

Journal of Visualized Experiments

Navigating MARRVEL, a web-based tool that integrates human genomics and model organism genetics information

--Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE59542R2
Full Title:	Navigating MARRVEL, a web-based tool that integrates human genomics and model organism genetics information
Keywords:	Human genomics; variant prioritization; model organisms; Genetics; rare and undiagnosed diseases; functional genomics; database integration; translational research; medical diagnosis; variant of unknown significance (VUS); gene of uncertain significance (GUS); web-based tool
Corresponding Author:	Julia Wang Baylor College of Medicine Houston, UNITED STATES
Corresponding Author's Institution:	Baylor College of Medicine
Corresponding Author E-Mail:	Julia.Wang@bcm.edu
Order of Authors:	Julia Wang Zhandong Liu Hugo J. Bellen Shinya Yamamoto
Additional Information:	
Question	Response
Please indicate whether this article will be Standard Access or Open Access.	Standard Access (US\$2,400)
Please indicate the city, state/province, and country where this article will be filmed . Please do not use abbreviations.	Houston, Texas, USA

Dear Dr. Stephanie R. Weldon,

It is our pleasure to submit our manuscript “**Navigating MARRVEL, a web-based tool that integrates human genomics and model organism genetics information**” to JoVE.

We believe that this manuscript will be useful to a wide range of clinicians and researchers who are interested in bridging human genetics and model organisms research.

In addition to this paper, we worked with Dr. Lyndsay Troyer on another manuscript titled “**In vivo functional study of disease-associated rare human variants using Drosophila**” which will be submitted very soon. These two articles complement each other to form a complete pipeline from analyzing the sequencing results of patients to functional validation in vivo. We hope the two articles will be useful for the readership of your journal.

Sincerely,

Shinya Yamamoto and Julia Wang

Sincerely,



Shinya Yamamoto, D.V.M., Ph.D.

Co-Director of MOSC-UDN

Assistant Professor

Department of Molecular and Human Genetics

Baylor College of Medicine

1250 Moursund Street, Ste 1050.16

Houston, TX, 77030

Office: (832) 824-8119

Laboratory: (832) 824-8723

E-mail: yamamoto@bcm.edu

TITLE:

Navigating MARRVEL, a Web-Based Tool that Integrates Human Genomics and Model Organism Genetics Information

AUTHORS AND AFFILIATIONS:

Julia Wang^{1,2}, Undiagnosed Diseases Network*, Zhandong Liu^{3,4}, Hugo J. Bellen^{1,4,5,6,7}, Shinya Yamamoto^{1,4,5,6}

¹ Program in Developmental Biology, Baylor College of Medicine, Texas, USA

² Medical Scientist Training Program, Baylor College of Medicine, Texas, USA

³ Department of Pediatrics, Baylor College of Medicine, Texas, USA

⁴ Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Texas, USA

⁵ Department of Molecular and Human Genetics, Baylor College of Medicine, Texas, USA

⁶ Department of Neuroscience, Baylor College of Medicine, Texas, USA

⁷ Howard Hughes Medical Institute, Baylor College of Medicine, Texas, USA

*Members of the Undiagnosed Diseases Network is provided in **Supplemental Table 1**.

Corresponding Author:

Shinya Yamamoto (yamamoto@bcm.edu)

Email Addresses of Co-Authors:

Julia Wang (julia.wang@bcm.edu)

Zhandong Liu (zhandonl@bcm.edu)

Hugo J. Bellen (hbell@bcm.edu)

KEYWORDS:

Human genomics, variant prioritization, model organisms, genetics, rare and undiagnosed diseases, functional genomics, database integration, translational research, medical diagnosis, variant of unknown significance, gene of uncertain significance, web-based tool.

SUMMARY:

Here, we present a protocol to access and analyze many human and model organism databases efficiently. This protocol demonstrates the use of MARRVEL to analyze candidate disease-causing variants identified from next-generation sequencing efforts.

ABSTRACT:

Through whole-exome/genome sequencing, human geneticists identify rare variants that segregate with disease phenotypes. To assess if a specific variant is pathogenic, one must query many databases to determine whether the gene of interest is linked to a genetic disease, whether the specific variant has been reported before, and what functional data is available in model organism databases that may provide clues about the gene's function in human. MARRVEL (Model organism Aggregated Resources for Rare Variant ExpLoration) is a one-stop data collection tool for human genes and variants and their orthologous genes in seven model

organisms including in mouse, rat, zebrafish, fruit fly, nematode worm, fission yeast, and budding yeast. In this Protocol, we provide an overview of what MARRVEL can be used for and discuss how different datasets can be used to assess whether a variant of unknown significance (VUS) in a known disease-causing gene or a variant in a gene of uncertain significance (GUS) may be pathogenic. This protocol will guide a user through searching multiple human databases simultaneously starting with a human gene with or without a variant of interest. We also discuss how to utilize data from OMIM, ExAC/gnomAD, ClinVar, Geno2MP, DGV and DECHIPHER. Moreover, we illustrate how to interpret a list of ortholog candidate genes, expression patterns, and GO terms in model organisms associated with each human gene. Furthermore, we discuss the value protein structural domain annotations provided and explain how to use the multiple species protein alignment feature to assess whether a variant of interest affects an evolutionarily conserved domain or amino acid. Finally, we will discuss three different use-cases of this website. MARRVEL is an easily accessible open access website designed for both clinical and basic researchers and serves as a starting point to design experiments for functional studies.

INTRODUCTION:

The use of next-generation sequencing technology is expanding in both research and clinical genetic laboratories¹. Whole-exome (WES) and whole-genome sequencing (WGS) analyses reveal numerous rare variants of unknown significance (VUS) in known disease-causing genes as well as variants in genes that are yet to be associated with a Mendelian disease (GUS: genes of uncertain significance). Presented with a list of genes and variants in a clinical sequence report, medical geneticists must manually visit multiple online resources to obtain more information to assess which variant may be responsible for a certain phenotype seen in the patient of interest. This process is time-consuming, and its efficacy is highly dependent on the expertise of the individual. Although several guideline papers have been published^{2,3}, interpretation of WES and WGS requires manual curation since there is yet to be a standardized methodology for variant analysis. For the interpretation of VUS, knowledge on the previously reported genotype-phenotype relationship, mode of inheritance, and allele frequencies in the general population become valuable. In addition, knowledge on whether the variant affects a critical protein domain, or an evolutionarily conserved residue may increase or decrease the likelihood of pathogenicity. To gather all of this information, one typically needs to navigate through 10-20 human and model organism databases since the information is scattered through the World Wide Web.

Similarly, model organism scientists who work on specific genes and pathways are often interested in connecting their findings to human disease mechanisms and wish to take advantage of the knowledge that is being generated in the human genomics field. However, due to the rapid expansion and evolution of data sets regarding the human genome, it has been challenging to identify databases that provide useful information. In addition, since most model organism databases are designed for researchers who work with the specific organism on a daily basis, it is very difficult, for example, for a mouse researcher to search for specific information in a *Drosophila* database and *vice versa*. Similar to the variant interpretation searches performed by medical geneticists, identifying useful human and other model organism information is time-consuming and heavily dependent on the background of the model organism researcher. MARRVEL (Model organism Aggregated Resources for Rare Variant ExpLoration)⁴ is a tool

designed for both groups of users to streamline their workflow.

MARRVEL (<http://marrvel.org>) was designed as a centralized search engine that collects data systematically in an efficient and consistent manner for clinicians and researchers. With information from 20 or more publicly available databases, this program allows users to quickly gather information and access a large number of human and model organism databases without reiterative searches. The search result pages also contain hyperlinks to the original sources of information, allowing individuals to access the raw data and gather additional information provided by the sources.

In contrast to many of the variant prioritization tools that require large sequencing data input in the form of VCF or BAM files and installations of often proprietary/commercial software, MARRVEL operates on any web-browser. It can be used at no cost and compatible with portable devices (e.g. smartphones, tablets) as long as one is connected to the internet. We chose this format since many clinicians and researchers typically need to search one or a few genes and variants at a time. Note that we are developing batch-download and API (application programming interface) features for MARRVEL to eventually allow users to curate hundreds of genes and variants at a time through customized query tools if necessary.

Due to the wide range of applications, in this protocol, we will describe a broadly encompassing approach on how to navigate through different datasets that MARRVEL displays. More targeted examples that are tailored towards specific users' needs will be described in Representative Results section. It is important to note that the output of MARRVEL still requires a certain level of background knowledge in either human genetics or model organisms to extract valuable information. We refer the readers to the table that lists primary papers that describe the function of each of the original databases that are curated by MARRVEL (**Table 1**). The following protocol is divided into three sections: (1) How to begin a search, (2) how to interpret MARRVEL human genetics outputs, and (3) how to make use of model organism data in MARRVEL. In the Representative Results section, more focused and specific approaches are described. MARRVEL is being actively updated so please refer to the current website's FAQ page for details about data sources. We strongly recommend the users of MARRVEL to sign up in order to receive update notifications through the e-mail submission form at the bottom of the MARRVEL home page.

PROTOCOL:

1. How to begin a search

1.1 For the human gene and variant-based search, go to steps 1.1.1.-1.1.2. For human gene-based search (no variant input), go to step 1.2. For model organism gene-based search, refer to steps 1.3.1.-1.3.2.

1.1.1. Go to the home page of MARRVEL⁴ at <http://marrvel.org/>. Start by entering a human gene symbol. Ensure that the candidate gene names are listed below the input box with each character entry. If the search comes back negative, make sure the gene symbol used is up to date using the

HUGO Gene Nomenclature Committee website⁵ (HGNC; <https://www.genenames.org/>).

1.1.2. Enter a human variant. The search bar is compatible with two types of variant nomenclature: genome location similar to how variants are displayed on ExAC and GnomAD⁶ and transcript-based nomenclature according to HGVS guidelines. Examples of such formats are shown in grey text within the search box. For genomic location nomenclature, use the coordinates according to hg19/GRCh37. Proceed to step 2.

NOTE: If a search returns an error, the most common problems are either the gene symbol is not up to date or the variant nomenclature is incorrect. In those cases, the HGNC (<https://www.genenames.org/>), Mutalyzer⁷ (<https://www.mutalyzer.nl/>), and TransVar⁸ (<https://bioinformatics.mdanderson.org/transvar/>) websites are great resources to correct the error. HGNC provides official gene symbols and their aliases for all human genes.

1.1.3. If still encountering error messages after confirming the gene name is up to date, use Mutalyzer and TransVar to check and convert variant nomenclature.

1.1.4. In some situations, such as a very recent gene symbol change in HGNC, try using a synonym for the gene and please contact the MARRVEL operating team using the “Feedback” tab so to update the source data, as MARRVEL may not provide the correct information due a lag in data update.

1.2. Enter a human gene symbol and leave the human variant search bar blank. If an error is encountered, go to HGNC (<https://www.genenames.org/>) to check for the official gene symbol or try an older gene symbol.

1.3.1 Click on **Model Organisms Search** tab on the top banner (**Figure 1**) or go to <http://marrvel.org/model>. Select the model organism of choice and enter a model organism gene symbol. Click on the gene symbol as the name is autocompleted and then click **Search**. If the search result is negative, check the official gene symbol that is used in model organism databases (**Table 1**).

1.3.1. If the search result is still negative, access DIOPT (DRSC Integrative Ortholog Prediction Tool, https://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl) and HCOP (<https://www.genenames.org/tools/hcop/>) to assess if there are no good predicted orthologs for the gene of interest. DIOPT is an ortholog prediction search engine run by the DRSC (*Drosophila* RNAi Screening Center) and HCOP is a similar suite developed by HGNC.

NOTE: Additional searches using BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) may allow users to find orthologs that may be missed by prediction algorithms used in DIOPT and HCOP.

1.3.2. Click on the **MARRVEL** it at the bottom for the predicted human ortholog of choice. Check the **DIOPT score**⁹ and **Best score from Human gene to model organism?** for the selection of the human gene. Proceed to step 2.

NOTE: **DIOPT score**⁹ (https://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl) is a value of how many ortholog prediction algorithms predict a pair of genes in two organisms to be orthologous to one another. For more information about these values and the specific algorithms used to calculate this score, refer to Hu et al⁹. When **Best score from Human gene to model organism?** is **Yes**, it indicates that the human gene is more likely a true human orthologs of the gene of interest but there could be exceptions, especially when multiple human genes are orthologous to multiple model organism genes due to gene duplication events during evolution. If the gene of interest is a member of a complex gene family that have undergone divergent evolution in multiple species, users should identify a publication that has performed an extensive phylogenetic analysis of the gene family of interest to identify the most likely ortholog candidate gene.

2. How to interpret MARRVEL human genetics outputs for a gene and variant search

NOTE: On the results page, there are seven human databases that are displayed (**Table 1, Figure 1**). For each output box, there is an **External link button (small box with a diagonal arrow)** on the upper right-hand corner that will link to the original database for more details.

2.1. Click **OMIM** (Online Mendelian Inheritance in Man, <https://www.omim.org/>)¹⁰, the first database that is displayed.

NOTE: OMIM is a manually curated database that aggregates and summarizes information on genetic diseases and traits in the human.

2.1.1. Use the **Human Gene Description** box from OMIM for a short summary of what is known about the gene and gene product.

2.1.2. Use the **Gene-Phenotype Relationships** box to determine if this gene is a known disease-causing gene or not. This box provides manually curated known disease or phenotype associations with the gene of interest.

2.1.3. Use the **Reported Alleles from OMIM** box to get a list of pathogenic variants curated by OMIM.

NOTE: Since manual curation of a publication regarding a new disease gene discovery is necessary for any gene-disease association to appear in OMIM, some time lag and/or missed publications may lead to misconception. It is recommended that users perform PubMed (<https://www.ncbi.nlm.nih.gov/pubmed/>) searches to look into recent literature as well (See 4.1.2.). For additional information curated in OMIM, refer to Amberger^{10,11}.

2.2. Click **ExAC** (Exome Aggregation Consortium, <http://exac.broadinstitute.org/>)⁶ and **gnomAD** (genome Aggregation Database, <http://gnomad.broadinstitute.org/>), large population genomics databases based on WES and WGS of people who are selected to exclude severe pediatric diseases.

NOTE: ExAC contains ~60,000 WES whereas gnomAD contains ~120,000 WES and ~15,000 WGS. Both ExAC and gnomAD can be used as a control population database, especially for severe pediatric disorders, but its interpretation requires some degree of caution. In general, gnomAD can be considered as an updated and expanded version of ExAC since most cohorts that are included in ExAC is also included in gnomAD. However since there are some exceptions (see cohort information in <http://exac.broadinstitute.org/about> and <http://gnomad.broadinstitute.org/about>, respectively), MARRVEL displays data from both sources.

