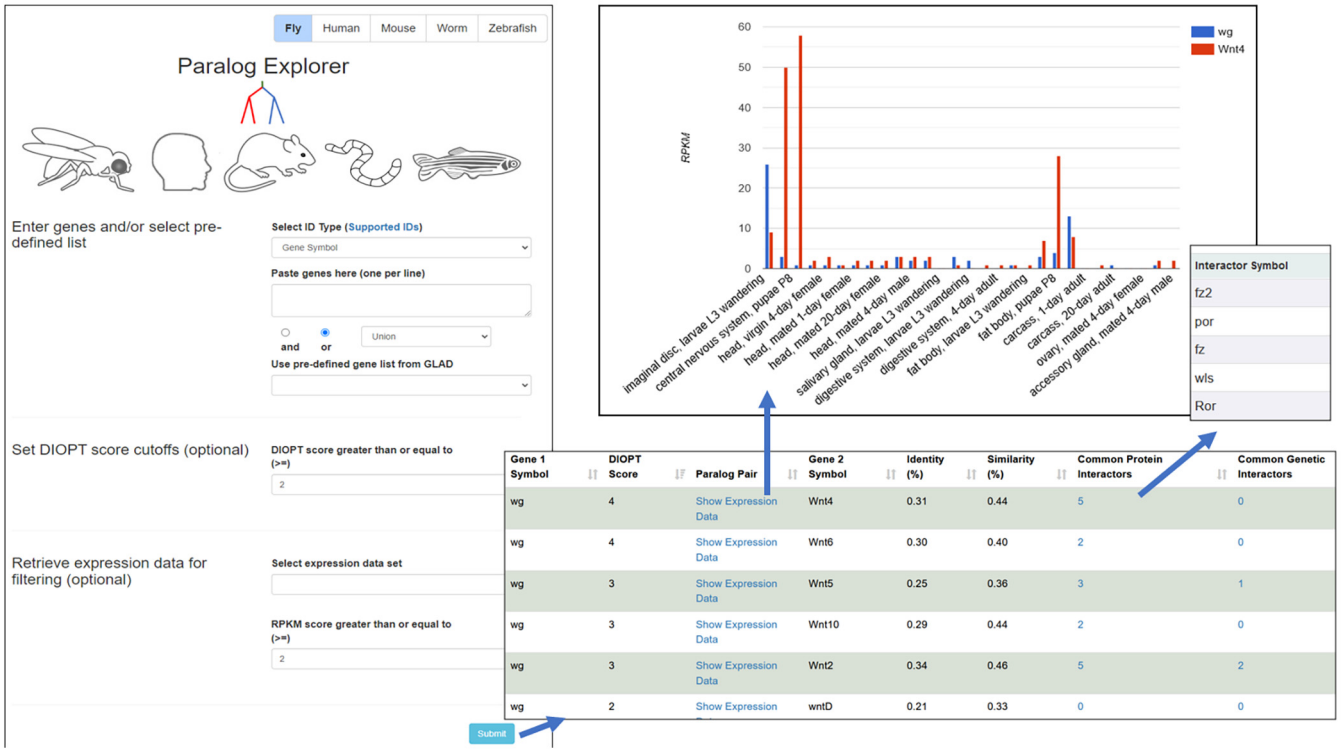

Fig. 2. Features included in the Paralog Explore user interface Paralog Explore is a web-based tool allowing user to select paralogs for input gene(s) along with interaction and expression data.

the genome. For example, the *Drosophila* zinc finger transcription factors *gcm* and *gcm2* share 48 % similarity at the amino acid level, are located just 26 kb apart from one another on the second chromosome, and have been experimentally shown to retain partially redundant functions [36].

In many other cases, a gene duplication event or the duplication of part or all of the entire genome may have occurred deep in evolutionary history, creating complex gene families composed of related genes at various degrees of sequence and functional similarities. For example, the *Hox* genes [37] and most of the major developmental signaling pathways [38] underwent duplication and diversification events very early in animal evolution, leading to a scenario today where all metazoan genomes contain varying copy numbers of each member of these gene families.

In still other cases, gene families may have dramatically expanded in certain animal lineages creating exceptionally large gene families with dozens or even hundreds of members, such as the over 900 odorant receptors encoded in the mouse genome [39].

Thus, in order to be useful to researchers with various interests, Paralog Explorer should quickly, accurately, and comprehensively identify paralogs at many different scales of similarity and genomic organization, and allow the user to investigate and rank the resulting hits based on their specific research context.

We sought to test the usefulness of Paralog Explorer to identify and characterize paralogs in three typical contexts, representing a range of gene similarity and paralog number: (1) amongst recently-diverged, highly conserved pairs/triplets of conserved paralogs; (2) amongst modestly-sized gene families that duplicated and diverged early in animal evolution and have been conserved as such in modern genomes; and (3) in a large gene family containing many dozens of paralogs.

To test the usefulness of Paralog Explorer on relatively simple cases, we examined a recently published list of 25 paralog pairs or triplets in the *Drosophila* genome that are closely related and physically linked in the genome, and for which there is evidence of transcriptional co-regulation via shared enhancers [40]. For each gene pair or triplet investigated by Levo *et al.,* we used Paralog Explorer to identify all predicted paralogs and ranked the results by DIOPT score. The results are presented in Table 2.

In 23 of 25 cases, Paralog Explorer identified the same top-scoring paralog as was identified by manual curation [40] (Table 2), and in the remaining two instances, additional examination provided an explanation. Among the former 23 cases, Paralog Explorer returned the predicted paralog as the best-scoring DIOPT hit and allowed the viewer to quickly confirm the chromosomal location of each gene, as well as to ascertain the co-expression patterns of the gene pairs in multiple high-throughput modENCODE datasets. In several instances, Paralog Explorer identified additional high-ranking paralogs that were not listed by Levo *et al.* but which appear to be *bona fide* paralogs. For example, *bowl* is a closely related paralog of *drm, sob,* and *odd*, and is also located in the same genomic region. Similarly, *comm3* is closely related to *comm* and *comm2*, and is located in the same genomic region (Table 2). Importantly, the existence of these additional paralogs may or may not reflect a functional conversation, but it allows researchers to systematically identify such genes for further study.

In addition to identifying the correct paralog as the top-scoring hit, Paralog Explorer also provides additional information that may be of interest. For nearly every gene query, Paralog Explorer identified a number of additional paralogs at varying degrees of similarity (Table 2). These results can be ranked by DIOPT score and/or by amino acid similarity, measures that are highly correlated with one another and serve as loose proxies for evolutionary conservation. Moreover, a user can also quickly determine whether such paralogs are physically linked in the genome and quickly access high-throughput co-expression datasets via hyperlinks.

Regarding the two cases for which Paralog Explorer did not return the same top hit as was identified via hand curation: in

Table 2
Performance of Paralog Explorer on a list of 25 curated paralog pairs/triplets from Levo *et al* (2022).

Paralog 1	Paralog 2	Paralog Explorer Top hit (DIOPT Score)	Note
drm	*sob / odd*	*sob (2), bowl (2)*, odd (1)*	bowl is located in same genomic region
slp1	*slp2*	*slp2 (6)*	16 additional lower-scoring paralogs (DIOPT <= 5)
H15	*mid*	*mid (9)*	7 additional lower-scoring paralogs (DIOPT <= 4)
gcm	*gcm2*	*gcm2 (4)*	Only hit
salr	*salm*	*salm (7)*	2 additional lower-scoring paralogs (DIOPT = 1)
nub	*pdm2*	*pdm2 (4)*	3 additional lower-scoring paralogs (DIOPT <= 3)
dnt	*drl*	*drl (9)*	18 additional lower-scoring paralogs (DIOPT <= 6)
inv	*en*	*en (5)*	16 additional lower-scoring paralogs (DIOPT = 1)
pyr	*ths*	*none*	pyr and ths are reciprocal best BLAST hit (33 % identity)
bab1	*bab2*	*bab2 (5)*	23 additional lower-scoring paralogs (DIOPT <= 4)
Doc1	*Doc2, Doc3*	*Doc3 (6), Doc2 (5)*	6 additional lower-scoring paralogs (DIOPT <=4)
scyl	*chrb*	*chrb (7)*	Only hit
toe	*eyg*	*eyg (4)*	22 additional lower-scoring paralogs (DIOPT <=3)
ara	*caup*	*caup (8)*	6 additional hits (DIOPT <= 7) including genomically linked *mirr*
comm	*comm2*	*comm2 (1), comm3* (1)*	All 3 genes genomically linked
kni	*knrl*	*knrl (4)*	2 additional paralogs (DIOPT <=3)
E5	*ems*	*ems (4)*	8 additional paralogs (DIOPT <= 2)
fd96Ca	*fd96Cb*	*fd96Cb (7)*	15 additional paralogs (DIOPT <= 4)
dan	*danr*	*danr (3)*	1 additional paralog (DIOPT = 1)
ac	*sc*	*l(1)sc* (7), sc (6)*	9 additional paralogs (DIOPT <= 5)
Vsx1	*Vsx2*	*Vsx2 (5)*	28 additional paralogs (DIOPT = 1)
btd	*Sp1*	*Sp1 (3)*	12 additional paralogs (DIOPT <= 3)
NetA	*NetB*	*NetB (6)*	8 additional paralogs (DIOPT <= 2)
disco	*disco-r*	*disco-r (8)*	1 additional paralog (DIOPT = 1)
B-H2	*B-H1*	*B-H1 (6)*	1 additional paralog (DIOPT = 1)