2.2.1. Use the **Control Population Gene Summary** box to obtain gene-level statistics such as the probability of finding the loss of function (LOF) alleles in the general population. This is called the pLI (probability of LOF Intolerance) score in ExAC and can be used to infer how likely a single copy of a LOF allele for a specific gene may cause a dominant disease through haplo-insufficient mechanisms.

NOTE: Looking at the pLI score of a gene has value, especially when dealing with dominant disorders that present as severe pediatric diseases associated with *de novo* variants. If a gene has a pLI score of 0.00, it means it is highly tolerant of LOF variants thus the gene unlikely cause disease via a dominant haploinsufficiency mechanism. This does not, however, necessarily rule out other dominant gain of function (GOF) or dominant negative mediated mechanisms may cause disease. In addition, genes that cause the recessive diseases may have low pLI scores since carriers are expected to be found in the general population. On the other hand, if a gene has a pLI score of 1.00, it is possible that the loss of one copy of this gene is detrimental for human health. Additional searches in websites such as DOMINO (<https://wwwfbm.unil.ch/domino/>) may also be used in combination to assess the likelihood of a variant in a specific gene causing a dominant disorder.

2.2.2. Use the next two boxes to obtain the allele frequencies of the variant of interest in ExAC and gnomAD, respectively to help interpret whether or not the variant may be pathogenic depending on if the patient has the dominant or recessive disease. This box will only be displayed when the user inputs variant information when initiating the search.

NOTE: If one hypothesizes a recessive disease scenario and the pLI score of the gene of interest is low, one should pay attention to the allele frequency listed here. Some geneticists may establish a cut-off point of 0.005 to 0.0001 as the maximum allele frequency for pathogenic variants that can cause a severe recessively inherited disease². On the other hand, if one hypothesizes a dominant disease scenario, it is less likely to find the identical or similar variant in a control population. Again, this requires caution because individuals with late-onset disorders, diseases with mild presentation, psychiatric disorders or diseases not screened by the ExAC/gnomAD researchers may be still included and the variant may still be a dominant pathogenic variant. Also, there have been some instances of variants linked to pediatric conditions found in a few individuals in these databases¹²⁻¹⁴, potentially due to incomplete penetrance or somatic mosaicism^{13,15,16}. In addition, although ExAC and gnomAD will display

variants that are found in a homozygous state, it will not indicate whether any of the variants are found in a compound heterozygous state. Finally, some variants found in these databases are tagged as low confidence due to technical challenges in sequencing (e.g. low sequence coverage, repetitive sequence). To look more carefully into these data sets, users are recommended to use the **external link** button to visit the original ExAC and gnomAD websites to gain additional information.

2.3. Click **Geno2MP** (Genotype to Mendelian Phenotype Browser, <http://geno2mp.gs.washington.edu/Geno2MP/>), a collection of WES-based data from the University of Washington Center for Mendelian Genetics. It contains about 9,600 exomes (as of 1/18/2019) of affected individuals and unaffected relatives with some phenotypic descriptions (**Figure 1**).

2.3.1. Use the **Disease population** box to obtain the allele frequency of the variant of interest in this cohort.

2.3.2. Use the **Gene-Phenotype Relationships** box to obtain HPO (human phenotype ontology)¹⁷ terms for the individuals with the variant of interest. This is one of many ways for one to look for patients that may have the same disease.

NOTE: If a gene of interest is suspected to be associated with a patient's disease and there are matches found in Geno2MP, additional important information may be present in the data source beyond what is displayed.

2.3.2.1. Click the **external link** button to the gene-specific page on Geno2MP, filter for mutations that are similar to those of the patient (e.g., missense, LOF), and carefully review the lists of variants. Take note of the variants with high CADD¹⁸ scores and click into the HPO profiles. For example, CADD scores higher than 20 are within the top 1% of all variants predicted to be deleterious, CADD scores that are higher than 10 are within the top 10%. HPO terms provide a standardized description of human phenotypes. Here, make sure to check if the variant was identified in an affected individual or in a relative.

2.3.2.2. If variants are found in patients that are affected in the same organ system as the patient, consider using the e-mail form to contact the physician that submitted these cases to Geno2MP using the feature provided on the Geno2MP website.

NOTE: Not all physicians respond to such queries, so one should explore other avenues of patient matchmaking. Other ways to gather a cohort of patients affected by the same diseases is to use tools such as GeneMatcher¹⁹ (<https://www.genematcher.org/>) and other databases that are part of the Matchmaker Exchange^{19,20} (<https://www.matchmakerexchange.org/>). See accompanying JoVE article for more information on matchmaking²¹.

2.4. Use the **ClinVar** (<https://www.ncbi.nlm.nih.gov/clinvar/>)²² database, supported by the National Institutes of Health (NIH), where researchers and clinicians submit variants with or

without determination of pathogenicity, for checking single nucleotide variants (SNV), small indels and larger copy number variations (CNV).

2.4.1. Use the top row to review a summary of the number of each type of variants reported in ClinVar (**Figure 1**).

2.4.2. Check the list of variants below in the box **Reported Alleles from ClinVar**.

NOTE: If a variant was included in the initial search, the highlighted variants in teal are all variants that include the genomic location of the variant of interest [including large CNVs, which are often labeled as; genomic coordinate...x1 (deletion) and ...x3 (duplication)].

2.5. Use **DGV**²³ (Database of Genomic Variants, <http://dgv.tcag.ca/dgv/app/home>) and **DECIPHER**²⁴ (Database of genomic variation and Phenotype in Humans using Ensembl Resources, <https://decipher.sanger.ac.uk/>), both collections of CNVs. DGV is the largest public-access collection of structural variants from more than 54,000 individuals. This database includes samples of reportedly healthy individuals, at the time of ascertainment, from up to 72 different studies. Similarly, the data displayed from DECIPHER includes common variants from the control population.

NOTE: Since MARRVEL does not have permission to display patient derived data from DECIPHER, users are encouraged to directly visit the DECIPHER website to access potentially pathogenic CNV information.

2.5.1. Click the **Copy Number Variation in Control Population (DGV Database)** box to obtain variants that contain the gene of interest. Information such as the size, subtype, and reference of the copy number variation can be found in the same box.

2.5.2. Click the **Common Copy Number Variants (DECIPHER Database)** box to obtain variants that contain the genomic location of the variant of interest. This information may help determine if the gene is duplicated or deleted in the control individuals.

NOTE: If the gene of interest is deleted in many individuals in the control population, it means that this gene is likely to be highly tolerant of LOF variants. Like low pLI scores, this suggests that a single copy loss of this gene is less likely to cause a severe disease via a haploinsufficiency mechanism. This does not, however, necessarily rule out other dominant gain of function or dominant negative mechanisms (e.g. antimorphic, hypermorphic and neomorphic alleles) caused by specific missense and truncation alleles. Possible limitations to these data include variation in source and method of the data acquired, lack of information regarding incomplete penetrance of pathogenic CNVs, and whether individuals developed certain diseases subsequent to data collection.

3. How to use model organism data in MARRVEL

3.1. Use the **Gene Function Table** to obtain the following information for eight model organisms including human (human, rat, mouse, zebrafish, *Drosophila*, *C elegans*, budding yeast and fission yeast):

3.1.1. **Gene name:** Since each gene name is hyperlinked to gene pages on respective model organism databases, click on these links to find out more about the phenotypic information and resources available for each model organism. For example on **FlyBase**²⁵ (<http://flybase.org/>), there will be a list of all alleles that have been generated, their respective phenotypes and the availability of each allele from public stock centers.

3.1.2. **PubMed link:** Click on the **PubMed link** to go to a list of publications that relates to the gene of interest in each organism. Without using these links, searching for the human gene directly in PubMed may lead to missing some publications that used an old gene alias to refer to the human gene. Similarly, model organism gene names may have fluctuated historically.

3.1.3. **DIOPT**⁹ score: Check this column for a score of how many ortholog prediction algorithms predict the gene is likely to be an ortholog of the human gene of interest. One may use a DIOPT score of 3 or above as a reasonable cut-off to identify solid ortholog candidates. However, there are cases where genuine orthologs only have a DIOPT score of 1 due to limited homology. At the top of the gene function table, un-check the “Show only best DIOPT score gene” box to display all candidates that typically include homologous genes that are not necessarily orthologs.

3.1.4. **Expression:** Check this column for the list of the tissues where the gene or protein of interest has been reported to be expressed in human or model organism databases. Human gene and protein expression data are from **GTEx**²⁶ (<https://gtexportal.org/>) and **Human Protein Atlas**²⁷ (<https://www.proteinatlas.org/>), respectively. Some have a button with pop-up links, such as for human and for fly that display the expression pattern using a heat map, whereas others are hyperlinked to respective model organism databases pages.

3.1.5. **Gene Ontology**²⁸ (GO) terms: Filter by **experimental evidence codes** and obtain from respective human or model organism databases. GO terms based on “computational analysis evidence codes” and “electronic annotation evidence codes” (predictions) are not displayed. Please visit each model organism website to gather this information if necessary.

3.1.6. Other links such as **Monarch Initiative**²⁹ (<https://monarchinitiative.org/>) and **IMPC**³⁰ (<http://www.mousephenotype.org/>): Use the **Monarch Initiative** hyperlink to navigate to the Phenogrid page for the specific human gene, a chart that provides a quick comparison between the phenotypes associated with the gene of interest to known human diseases and model organism mutants that have phenotypic overlaps. If a mouse gene has a knockout mouse made or planned by the International Mouse Phenotyping Consortium (IMPC), the “IMPC” links to the page that details the phenotype of the knockout mouse and its availability from public stock centers.

3.2. **Human Protein Domains:** Use the **human gene protein domains** box to obtain predicted

protein domains of the human gene. The data are derived from **DIOPT**, which uses **Pfam** (<https://pfam.xfam.org/>) and **CCD** (Conserved Domains Database, <https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>). A single residue maybe annotated more than once due to some overlap in domains annotated in the two sources.

3.3. Use the **Multiple Protein Alignment** box to obtain the amino acid multiple alignment generated by DIOPT⁹ which includes human (hs), rat (rn), mouse (mm), zebrafish (dr), fruit fly (dm), worm (ce), and yeasts (sc and sp). To highlight the amino acid of interest, scroll down to the bottom of the box and enter the amino acid numbers below and the amino acids of interest will be highlighted in teal. The alignment is provided by DIOPT and uses MAFFT aligner (Multiple alignment program for amino acid or nucleotide sequences, <https://mafft.cbrc.jp/alignment/software/>³¹).

NOTE: If the amino acid that is highlighted based on the number is not the one expected, it may be due to different splicing isoforms used for the alignment. In principle, DIOPT uses the longest isoform to display in this box. Also, for segments of genes that are not well conserved, alignment of multi-species sequences using default parameters may not be optimal. We recommend using other websites and software like Clustal Omega and ClustalW/X (<http://www.clustal.org/>)³² to optimize the alignment parameters and matrices accordingly.

REPRESENTATIVE RESULTS:

Human geneticists and model organism scientists each use MARRVEL in distinct ways, each with different desired outcomes. Below are three vignettes of possible uses for MARRVEL.

Evaluating pathogenicity of a variant in a dominant disease

Most of the users that visit MARRVEL use this website to analyze the likelihood that a rare human variant may cause a certain disease. For example, a missense (17:59477596 G>A, p.R20Q) variant in *TBX2* was found to segregate in an autosomal dominant manner in a small family with dysmorphic features and cleft palate, cardiac defects, skeletal and digit abnormalities, thyroid-related phenotypes, and immune defects¹². The mother and two children affected with these symptoms carried the variant, whereas the father did not. The 9-year-old son had the most severe phenotype, whereas the 36-year-old mother and the 6-year-old daughter had milder forms of this disease. To assess whether this variant is likely pathogenic, one can start a MARRVEL search by entering the gene and variants on the starting page on <http://MARRVEL.org>. Note that the variant search bar requires the removal of **Chr** in front of the variant if this is listed in the original clinical report to indicate “Chromosome”. At the time of the original study, the results page showed that there is no OMIM phenotype associated with this gene, and this variant is found only once in gnomAD but not in ExAC, ClinVar, or Geno₂MP. One may think this identification of one individual may be evidence against p.R20Q being a pathogenic variant, but it is important to note that the mother of the family exhibited a mild form of the disease. A variant found in 1/~150,000 individual is indeed a very rare variant and the identification of an individual with the identical variant may be explained by reduced expressivity or penetrance. In the Gene Function table, it is often helpful to check if the gene is expressed in relevant tissues in humans (via GTEx and Protein Atlas) in reference to the phenotypes of the patient. In this case, the

expression pattern matches since the patient has phenotypes in multiple tissues and the gene is also widely expressed, including cardiac, and immune-related organs.

Based on model organism information displayed in MARRVEL, one can quickly see that the gene is conserved from *C. elegans* and *Drosophila* to human and the amino acid of interest, p.R20 is also highly conserved throughout evolution as shown in **Figure 2** (note that rat *Tbx2* does not align well in this region, likely due to the transcript that is used for alignment). Phenotypic information in mouse and zebrafish indicates that this gene affects development or function of a number of tissues including the cardiovascular system, craniofacial/palate, and digits. In sum, these data suggest that this variant is possibly pathogenic and further functional study is valuable. Considering that the gene and variant are conserved in organisms like *C. elegans* and *Drosophila*, functional studies in invertebrate animals will be faster and cheaper compared to performing the same experiment in vertebrate model organisms such as zebrafish, mouse and rat. Please see the accompanying article by Harnish et al.²¹ regarding how we designed and performed functional assays for this case¹². The involvement of this gene/variant in this family's disease was further strengthened by identification of an unrelated 8-year-old male patient with overlapping phenotypes with a *de novo* missense variant in the same gene using GeneMatcher. The variants in the two families were both found to be functional using experiments in *Drosophila*, further supporting the pathogenicity of the rare variants in *TBX2*. The disease has recently been curated as 'Vertebral anomalies and variable Endocrine and T-cell Dysfunction (VETD, OMIM #618223)' in OMIM. See **Figure 3** for entire output for *TBX2* 17:59477596 G>A.

Evaluating pathogenicity of a variant in a recessive disease

There are significant differences between analyzing human variants in dominant and recessive diseases. For example, pLI score, minor allele frequency, and presence of deletions in the control population become less important because two alleles are necessary to reveal any phenotype.