* Additional paralog identified at the same or higher DIOPT score as the predicted hit.

one case, *ac* and *sc*, Paralog Explorer identified an additional paralog, *l(1)sc*, as the top hit for *ac*, and *sc* as the second-highest hit. Thus, in this case, Paralog Explorer revealed biologically-relevant information. In the other case, *pyr* and *ths,* Paralog Explorer failed to return this pair because the current algorithms integrated by DIOPT database do not identify this pair as paralogs due to the low homology of FGF ligands [41], despite the fact that they are reciprocal best BLAST hits with one another in the *Drosophila* genome (E-value e-08, 33 % amino acid identity). Because Paralog Explorer is based on the DIOPT database, this error was propagated.

To extend these observations beyond *Drosophila*, we turned to a curated set of 3,117 human paralog pairs that includes paralogs across a wide range of sequence similarity and presumed duplication age (see Methods and [30]). For each of these paralog pairs, we inputted the first gene as a query in Paralog Explorer and asked whether the literature-predicted paralog appeared as the first, second, or third highest-scoring DIOPT score. Paralog Explorer identified the predicted paralog among the three top-scoring hits in 3,059 cases (98.1 %). In 2,301 of these cases (73.8 %), the predicted paralog was the top DIOPT hit, in 616 cases (19.8 %) it was the second-highest hit, and in 142 cases (4.6 %) it was the third-highest hit. In 55 cases (1.8 %), the predicted paralog was identified but ranked less than third highest. We observed just three cases (0.1 %) for which the predicted paralog was not identified at all. Subsequent evaluation suggested that two of these might be nomenclature-related issues, while the other one belongs to a large family of zinc-finger proteins containing over 100 members (Supplemental file). Altogether, the results of our analysis with a curated set of human paralog pairs demonstrates that Paralog Explorer reliably identifies known paralogs.

Many genes belong to "gene families" comprised of multiple paralogs that duplicated and diverged at varying points during evolution, rather than as a simple pair or triplet of recently duplicated, highly-similar paralogs. For example, the TGF-β genes are a family of secreted signaling ligands that arose and diversified very early in animal evolution, and today are present in varying numbers of paralogous genes in metazoan genomes; in *Drosophila*, there are seven TGF-β genes. Phylogenetically, the seven *Drosophila* ligands fall into three sub-families: the BMP-family ligands *dpp, gbb,* and *scw*, the Activin-family ligands *daw, myo, Actβ*, and the *mav* gene which does cleanly fall into either sub-family [42]. We searched Paralog Explorer using the canonical *Drosophila* ligand *dpp*, and successfully recovered all six paralogous ligands (Fig. 3). Furthermore, we noted that DIOPT scores between paralogs was generally reflective of the phylogenetic structure of the gene family [43] (Fig. 3). For example, *gbb* and *scw* display the highest DIOPT score (5), and both individually score next-highest to *dpp,* resembling the taxonomic structure of these three BMP-family ligands. However, we emphasize that DIOPT scores do not directly reflect phylogenetic relationships, and can depart significantly in cases where there has been significant evolutionary change along a specific branch. For example, based on DIOPT score alone, the *Actβ* gene is most closely related to *daw* (DIOPT score = 4) and equally similar to *myo* and the other four ligands (DIOPT score = 2), whereas phylogenetic analyses reveals that *Actβ* falls into a monophyletic Activin-like group with both *daw* and *myo*, and is more closely related to both of these two paralogs than it is to the remaining four [43] (Fig. 3).

We expanded our search of gene families to include several other highly conserved signaling pathways: the seven *Drosophila* Wnt ligands, the three JAK/STAT ligands (*upd* genes), the three Pvf ligands, and the five Spatzle ligands. Each of these gene families play important roles during development, each one expanded very early in animal evolution, and each family has been expanded and/or contracted in various animal lineages. For each, we entered a single family member into Paralog Explorer, and in 100 % of these examples Paralog Explorer correctly returned the entire family of related paralogs (Table 3). We note that these gene families contain a broad range of divergence amongst family members, demonstrating that Paralog Explorer is able to robustly and accurately predict the full suite of paralogs for a given gene across a wide range of evolutionary divergence and amongst complex gene families. For the Wnt family ligands, we repeated the exercise of comparing DIOPT scores to known phylogenetic relationships [44] (Fig. 3). Again, dominant phylogenetic patterns of sequence conservation were reflected by DIOPT scores, while not precisely mirroring the known phylogenetic relationships. Specifically, reciprocal DIOPT scores identified *wg, wnt6,* and *wnt4* as closely related, and a close relationship between *wnt2* and *wnt5*, while the divergent *wntD* gene stood out as distinct from all other family members, all of which is reflective of known phylogenetic patterns [44]. Importantly, as with the example of TGF-β ligands shown above, results from Paralog Explorer should not be interpreted as

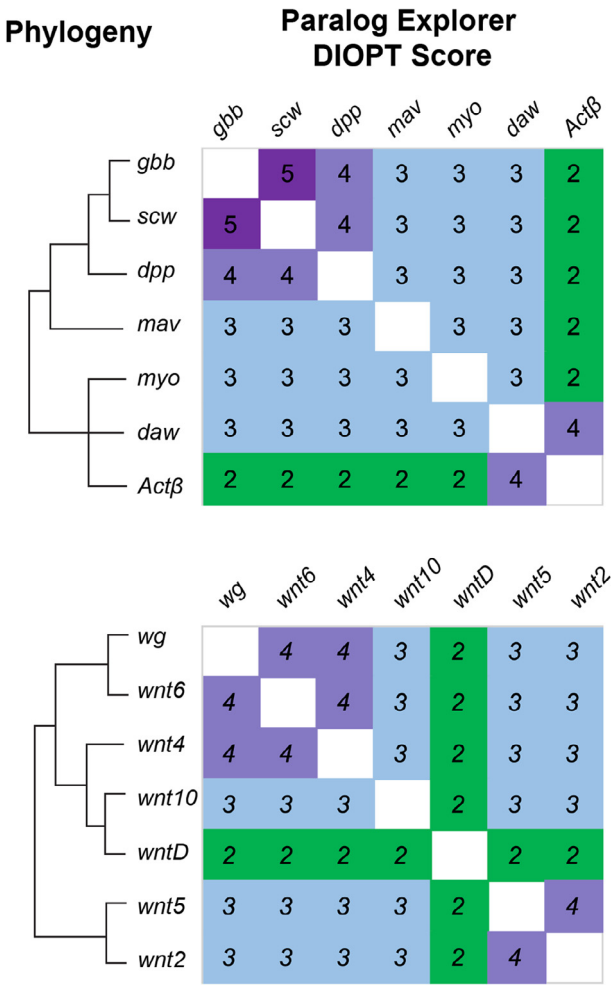

Fig. 3. Paralog Explorer identifies known members of paralogous gene families. Using a single gene as a query, Paralog Explorer correctly identifies the complete gene family of all known TGFβ ligands and Wnt ligands. The known gene family phylogeny is shown at left (see text for references), and a heatmap of the pairwise DIOPT scores is shown at right.

Table 3
Identification of multi-gene families using Paralog Explorer.