One example of analysis of a recessive disease is detailed in Yoon et al³³ and Wang et al⁴ which is summarized here. A 15-year-old girl exhibited developmental delay, microcephaly, ataxia, motor impairment, hypotonia, language impairments, brain abnormalities, and hypoplasia of the corpus callosum³³. The proband, her unaffected parents, and an unaffected sibling received WES. After filtering for variants that were both unique to the proband and rare in the population, variants in 13 different genes remained. Manual filtering and analysis of the 13 candidates by following the protocol described here resulted in the prioritization of one specific variant in *OGDHL* as a good candidate for functional studies. The key pieces of information that led to prioritizing p.S778L in *OGDHL* (10:50946295 G>A) over other variants include: (1) no previous disease association in OMIM, (2) variant not found in control populations, (3) gene ontology associated with microtubule and mitochondria, two systems that have many links to neurological disorders^{34,35}, (4) highly expressed in human cerebellum, a tissue severely affected in this patient, and (5) the variant of interest affecting a highly conserved amino acid (from yeast to human) and located within the catalytic domain⁴. pLI score for this gene is 0.00 but this doesn't affect the prioritization of this variant/gene for this case since we are suspecting a recessive mode of inheritance and that carriers of deleterious variants in this gene can present in the general population. See **Figure 4** for MARRVEL output for *OGDHL* 10:50946295 G>A.

Model organism studies performed in parallel showed that loss of *Ogdh* (also referred to as *Nc73EF*), the *Drosophila* ortholog of *OGDHL*, in the nervous system exhibits a neurodegenerative phenotype consistent with the proband's neurological disorder³³. Functional studies in *Drosophila* showed that the variant of interest (p.S778L) affects protein function, making this a strong candidate gene for this disease. Since then, this information about a potential pathogenic variant in *OGDHL* linked to a novel neurological disorder has been incorporated into OMIM (<https://www.omim.org/entry/617513>) very recently but have not yet been assigned a disease-phenotype number because only one case has been reported as of January 2019.

Is the human ortholog of a model organism gene of interest associated with genetic diseases?

Many model organism researchers may be interested to see whether the human ortholog of their gene of interest may have links to genetic diseases. In this example, we will search whether the human ortholog(s) of the fly *Notch* (N) gene has any relevance to genetic diseases. To do this, we will start with performing a "Model Organisms Search (1.3.1.-1.3.2.)" and select "*Drosophila melanogaster*" as the species name and "N" as the model organism gene name. The four predicted human orthologs for this fly gene will be displayed in the results window as *NOTCH1*, *NOTCH2*, *NOTCH3*, and *NOTCH4*. The four genes have different DIOPT scores (10/12 for *NOTCH1*, 8/12 for *NOTCH2* and *NOTCH3*, 5/12 for *NOTCH4*) due to the degree of homology between fly *N* and each human gene. Considering the "Best score from Human gene to Fly" is listed as "Yes" for all four genes, the reverse search from each human gene picks up the fly *N* gene as the most likely ortholog candidate. Indeed, the four human *NOTCH* genes are thought to have arisen from a single *Notch* gene during the two rounds of whole genome duplication events that happened in the vertebrate lineage after splitting from the invertebrate lineage³⁶. By clicking the "MARRVEL it" buttons for each human gene, one can obtain the human gene-based outputs for *NOTCH1-4*. On the results page of each gene, the top boxes for OMIM indicate that while *NOTCH1*, *2*, and *3* are associated with genetic diseases, *NOTCH4* is currently not associated with any human diseases. Note that there have been debates on whether variants in *NOTCH4* are associated with schizophrenia based on genome-wide association studies (GWAS)^{37,38}. Since OMIM generally does not curate GWAS data with some exceptions (e.g. *APOE*, *PTPN22*), this information is not available from the OMIM window. Similarly, since OMIM does not generally curate cancer-associated somatic mutation information, information on whether somatic mutations in these genes are associated with certain cancer types will not be listed with a few exceptions (e.g. *TP53*, *RB1*, *BRCA1*). By clicking the **PubMed** or **Monarch** box, one can identify some disease related papers that are not curated in OMIM. See **Figure 5** for the entire MARRVEL output for the fly gene *N* and human gene *NOTCH4*.

FIGURE AND TABLE LEGENDS:

Figure 1. A Representative output from a MARRVEL search. This specific example is showing a gene/variant search for "TBX2/17:59477596 G>A" (<http://marrvel.org/search/pair/TBX2/17:59477596%20G%3EA>). Sidebar on the left supports navigations through the data output. Note the "external link" signs here provide links to the appropriate pages of the UCSC genome browser (<https://genome.ucsc.edu/>). The tabs on the top

allow one to perform model organism gene-based searches, obtain additional information about MARRVEL and provide user feedbacks. The ‘Search Results’ panels display gene and variant information from the sources indicated in the image.

Figure 2. Summary of the model organism ortholog table and multi-species alignment for *TBX2*.

A) MARRVEL selects the top ortholog candidate for each species based on the DIOPT tool. For example, a DIOPT score of 10/12 shown for the *Drosophila bi* gene means 10 out of 12 orthology prediction programs used by DIOPT predicted that *bi* is the most likely fly ortholog of human *TBX2*. Since 25% of genes are duplicated in zebrafish compared to human, MARRVEL displays two paralogous genes (in this case *tbx2a* and *tbx2b*) when this is applicable. **B)** Snapshot of the multi-species alignment window. By selecting a specific organism [in this case human (hs)] and entering the amino acid of interest, one can highlight the specific amino acid in teal. In this example, p.R20 of human *TBX2* seems to be conserved in mouse (mm1), both zebrafish orthologs (dr1 and dr2), *Drosophila* (dm1) and *C. elegans* (ce1). Rat *Tbx2* does not seem to align well compared to other species, most likely due to the isoform used by the DIOPT to perform the multi-species alignment.

Figure 3: Entire output for *TBX2* 17:59477596 G>A.

Figure 4: MARRVEL output for *OGDHL* 10:50946295 G>A.

Figure 5: MARRVEL output for the fly gene *N* and human gene *NOTCH4*.

Table 1. List of Data Sources for MARRVEL. All databases where MARRVEL obtains data from are listed in this table. For each database, we list the type of database, URL/Link, rationale for including in MARRVEL, and primary references.

DISCUSSION:

Critical steps in this protocol include the initial input (steps 1.1-1.3) and subsequent interpretation of the output. The most common reason why search results are negative is because of the many ways that a gene and/or variant can be described. While MARRVEL is updated on a scheduled basis, these updates may cause disconnects between the different databases that MARRVEL links to. Thus, the first step in troubleshooting is invariably checking to see if alternative names of the gene or variant will lead to a successful search result. If it still cannot be resolved, please send a message to the development team using the feedback form in <http://marrvel.org/message>.

One limitation to MARRVEL is that it does not yet include all the useful databases necessary for gene and variant analysis. For example, pathogenicity prediction algorithms such as CADD¹⁸ are not currently provided. Similarly, protein structure information and protein-protein interaction information that may also provide structural and functional links to known disease-causing variants in genes are not currently displayed in MARRVEL. In our next major update, we plan to integrate this information into MARRVEL, in addition to incorporating more phenotypic information from model organism websites, IMPC, Monarch Initiative and Alliance of Genome Resources (AGR, <https://www.alliancegenome.org/>). Since MARRVEL was designed to facilitate

rare disease research, the program currently focuses on germline variants and does not provide access to somatic variant information. No cancer genetics related databases are integrated as of publication of this protocol. As MARRVEL is actively being developed and upgraded, we highly appreciate feedback, and strongly encourage the existing users to sign up for newsletters on <http://marrvel.org/message> for any future additional databases that become integrated.

Although data from MARRVEL can be used to prioritize variants that may be pathogenic. However, in order to demonstrate pathogenicity, one will need to identify other patients with similar genotypes and phenotypes or perform functional studies to provide solid evidence that the variant of interest has functional consequences that are relevant to the disease condition. For more information on additional information outside of MARRVEL that may be useful to judge if a variant is worth experimentally investigating in the model organism, please refer to the accompanying article Harnish *et al*²¹. In order to take the next steps in using model organisms to study human variants, human geneticists and model organism researchers must be able to connect and collaborate. GeneMatcher and other genomic consortia that are part of the Matchmaker Exchange consortium are resources that facilitate this next step. If the users reside in Canada, one can also register in the Rare Disease Models and Mechanisms Network (RDMM, <http://www.rare-diseases-catalyst-network.ca/>) to identify clinicians and/or model organism researchers that are willing to collaborate³⁹. Japan (J-RDMM, <https://irudbeyond.nig.ac.jp/en/index.html>), Europe (RDMM-Europe, <http://solve-rd.eu/rdmm-europe/>), and Australia (Australian Functional Genomics Network: <https://www.functionalgenomics.org.au/>) have recently adopted the Canadian RDMM model to facilitate similar collaborations within their countries/regions. Furthermore, by using tools such as BioLitMine (<https://www.flyrnai.org/tools/biolitmine/web/>) one can search for potential collaborators among Principal Investigators who have previously worked on the gene of interest.

Lastly, in addition to MARRVEL, there are a number of other cross-species data mining tools available including Gene2Function⁴⁰ (<http://www.gene2function.org/>), Monarch Initiative²⁹ (<https://monarchinitiative.org/>) and Alliance of Genome Resources (AGR, <https://www.alliancegenome.org/>). While Gene2Function provides access to cross-species data and Monarch Initiative provides phenotypic comparisons, MARRVEL has a larger emphasis on human variants and linking human genomic data with model organisms. AGR is an initiative that involves six model organism databases and the Gene Ontology Consortium that integrates data from different database in a uniform way to increase the accessibility of data accumulated by each database. These resources are complementary, and users should understand the strengths of each database to navigate the vast amount of knowledge that has been accumulated by researchers in the communities. As MARRVEL development continues, we plan to include more databases that are relevant to studying human variants in model organisms. The overarching goal of MARRVEL is to provide an easily accessible way for clinicians and researchers alike to analyze human genes and variants for further study by integrating useful information while keeping the interface as simple as we can.

ACKNOWLEDGMENTS:

We thank Drs. Rami Al-Ouran, Seon-Young Kim, Yanhui (Claire) Hu, Ying-Wooi Wan, Naveen

Manoharan, Sasidhar Pasupuleti, Aram Comjean, Dongxue Mao, Michael Wangler, Hsiao-Tuan Chao, Stephanie Mohr, and Norbert Perrimon for their support in the development and maintenance of MARRVEL. We are grateful to Samantha L. Deal and J. Michael Harnish for their input on this manuscript.

The initial development of MARRVEL was supported in part by the Undiagnosed Diseases Network Model Organisms Screening Center through the NIH Commonfund (U54NS093793) and through the NIH Office of Research Infrastructure Programs (ORIP) (R24OD022005). JW is supported by the NIH Eunice Kennedy Shriver National Institute of Child Health & Human Development (F30HD094503) and The Robert and Janice McNair Foundation McNair MD/PhD Student Scholar Program at BCM. HJB is further supported by the NIH National Institute of General Medical Sciences (R01GM067858) and is an Investigator of the Howard Hughes Medical Institute. ZL is supported by the NIH National Institute of General Medical Science (R01GM120033), National Institute of Aging (R01AG057339), and the Huffington Foundation. SY received additional support from the NIH National Institute on Deafness and other Communication Disorders (R01DC014932), the Simons Foundation (SFARI Award: 368479), the Alzheimer's Association (New Investigator Research Grant: 15-364099), Naman Family Fund for Basic Research and Caroline Wiess Law Fund for Research in Molecular Medicine.

DISCLOSURES:

The authors have nothing to disclose.

REFERENCES

- 1 Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *New England Journal of Medicine*. **369** (16), 1502-1511, (2013).
- 2 Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*. **17** (5), 405-424, (2015).
- 3 MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature*. **508** (7497), 469-476, (2014).
- 4 Wang, J. *et al.* MARRVEL: Integration of Human and Model Organism Genetic Resources to Facilitate Functional Annotation of the Human Genome. *American Journal of Human Genetics*. **100** (6), 843-853, (2017).
- 5 Povey, S. *et al.* The HUGO Gene Nomenclature Committee (HGNC). *Human Genetics*. **109** (6), 678-680, (2001).
- 6 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. **536** (7616), 285-291, (2016).
- 7 Wildeman, M., van Ophuizen, E., den Dunnen, J. T. & Taschner, P. E. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Human Mutation*. **29** (1), 6-13, (2008).
- 8 Zhou, W. *et al.* TransVar: a multilevel variant annotator for precision genomics. *Nature Methods*. **12** (11), 1002-1003, (2015).
- 9 Hu, Y. *et al.* An integrative approach to ortholog prediction for disease-focused and other

functional studies. *BMC Bioinformatics*. **12** 357, (2011).

10 Amberger, J. S. & Hamosh, A. Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes. *Current Protocols in Bioinformatics*. **58** 1 2 1-1 2 12, (2017).

11 Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research*. **47** (D1), D1038-D1043, (2019).

12 Liu, N. *et al.* Functional variants in TBX2 are associated with a syndromic cardiovascular and skeletal developmental disorder. *Human Molecular Genetics*. **27** (14), 2454-2465, (2018).

13 Ropers, H. H. & Wienker, T. Penetrance of pathogenic mutations in haploinsufficient genes for intellectual disability and related disorders. *European Journal of Medical Genetics*. **58** (12), 715-718, (2015).

14 Shashi, V. *et al.* De Novo Truncating Variants in ASXL2 Are Associated with a Unique and Recognizable Clinical Phenotype. *American Journal of Human Genetics*. **100** (1), 179, (2017).

15 Chen, R. *et al.* Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nature Biotechnology*. **34** (5), 531-538, (2016).

16 Halvorsen, M. *et al.* Mosaic mutations in early-onset genetic diseases. *Genetics in Medicine*. **18** (7), 746-749, (2016).

17 Kohler, S. *et al.* The Human Phenotype Ontology in 2017. *Nucleic Acids Research*. **45** (D1), D865-D876, (2017).

18 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*. **47** (D1), D886-D894, (2019).

19 Sobreira, N., Schiettecatte, F., Valle, D. & Hamosh, A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Human Mutation*. **36** (10), 928-930, (2015).

20 Sobreira, N. L. M. *et al.* Matchmaker Exchange. *Current Protocols in Human Genetics*. **95** 9 31 31-39 31 15, (2017).

21 Harnish, M., Deal, S., Wangler, M. & Yamamoto, S. In vivo functional study of disease-associated rare human variants using *Drosophila*. *Journal of Visualized Experiments*. , (2019).

22 Harrison, S. M. *et al.* Using ClinVar as a Resource to Support Variant Interpretation. *Current Protocols in Human Genetics*. **89** 8 16 11-18 16 23, (2016).

23 MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*. **42** (Database issue), D986-992, (2014).

24 Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*. **84** (4), 524-533, (2009).