Query Ligand	Known Paralogs	Identified via Paralog Explorer (DIOPT Score)	% of known identified
dpp	*gbb, scw, daw, mav, myo, Actβ*	*gbb (4), scw (4), daw (3), mav (3), myo (3), Actβ (2)*	100
wg	*wnt2, wnt4, wnt5, wnt6, wntD, wnt10*	*wnt2 (3), wnt4 (4), wnt5 (3), wnt6 (4), wntD (2), wnt10 (3)*	100
upd	*upd2, upd3*	*upd2 (2), upd3 (1)*	100
Pvf1	*Pvf2, Pvf3*	*Pvf2 (1), Pvf3 (1)*	100
spz	*NT1 (aka spz2), spz3, spz4, spz5, spz6*	*NT1 (1), spz3 (1), spz4 (1), spz5 (2), spz6 (1)*	100

directly reflective of phylogenetic relationships or functional conservation, but can provide potentially helpful information to generate hypothesis about genetic similarity amongst paralogs.

We then wished to know how well Paralog Explorer performed on very large gene families. We chose the odorant receptor (*Or*) gene family, of which there are 60 members in the *Drosophila* genome, as well as one pseudogene [45]. Remarkably, using *Or1a* as our query, Paralog Explorer returned exactly 59 paralogs, only failing to return the single pseudogenic member of this family noted in [45]. Thus, in even in the case of highly expanded gene families such as the *Or* genes, Paralog Explorer correctly identifies all known paralogs.

In addition to providing users with the ability to identify paralogs for individual queries or lists, Paralog Explorer also has the potential to assist in large-scale bioinformatic analyses. To demonstrate one such use case, we compared the paralog annotation with a synthetic lethality screen using CRISPR-Cas9 dual targeting [46] in human cell lines. Out of 406 heterogenous gene pairs, 21 pairs are annotated as paralogs in Paralog Explorer. Furthermore, 9 of the 21 paralog gene pairs (43 %) are scored as synthetic lethality interactors with one another in at least one cell line by the criteria of FDR < 0.1while only 20 out of 385 other gene pairs (5 %) scored. Paralogous gene pairs are much more likely to score in functional screens than are pairs of unrelated genes, and not surprisingly, more recent studies are focused on paralog gene pairs rather than randomly selected gene pairs [7,9–13]. Thus, we expect that Paralog Explorer will facilitate the experimental design of high-throughput screens and mapping of functionally related genes.

There is an important caveat to Paralog Explorer, which is likely common to all paralog prediction methods. Because hypotheses of paralogy are primarily drawn from sequence conservation, gene queries which contain individual protein domains that are highly conserved may return many putative paralogs, based on the presence of shared domains across proteins that are otherwise only distantly-related. Furthermore, the sequence length of these conserved domains will impact paralog predictions, such that longer domains are more likely to score higher while relatively short domains may not reach the threshold to score via DIOPT. While the presence of shared domains across proteins may in fact reflect a true evolutionary history of gene duplication, from a practical standpoint it can lead to complex results that require sophisticated manual analysis.

As one example, we examined the *Drosophila* Hox genes, which are transcription factors that include the highly conserved homeodomain, a ~ 60aa domain that is widespread across many transcription factors [37]. Inputting the anterior-most *Drosophila Hox* gene *lab* as a query, Paralog Explorer returns 23 predicted paralogs. These include the other Hox genes themselves (*pb, Dfd, Scr, Antp, Ubx, Abd-A, Abd-B*), as well as the paralogous homeodomain proteins *bcd, zen,* and *zen*-2 that are located nearby in the genome, all at a DIOPT score of 2. In addition, however, Paralog Explorer returns as its highest hit the homeobox gene *ro* (DIOPT score 3), which is not considered a Hox gene, as well as 12 additional genes, all but one of which are known homeobox-containing genes. Importantly this list is not comprehensive of all homeobox-containing genes. FlyBase identifies a total of 102 homeobox genes in the *Drosophila* genome, indicating that this hit list includes those that reach some similarity threshold based on DIOPT scores. Thus, for genes that contain specific highly conserved domains found in many genes, users should carefully analyze the results when forming hypotheses about paralogy.

Users of Paralog Explorer can rank paralogs based on the DIOPT score, which serves as a proxy for protein sequence similarity. Importantly, the tool also provides additional valuable ways to analyze paralog relationships aside from sequence similarity. For example, to generate hypotheses about which paralog pairs are likely to have overlapping functions, users can perform co-expression analysis, as genes showing synexpression (i.e., high degrees of correlated co-expression) often operate in similar pathways and/or processes. We analyzed the correlation of similarity in protein sequence and expression pattern by comparing the DIOPT scores and the percentage of gene pairs with high synexpression scores (Pearson correlation co-efficient score > 0.5) calculated based on cell line RNAseq data sets. RNAseq data from single cells or groups of homogeneous cell populations has been shown to be better than tissue-based datasets for synexpression analysis [47].

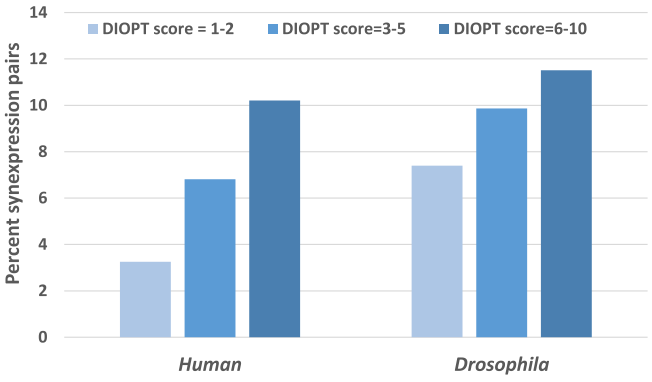

Fig. 4. The correlation of sequence-based scores and synexpression in cell lines
The correlation of similarity in protein sequence and expression pattern was analyzed by comparing the DIOPT scores and the percentage of paralog pairs with high synexpression scores (Pearson correlation co-efficient score > 0.5) calculated based on cell line RNA-seq data sets. We observed that the gene pairs of higher sequence similarity were more likely to have synexpression pattern for both *Drosophila* and human paralog pairs.

Table 4
Examples of functionally relevant human paralogs from the literature.

Paralog Pair	DIOPT score	*syn*-express score (cell lines)
SMARCA2-SMARCA4	8 (top1)	0.25 (top1)
STAG1-STAG2	7 (top2)	0.39 (top1)
DUSP4-DUSP6	4 (top3)	0.39 (top1)
DDX3X-DDX3Y	7 (top1)	−0.06

We found that gene pairs with higher sequence similarity were more likely to have higher synexpression scores for both *Drosophila* and human paralog pairs (Fig. 4).

We emphasize that paralog pairs that have high synexpression scores but are not necessarily the absolute top DIOPT-scoring pairs can also be functionally related. For example, the human genes DUSP4 and DUSP6 have been demonstrated to be functional paralogs that share a digenic dependence in MAPK pathway-driven cancers [9]. Both genes are part of a larger gene family. Based on DIOPT scores alone, the highest-ranking paralogs for DUSP4 are DUSP1 and DUSP10, while DUSP6 ranks third. However, DUSP4 displays the highest synexpression score with DUSP6 compared to any other DUSP, reflecting the fact that they are often co-expressed. We examined three other well-characterized human paralog pairs (SMARCA2/SMARCA4, DDX3X/DDX3Y, and STAG1/STAG2) [48–50] and found that the well-characterized paralog pairs are the top-ranked for: both DIOPT and synexpression (SMARCA2/SMARCA4); synexpression but not DIOPT score (STAG1/STAG2); and DIOPT score but not synexpression (DDX3X/DDX3Y). Thus, we designed Paralog Explorer to allow users to rank Paralog candidates according to multiple rubrics and thus to generate context-specific hypotheses about functional relevance (Table 4).

6. Discussion

Paralog Explorer is a tool which allows users to quickly and reliably identify paralogs of any gene(s) of interest, as well as relevant measures of their similarity, genomic location, co-expression patterns, genetic and protein interactions, and GO terms. It is important to note that identifying two or more genes as paralogs is a hypothesis about their evolutionary history – i.e. that they arose via gene duplication – rather than about molecular function and or whether they may be functionally redundant. Thus, we designed Paralog Explorer to be a flexible search tool that will allow researchers with diverse interests to generate hypotheses about paralogous genes.

We have shown that Paralog Explorer can reliably and robustly identify known paralogs across a wide range of sequence similarities. We emphasize that there is no "one size fits all" approach to deciding which paralogs are relevant for different biological questions. For this reason, Paralog Explorer allows users to rank results based on a number of measures, including DIOPT score, sequence similarity, chromosomal location and synexpression scores. We have shown that the DIOPT score is often a useful, though very coarse, proxy for phylogenetic proximity, and have also provided several examples for which the 'functionally relevant' paralog may not necessarily be the top-scoring DIOPT hit. Paralog Explorer is designed to accommodate this wide range of biological realities and to provide users with easily accessible bioinformatic information to help generate hypotheses.