25 Thurmond, J. *et al.* FlyBase 2.0: the next generation. *Nucleic Acids Research*. **47** (D1), D759-D765, (2019).

26 Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot

- analysis: multitissue gene regulation in humans. *Science*. **348** (6235), 648-660, (2015).
- 27 Ponten, F., Jirstrom, K. & Uhlen, M. The Human Protein Atlas--a tool for pathology. *Journal of Pathology*. **216** (4), 387-393, (2008).
- 28 The Gene Ontology, C. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*. 10.1093/nar/gky1055, (2018).
- 29 Mungall, C. J. *et al.* The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*. **45** (D1), D712-D722, (2017).
- 30 Meehan, T. F. *et al.* Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. *Nature Genetics*. **49** (8), 1231-1238, (2017).
- 31 Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform*. 10.1093/bib/bbx108, (2017).
- 32 Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*. **27** (1), 135-145, (2018).
- 33 Yoon, W. H. *et al.* Loss of Nardilysin, a Mitochondrial Co-chaperone for alpha-Ketoglutarate Dehydrogenase, Promotes mTORC1 Activation and Neurodegeneration. *Neuron*. **93** (1), 115-131, (2017).
- 34 Deal, S. & Yamamoto, S. Unraveling novel mechanisms of neurodegeneration through a large-scale forward genetic screen in Drosophila (In Press). *Frontiers in Genetics* **9**700, (2019).
- 35 Matamoros, A. J. & Baas, P. W. Microtubules in health and degenerative disease of the nervous system. *Brain Research Bulletin*. **126** (Pt 3), 217-225, (2016).
- 36 Theodosiou, A., Arhondakis, S., Baumann, M. & Kossida, S. Evolutionary scenarios of Notch proteins. *Molecular Biology and Evolution*. **26** (7), 1631-1640, (2009).
- 37 Shayevitz, C., Cohen, O. S., Faraone, S. V. & Glatt, S. J. A re-review of the association between the NOTCH4 locus and schizophrenia. *American Journal of Medical Genetics. Part B: Neuropsychiatric Genetics*. **159B** (5), 477-483, (2012).
- 38 Wang, Z. *et al.* A review and re-evaluation of an association between the NOTCH4 locus and schizophrenia. *American Journal of Medical Genetics. Part B: Neuropsychiatric Genetics*. **141B** (8), 902-906, (2006).
- 39 Oriel, C. & Lasko, P. Recent Developments in Using Drosophila as a Model for Human Genetic Disease. *International Journal of Molecular Sciences*. **19** (7), (2018).
- 40 Hu, Y., Comjean, A., Mohr, S. E., FlyBase, C. & Perrimon, N. Gene2Function: An Integrated Online Resource for Gene Function Discovery. *G3 (Bethesda)*. **7** (8), 2855-2858, (2017).

The image displays the MARRVEL1.1 web application interface, which is designed for genomic and clinical data analysis. The interface is divided into several key sections:

- Top Navigation Bar:** Contains the application name "MARRVEL1.1" and links for "Human Search", "Model Organisms Search", "About", "FAQ", and "Feedback".
- Side Bar (Left):** A dark grey sidebar with a green border containing the following sections:
 - INPUT:** Fields for "Variant" (Chr17:59477596 G>A) and "Gene" (TBX2).
 - DATABASES:** A list of databases including OMIM, ExAC / Geno2MP, ClinVar, and DGV / DECIPHER.
 - MODEL ORGANISMS:** Options for "Predicted Orthologs", "Protein Domain", and "Protein Alignment".
 - Bottom:** A "New search" button.
- Tabs (Center):** A central area with multiple tabs, each representing a different data source or analysis tool. Red boxes highlight specific tabs, and red arrows point to their corresponding labels on the right.
- Search Results (Right):** A list of labels corresponding to the highlighted tabs, including OMIM, ExAC/gnomAD, Geno2MP, ClinVar, DGV, DECIPHER, Gene function and expression information (Human & MO), Protein Domains, and Multi-species protein alignment.

A

Species	Gene Symbol	DIOPT Score
Human (hs)	<i>TBX2</i>	-
Mouse (mm)	<i>Tbx2</i>	13/13
Rat (dn)	<i>Tbx2</i>	7/11
Zebrafish (dr)	<i>tbx2b</i>	11/12
	<i>tbx2a</i>	9/12
<i>Drosophila</i> (dm)	<i>bi</i>	10/12
<i>C elegans</i> (ce)	<i>tbx-2</i>	8/12
Budding yeast (sc)	-	-
Fission yeast (sp)	-	-

B

Multiple Protein Alignment[?]

DIOPT V6

hs1

MR-----EP-ALAASAMAYHPF--HAPRPADF-----

mm1

MR-----EP-ALAASAMAYHPF--HAPRPADF-----

rn1

dr1

MR-----DP-VFTGTAMAYHPF--HAHRPTDF-----

dr2

MR-----DP-VFTANAMAYHPF--HAHRPADF-----

dm1

MRYDVQELLFHQSAEDPFARFANGMAYHPFLQLTQRPTDFSVSSLLTAGSNNNNSG

ce1

-----MAFNPF--ALGRP-DLLL-PFMGAG-----

Highlight from

20

to

20

for

☒ hs

☐ rn

☐ mm

☐ dr

☐ dm

☐ ce

INPUT

Variant

Chr17:59477596 G>A

Gene

TBX2

DATABASES

OMIM

ExAC / Geno2MP

ClinVar

DGV / DECIPHER

MODEL ORGANISMS

Predicted Orthologs

Protein Domain

Protein Alignment

Human Gene Description (OMIM)

T-BOX 2; TBX2

MIM number: 600747

Description: The TBX2 gene encodes a transcription factor that belongs to the family of T-box factor proteins that bind DNA, including TBX1 (602054) and TBX3 (601621). TBX2 is expressed in a variety of tissues and organs during embryogenesis (summary by Harrelson et al., 2004).

Gene-Phenotype Relationships

OMIM

Phenotype	Phenotype MIM number	Inheritance
Vertebral anomalies and variable endocrine and T-cell dysfunction	618223	Autosomal dominant

Reported Alleles From OMIM

TBX2

Phenotype	Mutation	dbSNP
VERTEBRAL ANOMALIES AND VARIABLE ENDOCRINE AND T-CELL DYSFUNCTION	TBX2, ARG20GLN	rs1364709483
VERTEBRAL ANOMALIES AND VARIABLE ENDOCRINE AND T-CELL DYSFUNCTION	TBX2, ARG305HIS	rs1555877071

Control Population Gene Summary

TBX2 (ExAC Gene Table)

Constraint from ExAC	Expected no. variants	Observed no. variants	ConstraintMetric
Synonymous	170.6	96	z=3.54
Missense	289.1	172	z=3.37
LoF	14.5	1	pLI=0.96
CNV	4.2	0	z=0.92

Disease Population (Geno2MP Database)

Chr17:59477596 G>A

No matches found

Gene-Phenotype Relationships (Geno2MP)

Chr17:59477596 G>A

No matches found

Population Allele Frequencies (GnomAD Database)

Chr17:59477596 G>A

Allele Count	2
Allele Number	179042
Homozygous count	0
Allele frequency	0.000011171
Gene	RP11-332H18.4

Benign

Likely Benign

Pathogenic

Likely Pathogenic

Risk Factor

Uncertain Significance

Conflicting Interpretations

0

0

18

0

0

2

0

Reported Alleles From ClinVar

TBX2 / Chr17:59477596 G>A

Variation	Location	Condition(s)	Frequency	Clinical Significance	Review Status
NC_000017.10:g.17711738_217748468del200036731	Chr17:17711738-217748468	Smith-Magenis syndrome		Pathogenic	criteria provided, single submitter
NC_000017.10:g.(?_56321134)_(62080001_?)dup	Chr17:56321134-62080001	See cases		Pathogenic	criteria provided, single submitter
GRCh38/hg38 17q23.2(chr17:61331901-61424019)x3	Chr17:59409262-59501380	See cases		Uncertain significance	no assertion criteria provided
GRCh38/hg38 17q23.1-25.1(chr17:36449220-75053130)x3	Chr17:57595736-73049225	See cases		Pathogenic	no assertion criteria provided
GRCh37/hg19 17q23-24(chr17:59209629-64222315)x3	Chr17:59209629-64222315	See cases		Pathogenic	criteria provided, single submitter
GRCh38/hg38 17q23.2-25.3(chr17:36449220-83086677)x3	Chr17:58617905-81044553	See cases		Pathogenic	criteria provided, single submitter
GRCh38/hg38 17q23.1-23.2(chr17:60043448-62148729)x1	Chr17:58120809-60226090	See cases		Pathogenic	criteria provided, single submitter
GRCh37/hg19 17q23(chr17:58934659-60395826)x1	Chr17:58934659-60395826	See cases		Pathogenic	criteria provided, single submitter

Copy Number Variation In Control Population (DGV Database)

TBX2

Show 10 entries

Search:

Position	Size	Type	Subtype	Frequency	Gain	Loss	Sample Size	References
59323466 17 59562984	239518	CNV	loss	0.02105263	0	2	95	17160897
59454601 17 59495300	40699	CNV	deletion	1.00000000	0	1	1	24416366
59456589 17 59491281	34692	CNV	gain	0.00049358	1	0	2026	19592680
59476600 17 59478200	1600	CNV	deletion	0.50000000	0	1	2	24896259
59486746 17 59487226	480	CNV	gain	0.09677419	3	0	31	20364138

Showing 1 to 5 of 5 entries

Previous1Next

Common Copy Number Variants (DECIPHER Database)

17:59477596

No matches found

Gene Function Table

TBX2

Show only best DIOPT v6 score gene

	Homolog	DIOPT Score	Expression	Molecular function	Cellular component	Biological process
Human	TBX2 PubMed Monarch	NA	<ul style="list-style-type: none">adrenal glandbreastcaudatecerebellumcerebral cortexcervix, uterinecolonepididymisfallopian tubekidneynasopharynxplacentarectumskeletal muscletestisthyroid glandurinary bladder Show all / GTEx	<ul style="list-style-type: none">contributes_toRNA polymerase II core promoter proximal region sequence-specific DNA bindingtranscriptional repressor activity, RNA polymerase II core promoter proximal region sequence-specific bindingprotein bindingcontributes_tosequence-specific DNA binding	<ul style="list-style-type: none">nucleus	<ul style="list-style-type: none">negative regulation of transcription from RNA polymerase II promotercell agingpositive regulation of cell proliferationnegative regulation of transcription, DNA-templatedcellular senescence
Rat	Tbx2 PubMed	7/11	Show all (9)	No term based on experiment	No term based on experiment	No term based on experiment
Mouse	Tbx2 PubMed IMPC	13/13	<ul style="list-style-type: none">extraembryonic componentembryo ectodermcardiovascular systembranchial archesalimentary system Show 12 more Open on MGI	<ul style="list-style-type: none">DNA bindingtranscription factor activity, sequence-specific DNA bindingprotein binding	<ul style="list-style-type: none">nucleustranscription factor complex	<ul style="list-style-type: none">heart morphogenesisoutflow tract septum morphogenesisoutflow tract morphogenesisendocardial cushion morphogenesisregulation of transcription from RNA polymerase II promoter involved in myocardial precursor cell differentiation Show 15 more
Zebrafish	tbx2b PubMed	11/12	<ul style="list-style-type: none">braincranial ganglionepiphysisretinal ganglion cell layerretinal neural layer Show 51 more Open on ZFIN	No term based on experiment	No term based on experiment	<ul style="list-style-type: none">heart loopingcardiac chamber developmentcell adhesionbrain developmentregulation of heart contraction Show 15 more
	tbx2a PubMed	9/12	<ul style="list-style-type: none">cardiac ventriclehindbrainhypothalamusolfactory placodeoptic vesicle Show 22 more Open on ZFIN	<ul style="list-style-type: none">protein binding	No term based on experiment	<ul style="list-style-type: none">cardiac chamber developmentembryonic heart tube developmentpharyngeal system development
Drosophila	bl PubMed	10/12	<ul style="list-style-type: none">EyeBrain Show all (25)	<ul style="list-style-type: none">DNA bindingprotein binding	<ul style="list-style-type: none">nucleus	<ul style="list-style-type: none">negative regulation of transcription from RNA polymerase II promotercompound eye morphogenesiswing disc morphogenesisimaginal disc-derived wing morphogenesisorgan growth Show 1 more
C Elegans	tbx-2 PubMed	8/12	Open on WormBase	<ul style="list-style-type: none">enzyme binding	<ul style="list-style-type: none">nematode larval developmentnucleuscytoplasm	<ul style="list-style-type: none">regulation of transcription from RNA polymerase II promoterregulation of protein localizationlocomotion

Human Gene Protein Domains

TBX2

DIOPT v6

Index	Domain name	Domain start	Domain stop	Domain description	Protein ID	External ID
hs1	TBOX	106	289	T-box DNA binding domain of the T-box family of transcriptional regulators. The T-box family is an ancient group that appears to play a critical role in development in all animal species. These genes were uncovered on the basis of similarity to the DNA...;	NP_005985.3	CDD:29144
hs1	Repression domain 1 (RD1)	518	601	propagated from UniProtKB/Swiss-Prot (Q13207.3)	NP_005985.3	
hs1	TBX	305	382	T-box transcription factor; pfam12598	NP_005985.3	CDD:204975

Multiple Protein Alignment

TBX2

DIOPT v6

hs1

mm1

rn1

dr1

dr2

dm1

ce1

hs1

mm1

rn1

dr1

dr2

dm1

ce1

Highlight from

Integer (1~n)

to

Integer (1~n)

for

hs

mm

rn

dr

dm

ce

Human Gene Description (OMIM®)

No summary description provided by OMIM

Gene-Phenotype Relationships®

OMIM

No OMIM gene-phenotype relationships found

Reported Alleles From OMIM®

OGDHL

No OMIM allelic variants found

Control Population Gene Summary

OGDHL (ExAC Gene Table)

Constraint from ExAC	Expected no. variants	Observed no. variants	ConstraintMetric ⓘ
Synonymous	189.3	196	z=-0.30
Missense	417.8	414	z=0.09
LoF ⓘ	31.1	13	pLI=0.00
CNV ⓘ	9.4	23	z=-1.07

Population Allele Frequencies (ExAC® Database)

Chr10:50946295 G>A

Allele count	1
Allele number	121070
Homozygous count	0
Allele frequency	0.00000825968
Gene	OGDHL

Population Allele Frequencies (GnomAD® Database)

Chr10:50946295 G>A

Allele Count	3
Allele Number	277128
Homozygous count	0
Allele frequency	0.000010825
Gene	OGDHL

Disease Population (Geno2MP® Database)

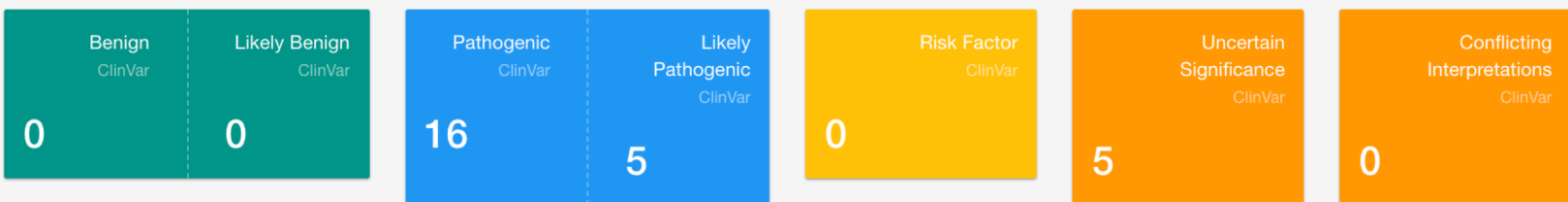
Chr10:50946295 G>A

No matches found

Gene-Phenotype Relationships (Geno2MP®)

Chr10:50946295 G>A

No matches found



Reported Alleles From ClinVar

OGDHL / Chr10:50946295 G>A

Variation	Location	Condition(s)	Frequency	Clinical Significance	Review Status
NC_000010.10:g.(?_46491169)_(51081560_?)del	Chr10:46491169-51081560	See cases		Pathogenic	criteria provided, single submitter
NC_000010.10:g.(?_46242057)_(51595050_?)dup	Chr10:46242057-51595050	See cases		Pathogenic	criteria provided, single submitter
GRCh38/hg38 10q11.21-21.3(chr10:42112187-67400675)x3	Chr10:42607635-69160433	See cases		Pathogenic	no assertion criteria provided
GRCh38/hg38 10q11.22-11.23(chr10:45710242-49929364)x1	Chr10:46205690-51330432	See cases		Pathogenic	no assertion criteria provided
GRCh38/hg38 10q11.22-11.23(chr10:45931517-50035809)x3	Chr10:46966533-51795569	See cases		Uncertain significance	no assertion criteria provided
GRCh38/hc38 10q11.22-11.23(chr10:45788078-	Chr10:46476965-	See cases		Uncertain	no assertion criteria provided

Copy Number Variation In Control Population (DGV® Database)

OGDHL

Show 10 entries

Search:

Position	Size	Type	Subtype	Frequency	Gain	Loss	Sample Size	References
46692238 10 51472468	4780230	OTHER	Inversion	0.00000000	0	0	9	18451855
49201820 10 51136375	1934555	CNV	loss	0.00003438	0	1	29084	25217958
50928727 10 51019190	90463	CNV	gain	0.00003438	1	0	29084	25217958
50949446 10 50953144	3698	CNV	duplication	0.00540541	1	0	185	20981092
50950375 10 50950772	397	CNV	insertion	1.00000000	1	0	1	19546169

Showing 1 to 5 of 5 entries

Previous1Next

Common Copy Number Variants (DECIPHER® Database)

10:50946295

No matches found

Gene Function Table

OGDHL

Show only best DIOPT v6 score gene

	Homolog	DIOPT Score ⓘ	Expression	Molecular function	Cellular component	Biological process
Human	OGDHL PubMed Monarch	NA	<ul style="list-style-type: none">cerebellumkidneyparathyroid gland Show all / GTEx	<ul style="list-style-type: none">protein binding	No term based on experiment	No term based on experiment
Rat	Ogdhl PubMed	11/11	Show all (6)	No term based on experiment	No term based on experiment	No term based on experiment
Mouse	Ogdhl PubMed IMPC	12/13	No data available Open on MGI	No term based on experiment	No term based on experiment	No term based on experiment
Zebrafish	ogdhl PubMed	12/12	No structure expressed in wild-type Open on ZFIN	No term based on experiment	No term based on experiment	No term based on experiment
Drosophila	Nc73EF PubMed	9/12	<ul style="list-style-type: none">HeadEyeBrainThoracic-Abdominal GanglionCropMidgutHindgutMalpighian TubulesFat BodySalivary GlandHeartCarcassOvaryTestisVirginFemale SpermathecaInseminatedFemale SpermathecaMale Accessory Gland Show all (25)	No term based on experiment	<ul style="list-style-type: none">microtubule associated complex	No term based on experiment
C Elegans	ogdh-1 PubMed	8/12	Open on WormBase	<ul style="list-style-type: none">oxoglutarate dehydrogenase (succinyl-transferring) activityoxidoreductase activity, acting on the aldehyde or oxo group of donors, disulfide as acceptorthiamine pyrophosphate binding	<ul style="list-style-type: none">mitochondrionmitochondrial matrixcytosolmitochondrial membraneoxoglutarate dehydrogenase complex	<ul style="list-style-type: none">generation of precursor metabolites and energyglycolytic processtricarboxylic acid cycle
Budding Yeast	KGD1 PubMed	9/11	Open on SGD	No term based on experiment	<ul style="list-style-type: none">mitochondrionmitochondrial oxoglutarate dehydrogenase complexmitochondrial nucleoid	No term based on experiment
Fission Yeast	SPBC3H7.03c PubMed	6/8	Open on PomBase	No term based on experiment	<ul style="list-style-type: none">mitochondrion	No term based on experiment

Human Gene Protein Domains®

OGDHL

DIOPT V6

Index	Domain name	Domain start	Domain stop	Domain description	Protein ID	External ID
hs1	Transket_pyr	636	852	Transketolase, pyrimidine binding domain; pfam02779	NP_060715.2	CDD:202390
hs1	TPP_E1_OGDC_like	251	514	Thiamine pyrophosphate (TPP) family, E1 of OGDC-like subfamily, TPP-binding module; composed of proteins similar to the E1 component of the 2-oxoglutarate dehydrogenase multienzyme complex (OGDC). OGDC catalyzes the oxidative decarboxylation of...; cd02016	NP_060715.2	CDD:48179
hs1	sucA	46	1001	2-oxoglutarate dehydrogenase E1 component; Reviewed; PRK09404	NP_060715.2	CDD:181824

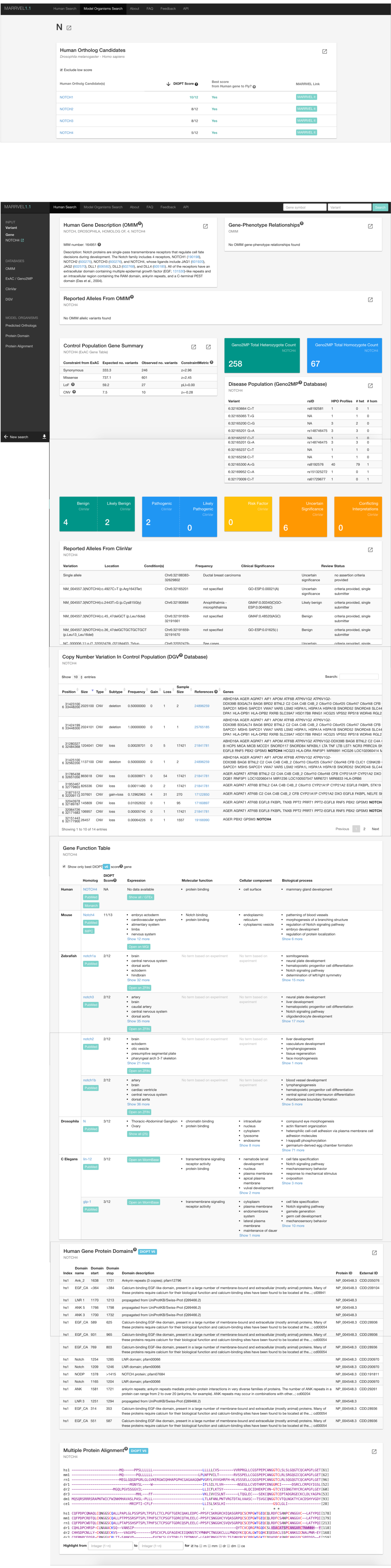
Multiple Protein Alignment®

OGDHL

DIOPT V6

hs1 -----MSQLRLLP---SRLGVQAARLLAAHDVPVFGWRSRSSGPPATFPSS-----KGGGGSSYMEEMYFAWLENPQSVHKSWDSFFREASEEAFSGSAQP---RPPSV---[93]
mm1 -----MSQLRLLP---FRLGPRATKLLATRAIPVFGRCRRSSGPPTTIPRS-----RSGVSSSYVEEMYFAWLENPQSVHKSWDSFFORASKEASVGAQP---QLPAV---[93]
rn1 -----MSQLRLLP---FRLGPQARKLLATRDIAAFGRRRSSGPPTTIPRS-----RGGVSPSYVEEMYFAWLENPQSVHKSWDNFFORATKEASVGAQP---QPPAV---[93]
dm1 -----MNQLRPLA---GALRSSSQWM-----RGHVGTKRVFDLR---RNCSSGVT---EPLAACSSSYVEEMYAWLEDHKNVHESWDAYFRNAE---ASSS---[82]
dr1 MHRHAHTAFSLALSP---MAHKNFATWLLKS-----SSSQMAKVTAATAAVRTYNSAAA---EPFANGSTASYVEEMYNAWLDRPTSVHTSWDAYFRSNS---YVSPNLL---AP---[97]
ce1 MHRASLICRLASPRINAINAS-----SGKSHISASTLVQH---RNOSVAAAVKHEPFLNGSSSIYIEQMYEAWLQDPSSVHTSWDAYFRNVE---AGAGPGQAFQAPPAT---[101]
sc1 -----MLRFVSS---QTCRYSSRGLLKTSL-----LNASTVKIVGRGLATTGT---DNFLTSTNATYIDEMYQAWKDPSSVHVSWDAYFKNMS---NPKIPATKAFQAPPS---[94]
sp1 -----MLRFIPS-SA---KARALRRSAVTAY---RLNRLTCLSSLQ---QNRTFATQPT---DDFLTGGGAADYVDEMYDAWKDPNSVHSAWQAYFKNVQ---ERGVSPSKAFQAPPL---[97]
* . * : * * * . : . * * * : * :
hs1 ---VHESRSVSSR-----T-----KTSKLVEDHLAVQSLIRAYQIRGHVVAQLDPLGILDADLD---SFVPSDLITTTIDKL-----AFYDLQEADLDKEFQ[174]
mm1 ---LQESRTSVSSC-----T-----KTSKLVEDHLAVQSLIRAYQIRGHVVAQLDPLGILDADLD---SFVPSDLITTTIDKLGSWDPSSLFSYAALASFPAYDLEADLDKEFR[193]
rn1 ---IQESRASVSSC-----T-----KTSKLVEDHLAVQSLIRAYQIRGHVVAQLDPLGILDADLD---SFVPSDLITTTIDKL-----AFYDLQEADLDKEFR[174]
dm1 ---VESGEKPLMLLQGRMSQTPA-----MSEKLVEDHLAVHTLIRAYQVRGHVVARLDPLGILTDADLD---SFVPSDLITTSIDKL-----ASYGLEESDLDKSFQ[173]
dr1 ---VQANTLPLTAFNFGGAVAGA-A-----PDSKTIDDLAVQATIRSYQIRGHVIAHLDPLEINTPEL-----PQN---SSTKSI-----Y-----ANFSFGEQDMDROFK[182]
ce1 ---AVAGLGVGDAKARUTTCACADATDRTNACURCTCNLVTALITDEVATGCTGATNADICMGARLRTTDEI-----FI-----CEVLEGEQDMDROFK[182]

Highlight from Integer (1~n) to Integer (1~n) for ☒ hs ☐ rn ☐ mm ☐ dr ☐ dm ☐ ce ☐ sc ☐ sp