We note that Paralog Explorer is built based on predictions in DIOPT, which in some instances may fail to include certain paralogs that are validated by experimental data or published literature. As DIOPT is updated and improved via user-submitted data, Paralog Explorer will be updated accordingly.

CRediT authorship contribution statement

Yanhui Hu: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Supervision, Project administration. **Ben Ewen-Campen:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing – original draft. **Aram Comjean:** Software, Visualization. **Jonathan Rodiger:** Software, Visualization. **Stephanie E. Mohr:** Writing – review & editing, Supervision. **Norbert Perrimon:** Writing – review & editing, Supervision, Project administration.

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.

Acknowledgements

We would like to thank the members of the Perrimon laboratory, the FlyBase consortium, the Drosophila RNAi Screening Center (DRSC), and the Transgenic RNAi Project (TRiP) for helpful suggestions. This work was supported by NIH/NIGMS grant P41 GM132087 and NCRR/ORIP grant R24-OD021997. N.P. is an investigator of Howard Hughes Medical Institute.

Availability

The online resource is available without restriction at https://www.flyrnai.org/tools/paralogs/web/.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.11.041.

References

[1] Ewen-Campen B et al. Accessing the Phenotype Gap: Enabling Systematic Investigation of Paralog Functional Complexity with CRISPR. Dev Cell 2017;43(1):6–9.
[2] Gu Z et al. Role of duplicate genes in genetic robustness against null mutations. Nature 2003;421(6918):63–6.