Type of database	Name of Database	URL/Link to Database	Rationale for Inclusion into MARRVEL	Reference (PMID)
Human Genetics	ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/	ClinVar is a public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. Variants with interpretations reported by researchers and clinicians are valuable for analyzing how likely a variant is pathogenic.	PMID: 29165669
Human Genetics	DECIPHER	https://decipher.sanger.ac.uk/	The DECIPHER data displayed on MARRVEL includes common variants from the control population. The data displayed includes structural variants that cover the genomic location of the input variant. DECIPHER also contains variant and phenotypic information for affected individuals but can only be accessed directly through their website.	PMID: 19344873
Human Genetics	DGV	http://dgv.tcag.ca/dgv/app/home	To our knowledge, DGV is the largest public-access collection of structural variants from more than 54,000 individuals. The database includes samples of reportedly healthy individuals, at the time of ascertainment, from up to 72 different studies. Possible limitations to this data include variation in source and method of the data acquired the lack of information regarding incomplete penetrance of pathogenic CNVs, and whether individuals will develop associated diseases subsequent to data collection.	PMID: 24174537
Orthology Prediction	DIOPT	https://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl	DIOPT provided multiple protein sequence alignment of the best predicted orthologs in six model organisms against the protein sequence of the human gene of interest. The alignment will provide information on the conservation of specific amino acids as well as functional protein domains.	PMID: 21880147
Human Gene/Transcript Nomenclature	Ensembl	https://useast.ensembl.org/	Ensembl gene IDs are used to link the different databases.	PMID: 29155950
Human Genetics	ExAC	http://exac.broadinstitute.org/	ExAC contains more than 60,000 exomes and is, other than gnomAD (http://gnomad.broadinstitute.org/), the largest public collection of exomes that have been selected against individuals with severe early-onset Mendelian phenotypes. For MARRVEL's purposes, ExAC and gnomAD serves as the best control population dataset to calculate minor allele frequency. We provide two sets of outputs from ExAC. The first output is the gene-centric overview of the expected versus observed number of missense and loss of function (LOF) alleles. A metric called pLI (probability of LOF intolerance) ranges between 0.00 and 1.00 reflects the selective pressure on certain variants before reproductive age. pLI score of 1.00 means that this gene is very intolerant of any LOF variants and haploinsufficiency of this gene may cause disease in human. The second output is data from ExAC that pertains to the specific variant. If identical variant is seen in ExAC, MARRVEL will display the minor allele frequency.	PMID: 27535533
Primary Model Organism Databases	FlyBase (<i>Drosophila</i>)	http://flybase.org	MARRVEL collects and displays data from multiple model organism databases. We provide a summary of the molecular, cellular and biological function of the gene using GO terms. The most likely ortholog is derived by DIOPT.	PMID:26467478
Model Organism Database Integration Tools	Gene2Function	http://www.gene2function.org/search/	MARRVEL collaborates with DIOPT and Gene2Function to provide the "Model Organism Search" feature. Hyperlink is provided for users to access their website that integrates a number of MO databases and displays them in a different style from how MARREL does.	PMID: 28663344
Human Genetics	Geno2MP	http://geno2mp.gs.washington.edu/Geno2MP/	Geno2MP is a collection of samples from the University of Washington Center for Mendelian Genetics. It contains ~9,650 exomes of affected individuals and unaffected relatives. This database links the phenotypic as well as mode of inheritance information to specific alleles. For phenotype, by comparing the affected organ system of the patient of interest to the affected individuals in Geno2MP, one may find potential matches. A match in allele, mode of inheritance, and phenotype provides an increased probability that the variant likely pathogenic. However, due to small sample size a negative association does not necessarily decrease a variant's pathogenic priority. A mechanism to contact the primary physician of a patient of interest is provided in the original source.	N/A
Human Genetics	gnomAD	http://gnomad.broadinstitute.org/	gnomAD contains a total of 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies. Significant portion of ExAC data is integrated into gnomAD. In MARRVEL we currently display the population frequencies that pertains to specific variant.	PMID: 27535533
Gene Ontology	GO Central	http://www.geneontology.org/	MARRVEL displays only gene ontology (GO) terms (molecular function, cellular component, and Biological Process) derived from experimental evidence for each gene. They are filtered by "experimental evidence" and "GO terms based on computational analysis evidence."	PMID: 10802651, 25428369
Human Gene/Protein Expression	GTEx	https://gtexportal.org/home/	MARRVEL displays both mRNA and protein expression pattern in human tissues of each gene. The expression pattern can add insight into the phenotypes observed in patients and/or model organisms.	PMID: 29019975, 23715323
Human Gene Nomenclature	HGNC	https://www.genenames.org/	HGNC official gene symbols are used for MARRVEL searches.	PMID: 27799471
Primary Model Organism Databases	IMPC (mouse)	http://www.mousephenotype.org/	MARRVEL provides a hyperlink to corresponding mouse gene pages on the IMPC website. If there has been a knock-out mouse made by the IMPC, an exhaustive list of assays and their results are made available publicly and can provide insight into the phenotype when a gene is lost. Some information is curated in MGI but there maybe a time lag.	PMID: 27626380
Primary Model Organism Databases	MGI (mouse)	http://www.informatics.jax.org/	MARRVEL collects and displays data from multiple model organism databases. We provide a summary of the molecular, cellular and biological function of the gene using GO terms. The most likely ortholog is derived by DIOPT.	PMID:25348401
Model Organism Database Integration Tools	Monarch Initiative	https://monarchinitiative.org/	MARRVEL provides a link to the Phenogrid of a human gene on Monarch Initiative. This grid provides comparisons between the phenotype of model organisms and known human diseases.	PMID: 27899636
Human Variant Nomenclature	Mutalyzer	https://mutalyzer.nl/	MARRVEL uses Mutalyzer's API to convert different variant nomenclatures to genomic location.	PMID: 18000842
Human Genetics	OMIM	https://omim.org/	The three main pieces of information that we draw from OMIM are: gene function, associated phenotypes, and reported alleles. It is helpful to know if a gene is associated with a known Mendelian phenotype (# entries) whose molecular basis is known. Genes without this knowledge are candidates for novel gene discovery. For genes that are this category, if the patient's phenotype does not match the reported disease and phenotype as well as those of the patients in the literature, then this increases the opportunity to provide a phenotypic expansion for the gene of interest.	PMID: 28654725
Primary Model Organism Databases	PomBase (fission yeast)	https://www.pombase.org/	MARRVEL collects and displays data from multiple model organism databases. We provide a summary of the molecular, cellular and biological function of the gene using GO terms. The most likely ortholog is derived by DIOPT.	PMID:22039153
Literature	PubMed	https://www.ncbi.nlm.nih.gov/pubmed/	MARRVEL provides a hyperlink to "Gene" based PubMed search. Clicking this link will allow one to search biomedical papers that refers to the gene of interest based on previous gene names and symbols.	N/A
Primary Model Organism Databases	RGD (rat)	https://rgd.mcw.edu/	MARRVEL collects and displays data from multiple model organism databases. We provide a summary of the molecular, cellular and biological function of the gene using GO terms. The most likely ortholog is derived by DIOPT.	PMID:25355511
Primary Model Organism Databases	SGD (budding yeast)	https://www.yeastgenome.org/	MARRVEL collects and displays data from multiple model organism databases. We provide a summary of the molecular, cellular and biological function of the gene using GO terms. The most likely ortholog is derived by DIOPT.	PMID: 22110037
Human Gene/Protein Expression	The Human Protein Atlas	https://www.proteinatlas.org/	MARRVEL displays both mRNA and protein expression pattern in human tissues of each gene. The expression pattern can add insight into the phenotypes observed in patients and/or model organisms.	PMID: 21752111
Primary Model Organism Databases	WormBase (<i>C. elegans</i>)	http://wormbase.org	MARRVEL collects and displays data from multiple model organism databases. We provide a summary of the molecular, cellular and biological function of the gene using GO terms. The most likely ortholog is derived by DIOPT.	PMID:26578572
Primary Model Organism Databases	ZFIN (zebrafish)	https://zfin.org/	MARRVEL collects and displays data from multiple model organism databases. We provide a summary of the molecular, cellular and biological function of the gene using GO terms. The most likely ortholog is derived by DIOPT.	PMID:26097180



1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

ARTICLE AND VIDEO LICENSE AGREEMENT

Title of Article:

Navigating MARRVEL, a web-based tool that integrates human genomics and

Author(s):

Julia Wang, UDN, Zhenyao Lin, Hugo Beller, Shiyu Li, model organism genetics

Item 1 (check one box): The Author elects to have the Materials be made available (as described at

<http://www.jove.com/author>) via: ☒ Standard Access ☐ Open Access

Item 2 (check one box):



The Author is NOT a United States government employee.



The Author is a United States government employee and the Materials were prepared in the course of his or her duties as a United States government employee.



The Author is a United States government employee but the Materials were NOT prepared in the course of his or her duties as a United States government employee.

ARTICLE AND VIDEO LICENSE AGREEMENT

1. **Defined Terms.** As used in this Article and Video License Agreement, the following terms shall have the following meanings: “**Agreement**” means this Article and Video License Agreement; “**Article**” means the article specified on the last page of this Agreement, including any associated materials such as texts, figures, tables, artwork, abstracts, or summaries contained therein; “**Author**” means the author who is a signatory to this Agreement; “**Collective Work**” means a work, such as a periodical issue, anthology or encyclopedia, in which the Materials in their entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole; “**CRC License**” means the Creative Commons Attribution-Non Commercial-No Derivs 3.0 Unported Agreement, the terms and conditions of which can be found at: <http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode>; “**Derivative Work**” means a work based upon the Materials or upon the Materials and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Materials may be recast, transformed, or adapted; “**Institution**” means the institution, listed on the last page of this Agreement, by which the Author was employed at the time of the creation of the Materials; “**JoVE**” means MyJoVE Corporation, a Massachusetts corporation and the publisher of *The Journal of Visualized Experiments*; “**Materials**” means the Article and / or the Video; “**Parties**” means the Author and JoVE; “**Video**” means any video(s) made by the Author, alone or in conjunction with any other parties, or by JoVE or its affiliates or agents, individually or in collaboration with the Author or any other parties, incorporating all or any portion of the Article, and in which the Author may or may not appear.

2. **Background.** The Author, who is the author of the Article, in order to ensure the dissemination and protection of the Article, desires to have the JoVE publish the Article and create and transmit videos based on the Article. In furtherance of such goals, the Parties desire to memorialize in this Agreement the respective rights of each Party in and to the Article and the Video.

3. **Grant of Rights in Article.** In consideration of JoVE agreeing to publish the Article, the Author hereby grants to JoVE, subject to **Sections 4 and 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Article in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Article into other languages, create adaptations, summaries or extracts of the Article or other Derivative Works (including, without limitation, the Video) or Collective Works based on all or any portion of the Article and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. If the “Open Access” box has been checked in **Item 1** above, JoVE and the Author hereby grant to the public all such rights in the Article as provided in, but subject to all limitations and requirements set forth in, the CRC License.

ARTICLE AND VIDEO LICENSE AGREEMENT

4. **Retention of Rights in Article.** Notwithstanding the exclusive license granted to JoVE in **Section 3** above, the Author shall, with respect to the Article, retain the non-exclusive right to use all or part of the Article for the non-commercial purpose of giving lectures, presentations or teaching classes, and to post a copy of the Article on the Institution's website or the Author's personal website, in each case provided that a link to the Article on the JoVE website is provided and notice of JoVE's copyright in the Article is included. All non-copyright intellectual property rights in and to the Article, such as patent rights, shall remain with the Author.

5. **Grant of Rights in Video – Standard Access.** This **Section 5** applies if the "Standard Access" box has been checked in **Item 1** above or if no box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby acknowledges and agrees that, Subject to **Section 7** below, JoVE is and shall be the sole and exclusive owner of all rights of any nature, including, without limitation, all copyrights, in and to the Video. To the extent that, by law, the Author is deemed, now or at any time in the future, to have any rights of any nature in or to the Video, the Author hereby disclaims all such rights and transfers all such rights to JoVE.

6. **Grant of Rights in Video – Open Access.** This **Section 6** applies only if the "Open Access" box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby grants to JoVE, subject to **Section 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Video in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Video into other languages, create adaptations, summaries or extracts of the Video or other Derivative Works or Collective Works based on all or any portion of the Video and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. For any Video to which this Section 6 is applicable, JoVE and the Author hereby grant to the public all such rights in the Video as provided in, but subject to all limitations and requirements set forth in, the CRC License.

7. **Government Employees.** If the Author is a United States government employee and the Article was prepared in the course of his or her duties as a United States government employee, as indicated in **Item 2** above, and any of the licenses or grants granted by the Author hereunder exceed the scope of the 17 U.S.C. 403, then the rights granted hereunder shall be limited to the maximum rights permitted under such

statute. In such case, all provisions contained herein that are not in conflict with such statute shall remain in full force and effect, and all provisions contained herein that do so conflict shall be deemed to be amended so as to provide to JoVE the maximum rights permissible within such statute.

8. **Likeness, Privacy, Personality.** The Author hereby grants JoVE the right to use the Author's name, voice, likeness, picture, photograph, image, biography and performance in any way, commercial or otherwise, in connection with the Materials and the sale, promotion and distribution thereof. The Author hereby waives any and all rights he or she may have, relating to his or her appearance in the Video or otherwise relating to the Materials, under all applicable privacy, likeness, personality or similar laws.

9. **Author Warranties.** The Author represents and warrants that the Article is original, that it has not been published, that the copyright interest is owned by the Author (or, if more than one author is listed at the beginning of this Agreement, by such authors collectively) and has not been assigned, licensed, or otherwise transferred to any other party. The Author represents and warrants that the author(s) listed at the top of this Agreement are the only authors of the Materials. If more than one author is listed at the top of this Agreement and if any such author has not entered into a separate Article and Video License Agreement with JoVE relating to the Materials, the Author represents and warrants that the Author has been authorized by each of the other such authors to execute this Agreement on his or her behalf and to bind him or her with respect to the terms of this Agreement as if each of them had been a party hereto as an Author. The Author warrants that the use, reproduction, distribution, public or private performance or display, and/or modification of all or any portion of the Materials does not and will not violate, infringe and/or misappropriate the patent, trademark, intellectual property or other rights of any third party. The Author represents and warrants that it has and will continue to comply with all government, institutional and other regulations, including, without limitation all institutional, laboratory, hospital, ethical, human and animal treatment, privacy, and all other rules, regulations, laws, procedures or guidelines, applicable to the Materials, and that all research involving human and animal subjects has been approved by the Author's relevant institutional review board.

10. **JoVE Discretion.** If the Author requests the assistance of JoVE in producing the Video in the Author's facility, the Author shall ensure that the presence of JoVE employees, agents or independent contractors is in accordance with the relevant regulations of the Author's institution. If more than one author is listed at the beginning of this Agreement, JoVE may, in its sole discretion, elect not take any action with respect to the Article until such time as it has received complete, executed Article and Video License Agreements from each such author. JoVE reserves the right, in its absolute and sole discretion and without giving any reason therefore, to accept or decline any work submitted to JoVE. JoVE and its employees, agents and independent contractors shall have

ARTICLE AND VIDEO LICENSE AGREEMENT

full, unfettered access to the facilities of the Author or of the Author's institution as necessary to make the Video, whether actually published or not. JoVE has sole discretion as to the method of making and publishing the Materials, including, without limitation, to all decisions regarding editing, lighting, filming, timing of publication, if any, length, quality, content and the like.

11. **Indemnification.** The Author agrees to indemnify JoVE and/or its successors and assigns from and against any and all claims, costs, and expenses, including attorney's fees, arising out of any breach of any warranty or other representations contained herein. The Author further agrees to indemnify and hold harmless JoVE from and against any and all claims, costs, and expenses, including attorney's fees, resulting from the breach by the Author of any representation or warranty contained herein or from allegations or instances of violation of intellectual property rights, damage to the Author's or the Author's institution's facilities, fraud, libel, defamation, research, equipment, experiments, property damage, personal injury, violations of institutional, laboratory, hospital, ethical, human and animal treatment, privacy or other rules, regulations, laws, procedures or guidelines, liabilities and other losses or damages related in any way to the submission of work to JoVE, making of videos by JoVE, or publication in JoVE or elsewhere by JoVE. The Author shall be responsible for, and shall hold JoVE harmless from, damages caused by lack of sterilization, lack of cleanliness or by contamination due to the making of a video by JoVE its employees, agents or independent contractors. All sterilization, cleanliness or decontamination procedures shall be solely the responsibility of the Author and shall be undertaken at the Author's

expense. All indemnifications provided herein shall include JoVE's attorney's fees and costs related to said losses or damages. Such indemnification and holding harmless shall include such losses or damages incurred by, or in connection with, acts or omissions of JoVE, its employees, agents or independent contractors.

12. **Fees.** To cover the cost incurred for publication, JoVE must receive payment before production and publication the Materials. Payment is due in 21 days of invoice. Should the Materials not be published due to an editorial or production decision, these funds will be returned to the Author. Withdrawal by the Author of any submitted Materials after final peer review approval will result in a US\$1,200 fee to cover pre-production expenses incurred by JoVE. If payment is not received by the completion of filming, production and publication of the Materials will be suspended until payment is received.

13. **Transfer, Governing Law.** This Agreement may be assigned by JoVE and shall inure to the benefits of any of JoVE's successors and assignees. This Agreement shall be governed and construed by the internal laws of the Commonwealth of Massachusetts without giving effect to any conflict of law provision thereunder. This Agreement may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to be one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

A signed copy of this document must be sent with all new submissions. Only one Agreement required per submission.

CORRESPONDING AUTHOR:

Name:

Shinya Yamamoto

Department:

Molecular & Human Genetics

Institution:

Baylor College of Medicine

Article Title:

Navigating MAPVEL, a web-based tool that integrates human

genomics and
model organism
genetics

Signature:



Date:

12/12/2019

Please submit a signed and dated copy of this license by one of the following three methods:

- 1) Upload a scanned copy of the document as a pdf on the JoVE submission site;
- 2) Fax the document to +1.866.381.2236;
- 3) Mail the document to JoVE / Attn: JoVE Editorial / 1 Alewife Center #200 / Cambridge, MA 02139

For questions, please email submissions@jove.com or call +1.617.945.9051

Editorial comments:

1. The editor has formatted the manuscript to match the journal's style. Please retain the same.

We have retained the format.

2. Please include a figure/table for each of the cases described in the representative result.

Supplemental Figures 3, 4, and 5 are uploaded.