[3] De Kegel B, Ryan CJ. Paralog buffering contributes to the variable essentiality of genes in cancer cell lines. PLoS Genet 2019;15(10):e1008466.
[4] Wang T et al. Identification and characterization of essential genes in the human genome. Science 2015;350(6264):1096–101.
[5] Ohno S. *Evolution by Gene Duplication*. 1970: Springer Berlin, Heidelberg. 798812.
[6] Ibn-Salem J, Muro EM, Andrade-Navarro MA. Co-regulation of paralog genes in the three-dimensional chromatin architecture. Nucleic Acids Res 2017;45(1):81–91.
[7] Chow RD et al. In vivo profiling of metastatic double knockouts through CRISPR-Cpf1 screens. Nat Methods 2019;16(5):405–8.
[8] Han K et al. Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. Nat Biotechnol 2017;35(5):463–74.
[9] Ito T et al. Paralog knockout profiling identifies DUSP4 and DUSP6 as a digenic dependence in MAPK pathway-driven cancers. Nat Genet 2021;53(12):1664–72.
[10] Thompson NA et al. Combinatorial CRISPR screen identifies fitness effects of gene paralogues. Nat Commun 2021;12(1):1302.
[11] De Kegel B et al. Comprehensive prediction of robust synthetic lethality between paralog pairs in cancer cell lines. Cell Syst 2021;12(12):1144–1159 e6.
[12] Dede M et al. Multiplex enCas12a screens detect functional buffering among paralogs otherwise masked in monogenic Cas9 knockout screens. Genome Biol 2020;21(1):262.
[13] Koferle A et al. Interrogation of cancer gene dependencies reveals paralog interactions of autosome and sex chromosome-encoded genes. Cell Rep 2022;39(2):110636.
[14] Ewen-Campen B et al. No Evidence that Wnt Ligands Are Required for Planar Cell Polarity in Drosophila. Cell Rep 2020;32(10):108121.
[15] Guo LY et al. Multiplexed genome regulation in vivo with hyper-efficient Cas12a. Nat Cell Biol 2022;24(4):590–600.
[16] Parvez S et al. MIC-Drop: A platform for large-scale in vivo CRISPR screens. Science 2021;373(6559):1146–51.
[17] Port F, Bullock SL. Augmenting CRISPR applications in Drosophila with tRNA-flanked sgRNAs. Nat Methods 2016;13(10):852–4.
[18] Port F, Starostecka M, Boutros M. Multiplexed conditional genome editing with Cas12a in Drosophila. Proc Natl Acad Sci U S A 2020;117(37):22890–9.
[19] Hu Y et al. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinf 2011;12:357.
[20] Barshir R et al. Role of duplicate genes in determining the tissue-selectivity of hereditary diseases. PLoS Genet 2018;14(5):e1007327.
[21] Jubran J et al. Dosage-sensitive molecular mechanisms are associated with the tissue-specificity of traits and diseases. Comput Struct Biotechnol J 2020;18:4024–32.
[22] Hu Y et al. Molecular Interaction Search Tool (MIST): an integrated resource for mining gene and protein interaction data. Nucleic Acids Res 2018;46(D1):D567–74.
[23] Boley N et al. Navigating and mining modENCODE data. Methods 2014;68(1):38–47.
[24] Brown JB et al. Diversity and dynamics of the Drosophila transcriptome. Nature 2014;512(7515):393–9.
[25] Graveley BR et al. The developmental transcriptome of Drosophila melanogaster. Nature 2011;471(7339):473–9.
[26] mod EC et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 2010;330(6012):1787–97.
[27] Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45(6):580–5.
[28] Barretina J et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483(7391):603–7.
[29] Hu Y et al. GLAD: an online database of gene list annotation for drosophila. J Genomics 2015;3:75–81.
[30] Dandage R, Landry CR. Paralog dependency indirectly affects the robustness of human cells. Mol Syst Biol 2019;15(9):e8871.
[31] Ashburner M et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25(1):25–9.
[32] Gene Ontology C. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 2021;49(D1):D325–34.
[33] Gramates LS et al. FlyBase: a guided tour of highlighted features. Genetics 2022;220(4).
[34] Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 2005;39:309–38.
[35] Pereira-Leal JB et al. Evolution of protein complexes by duplication of homomeric interactions. Genome Biol 2007;8(4):R51.
[36] Alfonso TB, Jones BW. gcm2 promotes glial cell differentiation and is required with glial cells missing for macrophage development in Drosophila. Dev Biol 2002;248(2):369–83.
[37] Singh NP, Krumlauf R. Diversification and functional evolution of HOX proteins. Front Cell Dev Biol 2022;10:798812.
[38] Pires-daSilva A, Sommer RJ. The evolution of signalling pathways in animal development. Nat Rev Genet 2003;4(1):39–49.
[39] Godfrey PA, Malnic B, Buck LB. The mouse olfactory receptor gene family. Proc Natl Acad Sci U S A 2004;101(7):2156–61.
[40] Levo M et al. Transcriptional coupling of distant regulatory genes in living embryos. Nature 2022;605(7911):754–60.
[41] Stathopoulos A et al. pyramus and thisbe: FGF genes that pattern the mesoderm of Drosophila embryos. Genes Dev 2004;18(6):687–99.
[42] Upadhyay A et al. TGF-beta family signaling in drosophila. Cold Spring Harb Perspect Biol 2017;9(9).
[43] Van der Zee M, da Fonseca RN, Roth S. TGFbeta signaling in Tribolium: vertebrate-like components in a beetle. Dev Genes Evol 2008;218(3–4):203–13.
[44] Janssen R et al. Conservation, loss, and redeployment of Wnt ligands in protostomes: implications for understanding the evolution of segment formation. BMC Evol Biol 2010;10:374.
[45] Guo S, Kim J. Molecular evolution of Drosophila odorant receptor genes. Mol Biol Evol 2007;24(5):1198–207.
[46] Najm FJ et al. Orthologous CRISPR-Cas9 enzymes for combinatorial genetic screens. Nat Biotechnol 2018;36(2):179–89.
[47] Hu Y et al. The Drosophila Gene Expression Tool (DGET) for expression analyses. BMC Bioinf 2017;18(1):98.
[48] Allen MD, Bycroft M, Zinzalla G. Structure of the BRK domain of the SWI/SNF chromatin remodeling complex subunit BRG1 reveals a potential role in protein-protein interactions. Protein Sci 2020;29(4):1047–53.
[49] Bailey ML et al. Paralogous synthetic lethality underlies genetic dependencies of the cancer-mutated gene STAG2. Life Sci Alliance 2021;4(11).
[50] Rosner A, Rinkevich B. The DDX3 subfamily of the DEAD box helicases: divergent roles as unveiled by studying different organisms and in vitro assays. Curr Med Chem 2007;14(23):2517–25.