3. Once done please ensure that the highlight is no more than 2.75 pages including headings and spacings.

Completed.

4. Please expand the journal title in the reference section.

Completed.

Members of the Undiagnosed Diseases Network

Maria T. Acosta
David R. Adams
Aaron Aday
Mercedes E. Alejandro
Patrick Allard
Euan A. Ashley
Mahshid S. Azamian
Carlos A. Bacino
Eva Baker
Ashok Balasubramanyam
Dustin Baldrige
Gabriel F. Batzli
Alan H. Beggs
Hugo J. Bellen
Jonathan A. Bernstein
Gerard T. Berry
Anna Bican
David P. Bick
Camille L. Birch
Carsten Bonnenmann
Devon Bonner
Braden E. Boone
Bret L. Bostwick
Lauren C. Briere
Elly Brokamp
Donna M. Brown
Matthew Brush
Elizabeth A. Burke
Lindsay C. Burrage
Manish J. Butte
Hsiao-Tuan Chao
Gary D. Clark
Terra R. Coakley
Joy D. Cogan
F. Sessions Cole
Heather A. Colley
Cynthia M. Cooper
Heidi Cope
William J. Craigen
Precilla D'Souza
Mariska Davids
Jean M. Davidson
Jyoti G. Dayal
Esterban C. Dell'Angelica
Shweta U. Dhar
Laurel A. Donnell-Fink
Naghmeah Dorrani
Daniel C. Dorset
Emilie D. Douine

David D. Draper
Annika M. Dries
Laura Duncan
David J. Eckstein
Lisa T. Emrick
Christine M. Eng
Gregory M. Enns
Cecilia Esteves
Tyra Estwick
Liliana Fernandez
Carlos Ferreira
Elizabeth L. Fieg
Paul G. Fisher
Brent L. Fogel
Noah D. Friedman
William A. Gahl
Rena A. Godfrey
Alica M. Goldman
David B. Goldstein
Jean-Philippe F. Gourdine
Catherine A. Groden
Andrea L. Gropman
Melissa Haendel
Rizwan Hamid
Neil A. Hanchard
Frances High
Ingrid A. Holm
Jason Hom
Alden Huang
Yong Huang
Fariha Jamal
Yong-hui Jiang
Jean M. Johnston
Angela L. Jones
Lefkothea Karaviti
Emily G. Kelley
David M. Koeller
Isaac S. Kohane
Jennefer N. Kohler
Donna M. Krasnewich
Susan Korrick
Mary Koziura
Joel B. Krier
Jennifer E. Kyle
Seema R. Lalani
C. Christopher Lau
Jozef Lazar
Kimberly LeBlanc
Brendan H. Lee
Hane Lee
Shawn E. Levy

Richard A. Lewis
Sharyn A. Lincoln
Pengfei Liu
Sandra K. Loo
Joseph Loscalzo
Richard L. Maas
Ellen F. Macnamara
Calum A. MacRae
Valerie V. Maduro
Marta M. Majcherska
May Christine V. Malicdan
Laura A. Mamounas
Teri A. Manolio
Thomas C. Markello
Ronit Marom
Martin G. Martin
Julian A. Martínez-Agosto
Shruti Marwaha
Thomas May
Allyn McConkie-Rosell
Colleen E. McCormack
Alexa T. McCray
Jason D. Merker
Thomas O. Metz
Matthew Might
Paolo M. Moretti
Marie Morimoto
John J. Mulvihill
David R. Murdock
Avi Nath
Stan F. Nelson
J. Scott Newberry
John H. Newman
Sarah K. Nicholas
Donna Novacic
James P. Orengo
Stephen Pak
J. Carl Pallais
Christina GS. Palmer
Jeanette C. Papp
Neil H. Parker
Loren DM. Pena
John A. Phillips III
Jennifer E. Posey
John H. Postlethwait
Lorraine Potocki
Barbara N. Pusey
Genecee Renteria
Chloe M. Reuter
Lynette Rives
Amy K. Robertson

Lance H. Rodan
Jill A. Rosenfeld
Robb K. Rowley
Jacinda B. Sampson
Susan L. Samson
Timothy Schedl
Kelly Schoch
Daryl A. Scott
Lisa Shakachite
Prashant Sharma
Vandana Shashi
Kathleen Shields
Jimann Shin
Rebecca Signer
Catherine H. Sillari
Edwin K. Silverman
Janet S. Sinsheimer
Kevin S. Smith
Lilianna Solnica-Krezel
Rebecca C. Spillmann
Joan M. Stoler
Nicholas Stong
Jennifer A. Sullivan
David A. Sweetser
Cecelia P. Tamburro
Queenie K.-G. Tan
Cynthia J. Tift
Camilo Toro
Alyssa A. Tran
Tiina K. Urv
Tiphonie P. Vogel
Daryl M. Waggott
Colleen E. Wahl
Nicole M. Walley
Chris A. Walsh
Melissa Walker
Jennifer Wambach
Jijun Wan
Lee-kai Wang
Michael F. Wangler
Patricia A. Ward
Katrina M. Waters
Bobbie-Jo M. Webb-Robertson
Daniel Wegner
Monte Westerfield
Matthew T. Wheeler
Anastasia L. Wise
Lynne A. Wolfe
Jeremy D. Woods
Elizabeth A. Worthey
Shinya Yamamoto

John Yang
Amanda J. Yoon
Guoyun Yu
Diane B. Zastrow
Chunli Zhao

MARRVEL1.1Human SearchModel Organisms SearchAboutFAQFeedbackAPI

Step 1.3

MARRVEL

Step 1.1

Human Gene Symbol:

Please use official HGNC Gene Symbol. E.g. FBXL4

Human Variant (hg19):

E.g. 6:99365567 T>C or NC_000006.11:g.99365567T>C

Example: 6:99365567 T>C / FBXL4 or 6:99365567 T>C or FBXL4 or NM_012160.3:c.541A>G

Search

MARRVEL1.1Human SearchModel Organisms SearchAboutFAQFeedbackAPI

Step 1.3.2

MARRVEL

Step 1.3.1

Model organism

Fly (Drosophila melanogaster)

Official gene symbol

N

Example: N

Search

MARRVEL1.1Human SearchModel Organisms SearchAboutFAQFeedbackAPI

N

Human Ortholog Candidates

Drosophila melanogaster - Homo sapiens

☒ Exclude low score

Human Ortholog Candidate(s)	↓ DIOPT Score	Best score from Human gene to Fly?	MARRVEL Link
NOTCH1	10/12	Yes	MARRVEL.it
NOTCH2	8/12	Yes	MARRVEL.it
NOTCH3	8/12	Yes	MARRVEL.it
NOTCH4	5/12	Yes	MARRVEL.it

Step 1.3.2

MARRVEL1.1

Human SearchModel Organisms SearchAboutFAQFeedbackAPI

Gene symbolVariantSearch

INPUT

Variant

Gene

NOTCH1

DATABASES

OMIM

ExAC / Geno2MP

ClinVar

DGV

MODEL ORGANISMS

Predicted Orthologs

Protein Domain

Protein Alignment

New search

Gene Function Table

NOTCH1

☒ Show only best DIOPT score gene

	Homolog	DIOPT Score	Expression	Molecular function	Cellular component	Biological process
Human	<div>NOTCH1</div> <div>PubMed</div> <div>Monarch</div>	NA	<div><ul style="list-style-type: none">adrenal glandappendixcolongallbladderlungrectumstomachtestis</div> <div>Show all / GTEX</div>	<div><ul style="list-style-type: none">protein binding</div>	<div><ul style="list-style-type: none">MAML1-RBP-Jkappa- ICN1 complexreceptor complex</div>	<div><ul style="list-style-type: none">aortic valve morphogenesispulmonary valve morphogenesismitral valve formationNotch signaling pathwayheart development</div> <div>Show 16 more</div>
Rat	<div>Notch1</div> <div>PubMed</div>	11/11	<div>Show all (10)</div>	<div>No term based on experiment</div>	<div><ul style="list-style-type: none">acrosomal vesiclenucleusGolgi apparatusplasma membrane</div>	<div><ul style="list-style-type: none">positive regulation of neuroblast proliferationregulation of cardioblast proliferationNotch signaling pathwaypositive regulation of transcription of Notch receptor targetspermatogenesis</div> <div>Show 13 more</div>
Mouse	<div>Notch1</div> <div>PubMed</div> <div>IMPC</div>	13/13	<div><ul style="list-style-type: none">extraembryonic componentembryo ectodermembryo mesodermcardiovascular systembranchial arches</div> <div>Show 18 more</div>	<div><ul style="list-style-type: none">core promoter bindingtranscriptional activator activity, RNA polymerase II transcription factor bindingchromatin bindingtranscription factor activity, sequence-specific DNA bindingenzyme inhibitor activity</div>	<div><ul style="list-style-type: none">nucleuscytoplasmendoplasmic reticulumadherens junctioncell surface</div>	<div><ul style="list-style-type: none">negative regulation of transcription from RNA polymerase II promoterin utero embryonic developmentcell fate specificationepithelial to mesenchymal</div>

Title:

In vivo* functional study of disease-associated rare human variants using *Drosophila

Authors and Affiliations:

J. Michael Harnish^{1*}, Samantha L. Deal^{2*}, Undiagnosed Diseases Network[%], Hsiao-Tuan Chao^{3,4}, Michael F. Wangler^{1,2,4}, Shinya Yamamoto^{1,2,4,5#}

¹ Department of Molecular and Human Genetics, Baylor College of Medicine (BCM), Houston, TX 77030, USA

² Program in Developmental Biology, BCM, Houston, TX 77030, USA

³ Department of Pediatrics, Section of Neurology and Developmental Neuroscience, BCM, Houston, TX 77030, USA

⁴ Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX 77030, USA

⁵ Department of Neuroscience, BCM, Houston, TX 77030, USA

* These authors contributed equally

Corresponding Author

E-mail: yamamoto@bcm.edu

phone: +1-832-824-8119

e-mail addresses of co-authors: JMH Jacob.Harnish@bcm.edu; SLD Samantha.Deal@bcm.edu; HTC hc140077@bcm.edu; MFW mw147467@bcm.edu

% List of members of the 'Undiagnosed Diseases Network' is provided in Supplemental Table 1.

Keywords:

Human genetics and genomics, Mendelian diseases, rare and undiagnosed diseases, Undiagnosed Diseases Network (UDN), *Drosophila melanogaster*, variant of unknown significance (VUS), gene of uncertain significance (GUS), functional genomics, transgenic flies, UAS/GAL4 system, T2A-GAL4, electroretinogram (ERG)

Summary:

The goal of this protocol is to outline a process to design and perform *in vivo* experiments in *Drosophila melanogaster* to assess the functional consequences of rare variants that are associated with human diseases.

Abstract:

Advances in sequencing technology have made whole-genome and whole-exome datasets more easily accessible for both clinical diagnosis and cutting-edge human genetic research. Although a number of *in silico* algorithms have been developed to predict the pathogenicity of variants identified in these datasets, functional studies are critical to determine if specific genomic variants may affect protein function, especially for missense variants. In the Undiagnosed Disease Network (UDN) and other rare disease research consortiums, model organisms (MO) including *Drosophila*, *C. elegans*, zebrafish, and mice are actively being used to assess the function of putative human disease-causing variants. Here, we describe a protocol for functional assessment of rare variants that is being used in the UDN Model Organisms Screening Center *Drosophila* Core. The workflow begins with gathering human and MO information from multiple public databases using the MARRVEL tool to assess whether the variant is likely to contribute to patients' conditions and design effective experiments based on available knowledge and resources. Next, we generate genetic tools (T2A-GAL4 lines and UAS-human cDNAs) to assess the function of variants of interest in *Drosophila*. Upon development of these reagents, we perform a two-pronged functional assay based on rescue and over-expression experiments. In the rescue branch, we attempt to "humanize" the endogenous fly gene by replacing the orthologous *Drosophila* gene with reference or variant human transgenes. In the over-expression branch, we exogenously drive the reference and variant human proteins in a variety of tissues. In both cases, any scorable phenotype (e.g. lethality, eye morphology, electroretinogram) can be used as outputs irrespective of the disease of interest. Differences observed between reference and variant alleles suggest functional differences caused by the missense variant, thus suggesting pathogenicity. This protocol allows rapid *in vivo* assessments of putative human disease-causing variants for genes of both known and unknown function.

Introduction:

Patients with rare diseases often undergo an arduous journey referred to as the ‘diagnostic odyssey’ to obtain an accurate diagnosis¹. Most rare diseases are thought to have a strong genetic origin, making genetic and genomics analyses critical elements of the clinical workup. In addition to candidate gene panel sequencing and copy number variation analysis based on chromosomal microarrays, whole-exome (WES) and whole-genome sequencing (WGS) technologies have become increasingly valuable tools over the past decade^{2,3}. Currently, the diagnostic rate of identifying a known pathogenic variant in WES and WGS is ~25% (higher for pediatrics cases)^{4,5}. For most cases that remain undiagnosed after receiving clinical WES/WGS, the issue is that there are too many candidate genes and variants. Next generation sequencing often identifies novel or ultra-rare variants in many genes in an individual’s exome or genome, and interpreting whether these variants may contribute to disease phenotypes is challenging. For example, although most nonsense or frameshift mutations in genes are thought to be loss-of-function alleles due to nonsense mediated decay of the encoded transcript, truncating mutations found in the last exons escape this process and may function as benign or gain-of-function alleles⁶. Moreover, predicting the effect of a missense allele is a daunting task since it can result in a number of different genetic scenarios as first described by Herman Muller in the 1930s; amorph, hypomorph, hypermorph, antimorph, neomorph, or isomorph⁷. Numerous *in silico* programs and methodologies have been developed to predict the pathogenicity of missense variants based on evolutionary conservation, type of amino acid changes, position within a functional domain, allele frequency in the general population, and other parameters⁸. However, these programs are not a comprehensive solution to solving the complicated problem of variant interpretation. Interestingly, a recent study demonstrated that five broadly used variant pathogenicity prediction algorithms [Polyphen (genetics.bwh.harvard.edu/pph2)⁹, SIFT (sift.bii.a-star.edu.sg)¹⁰, CADD (cadd.gs.washington.edu)¹¹, PROVEAN (provean.jcvi.org/index.php)¹², and Mutation Taster (www.mutationtaster.org)] agree on pathogenicity ~80% of the time⁸. However, even when all algorithms agree, they return an incorrect prediction of pathogenicity up to 11% of the time. This not only leads to flawed clinical interpretation, but may also dissuade researchers from following up on new variants by falsely listing them as benign. One way to complement the current limitation of *in silico* modeling is to provide experimental data that demonstrates the effect of variant function *in vitro*, *ex vivo* (e.g. cultured cells, organoids), or *in vivo*.

In vivo functional studies of rare disease associated variants in MO have unique strengths¹³, and have been adopted by many rare disease research initiatives around the world including the Undiagnosed Diseases Network (UDN) in the United States (undiagnosed.hms.harvard.edu) and the Rare Diseases Models & Mechanisms (RDMM) Networks in Canada (www.rare-diseases-catalyst-network.ca), Japan (irudbeyond.nig.ac.jp), Europe (solve-rd.eu) and Australia (www.functionalgenomics.org.au)¹⁴. In addition to these coordinated efforts to integrate MO

researchers into the workflow of rare disease diagnosis and mechanistic studies at a national scale, a number of individual collaborative studies between clinical and MO researches have led to discovery and characterization of new human disease-causing genes and variants⁸²⁻⁸⁴. In the UDN, a centralized Model Organisms Screening Center (MOSC) receives submissions of candidate genes and variants with a description of the patients' condition and assesses whether the variant is likely to be pathogenic using informatics tools and *in vivo* experiments. In Phase I (2015-2018) of the UDN, the MOSC comprised of a *Drosophila* Core [Baylor College of Medicine (BCM)] and a Zebrafish Core (University of Oregon) that worked collaboratively to assess cases. Using informatics analysis and a number of different experimental strategies in *Drosophila* and zebrafish, the MOSC has so far contributed to the diagnosis of 132 patients, the identification of 31 new syndromes⁵⁵, the discovery of several new human disease genes (e.g. *EBF3*¹⁵, *ATP5F1D*¹⁶, *TBX2*¹⁷, *IRF2BPL*¹⁸, *COG4*¹⁹, *WDR37*²⁰) and phenotypic expansion of known disease genes (e.g. *CACNA1A*²¹, *ACOX1*²²). In addition to projects within the UDN, MOSC *Drosophila* Core researchers have contributed to new disease gene discoveries in collaboration with the Center for Mendelian Genomics and other initiatives (e.g. *ANKLE2*²³, *TM2D3*²⁴, *NRD1*²⁵, *OGDHL*²⁵, *ATAD3A*²⁶, *ARIH1*²⁷, *MARK3*²⁸, *DNMBP*²⁹) using the same set of informatics and genetic strategies that were developed for the UDN. Given the significance of MO studies on rare disease diagnosis, the MOSC was expanded to include a *C. elegans* Core and an additional Zebrafish core (both at Washington University at St Louis) for the Phase II (2018-2022) of the UDN.

In this manuscript, we describe an *in vivo* functional study **Protocol** that is actively being used in the UDN MOSC *Drosophila* Core to determine if missense variants have a functional consequence on the protein of interest using transgenic flies that express human proteins. The goal of this **Protocol** is to help MO researchers to work collaboratively with the clinical research groups to provide experimental evidence that the variant of interest has functional consequences and facilitate the clinical diagnosis. This **Protocol** will be most useful in a scenario in which a *Drosophila* researcher is approached by a clinical investigator who has a rare disease patient with a specific candidate variant in a gene of interest. This **Protocol** can be broken down into three elements; (1) Gathering information to assess the likelihood of the variant of interest being responsible for the patient phenotype and the feasibility of a functional study in *Drosophila*, (2) Gathering existing genetic tools and establishing new ones, and (3) Performing functional studies *in vivo*. The third element can further be subdivided into two sub-elements based on how one can assess the function of a variant of interest (rescue experiment or over-expression based strategies). It is important to note that this **Protocol** can be adapted and optimized to many scenarios outside of rare monogenic disease research (e.g. common diseases, gene-environment interaction, pharmacological and genetic screens to identify therapeutic targets). The ability to determine the functionality and pathogenicity of variants will not only benefit the patient of interest via providing an accurate molecular diagnosis but will also have broader impacts on translational and basic scientific research.

Protocol:

1. Gather human and MO information to assess the likelihood of a variant of interest being responsible for disease phenotypes and the feasibility of functional studies in *Drosophila*

1.1 Perform extensive database and literature searches to determine whether the specific genes and variants of interest are good candidates to explain the phenotype of the patient of interest. Some key questions that should be explored include:

- Q1)** Has this gene been previously implicated in other genetic disorders (phenotypic expansion of known disease gene) or is this an entirely new disease candidate gene [gene of uncertain significance (GUS)]?
- Q2)** At what frequency have the alleles of interest been seen in disease or control population databases?
- Q3)** Are there copy number variations that include this gene in disease or control population databases?
- Q4)** What are the orthologous genes in different MO species (e.g. mouse, zebrafish, *Drosophila*, *C. elegans*, yeast) and what are known about their functions and expression patterns?
- Q5)** Is the variant in a functional domain of the protein and is the amino acid of interest evolutionarily conserved?

Note: Answers to these five questions (**A1-5**) can be obtained by accessing a number of human and MO databases individually. Alternatively, one can quickly obtain a summary of these results using the MARRVEL (Model organism Aggregated Resources for Rare Variant ExpLoration, <http://marrvel.org/>) tool³⁰, which is described in-depth in an accompanying JoVE article³¹. See “Representative Results” section for specific examples. Additional internet-based resources such as the Monarch Initiative website (<https://monarchinitiative.org/>)³² and Gene2Function (<http://www.gene2function.org/search/>)³³ may also provide useful information.

1.2 Gather additional information related to the following questions to further assess whether the variant is a good disease candidate or not from a protein function and structure point of view.

- Q6)** Is the variant of interest predicted to be damaging based on *in silico* prediction algorithms?
- Q7)** Does the human gene/protein of interest or its MO orthologs genetically or physically interact with genes/proteins previously linked to genetic diseases? If so, do these diseases have overlapping phenotypes with our patient of interest?
- Q8)** Has the three-dimensional structure of the protein of interest been solved or modeled? If so, where does the variant of interest map relative to its key functional domains?

Note: The following databases and tools can be useful to gather the answers for these questions (**A6-8**). MARRVEL tool will be upgraded to incorporate this information in future updates³¹.

A6) A number of variant pathogenicity algorithms have been developed by many research groups over the past ~15 years. More recent programs, including the two listed below, combine multiple variant pathogenicity prediction algorithms and machine learning approaches to generate a pathogenicity score. For more information on variant prediction algorithms and their performance, we refer the readers to Ghosh *et al*⁸.

- CADD (Combined Annotation-Dependent Depletion): Integrative annotation tool built from more than 60 genomic features, which provides scores for human SNVs as well as short insertions and deletions. (cadd.gs.washington.edu)¹¹
- REVEL (Rare Exome Variant Ensemble Learner): Combines multiple variant pathogenicity algorithms (MutPred, FATHMM, VEST, PolyPhen, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP, SiPhy, phyloP, and phastCons) to provide an integrated score for all possible human missense variants. (sites.google.com/site/revelgenomics)³⁴

A7) Several tools have been developed to analyze genetic and protein-protein interactions based on MO publications as well as large-scale proteomics from multiple species screens.

- STRING (Search Tool for Recurring Instances of Neighboring Genes) (string-db.org)³⁵: A database for known and predicted protein-protein interactions. It integrates genetic interaction and co-expression datasets as well as text-mining tools to identify genes and proteins that may function together in a variety of organisms.
- MIST (Molecular Interaction Search Tool) (fgertools.hms.harvard.edu/MIST)³⁶: A database that integrates genetic and protein-protein interaction data from core genetic MOs (yeast, *C. elegans*, *Drosophila*, zebrafish, frog, rat and mouse) and humans. Prediction of interactions inferred from orthologous genes/proteins (interlogs) are also displayed.

A8) Protein structures that have been solved by X-ray crystallography, nuclear magnetic resonance (NMR) and cryo-electron microscopy can be found in public databases including the PDB (Protein Data bank) (www.wwpdb.org) and EMDatabank (www.emdatabank.org)³⁷. Although there is no single database for predicted/modeled protein structures, a number of algorithms including SWISS-MODEL (swissmodel.expasy.org)³⁸, Modeller (salilab.org/modeller)³⁹ and Phyre2 (www.sbg.bio.ic.ac.uk/phyre2)⁴⁰ are available for users to perform protein modeling.

1.3 Communicate with your clinical collaborator to discuss the information you gathered from the informatics analysis in **1.1** and **1.2**. If you and your clinical collaborator both feel that the variant and gene of interest are good candidates to explain the phenotypes seen in the patient of interest, proceed to **Section 2**. If you have specific questions about the patient's genotype and phenotype, make sure to ask them before moving forward. If you feel the variant of interest is unlikely to explain the patient's phenotype of interest (e.g. identical variant found in high frequency in control population), you must discuss this with your clinical collaborators to determine whether the variant is really a good candidate that is worth investing time and effort.

Note: If your clinical collaborators can identify other patients who have similar genotypes and phenotypes with your patient of interest, this will significantly increase the likelihood that the variant of interest is pathogenic. We encourage the MO researchers to work closely with clinical researchers to search for additional patients with variants in a specific gene of interest. The following two tools, for example, allow one to search through a cohort of patients enrolled in diverse clinical studies to identify patients carrying identical or similar variants.

- Geno₂MP (Genotypes to Mendelian Phenotypes) (geno2mp.gs.washington.edu): A de-identified database of 9,650 individuals (as of Sep 2018) enrolled in the University of Washington Center for Mendelian Genomics study⁴¹ that recruits patients suspected to have genetic disorders as well as their relatives. One can search for variants in a specific gene of interest and determine whether a specific variant is found in a patient with certain phenotypes (classified by the main organ system affected) or in unaffected family members (can be considered as ‘control’ for severe dominant disorders). If there are interesting patients in this database, one can use the “Contact” feature to reach out to the primary physicians who deposited the case via e-mail.
- GeneMatcher (www.genematcher.org): A matchmaking website for clinicians, basic researchers and patients who share interest in the same gene. Upon registration and submission of a gene of interest, GeneMatcher will provide the contact information of other submitters who expressed interest in the same gene. Since most users are clinicians and human geneticists, one can contact them via e-mail to see whether their patient’s genotype and phenotype matches your patient of interest. Since GeneMatcher is part of Matchmaker Exchange (www.matchmakerexchange.org)⁴², one can also search additional matchmaking databases around the world including Australian Genomics Health Alliance Patient Archive (mme.australiangenomics.org.au/#/home), Broad Matchbox (seqr.broadinstitute.org/matchmaker/matchbox), DECIPHER (decipher.sanger.ac.uk), MyGene2 (www.mygene2.org/MyGene2) and PhenomeCentral (phenomecentral.org) by gene submissions through GeneMatcher.

2. Gather existing genetic tools and establish new reagents to study a specific variant of interest

Once you determined that the variant of interest is a good candidate to pursue experimentally, the next step is to gather or generate reagents to perform *in vivo* functional studies. For functional studies described in this protocol, one will need a few key reagents. 1) UAS-Human cDNA transgenic strains that carry the reference or variant sequence, 2) a LOF allele of a fly gene of interest, and 3) a GAL4 line that can be used for rescue experiments.

2.1. Generation of UAS-human cDNA constructs and transgenic flies

2.1.1. Identify and obtain the appropriate human cDNA constructs. Many clones are available from the MGC (Mammalian Gene Collection)⁴³ and can be purchased from selected vendors (genecollections.nci.nih.gov/MGC). Many cDNAs are available in Gateway compatible reagents⁴⁴, which simplifies the subcloning step. If the cDNAs is not included in the MGC or if one wishes to use a specific splice isoform not distributed by non-profit vendors, one can looking into

commercial vendors or use gene synthesis services. We typically select a cDNA corresponding to the longest isoform [(typically referred to as the canonical isoform in Ensembl (useast.ensembl.org) or RefSeq (www.ncbi.nlm.nih.gov/refseq), two major databases that curate cDNA isoforms] if there are multiple options. One should especially be careful when a variant of interest only affects a subset of splicing isoforms.

Note: cDNAs may come in an “open (no stop codon)” or “closed (with endogenous or artificial stop codon)” format. Open clones allow C’ of tagging of proteins when subcloned into a plasmid with C’ tags (e.g. 3xHA tags for pUASg-HA.attB or pGW-HA.attB), whereas proteins will not be tagged if closed clones are subcloned into the same vector. While protein tags may offer biochemical (e.g. western blot) and cell biological (e.g. immunostaining) ways of monitoring the expression of the protein of interest, it may interfere with protein function in some cases.

2.1.2 Sub-clone the reference and variant cDNA into the *Drosophila* transgenic vector. The ϕ C31-mediated transgenesis system is best suited for functional studies using human cDNAs since the reference and variant cDNAs can be integrated into the same location in the genome⁴⁵. For this project, the MOSC *Drosophila* Core routinely use the pGW-HA.attB vector⁴⁶. This is a Gateway compatible vector that contains 5xUAS sites, an *hsp70* promoter, C’ 3xHA tag with a protein linker sequence, *tubulin α 1* 3’UTR and mini-*white*⁺ gene as a transgenesis marker. There are two altered *FRT* (*Flippase Recombination Target*) sites (*FRT5* and *FRT2*) flanking the open reading frame that can further be used to modify the transgene after genomic integration⁴⁷.

If the human cDNA is in Gateway compatible vectors (e.g. pDONR221, pDONR223, pENTR221, pENTR223.1), one can skip to **2.1.4** that explains LR reactions to subclone the cDNAs into pGW-HA.attB. If the human cDNA plasmids are not in a Gateway compatible vector, one can first subclone the fragments into a suitable entry vector (e.g. pDONR221) via the following protocol.

2.1.2a. Subclone the human cDNAs into a gateway compatible plasmid

2.1.2a.1 Perform an overhang PCR to introduce *attB1* and *attB2* arms. The forward primer should have the *attB1* sequence 5’-GGGGACAAGTTTGTACAAAAAGCAGGCTTCACC-3’ followed by the first 22 nucleotides of the target cDNA. The reverse primer should have the *attB2* sequence 5’-GGGGACCACTTTGTACAAGAAAGCTGGGTCCTA-3’ followed by the reverse complement of the last 25 nucleotides of your cDNA of interest. One can excluding the stop codon if you wish to “open” the clone to add a C’ tag, or add a stop codon if you wish to “close” an open clone.

2.1.2a.2 Prepare a 100 μ L Q5 PCR mix consisting of 50 μ L Q5 mastermix (NEB #M0491), 36 μ L milliQ water, 5 μ L of each forward and reverse primers listed in **2.2.1** diluted to 10 μ M, and 4 μ L of target cDNA (150 ng/ μ L).

2.1.2a.3 Perform the PCR using standard Q5 mutagenesis protocol (NEB #M0491)

2.1.2a.4 Isolate the target cDNA with added homology arms via gel electrophoresis and gel extraction. Create 1% agarose gel and perform electrophoresis using standard methods. Cut out

289 the band that corresponds to the size of your cDNA plus the additional length of the homology
290 arms. Extract DNA from the gel through standard methods. Commercial gel extraction kits are
291 available from with several companies (Qiagen #28704).

292 **2.1.2a.5** Perform a BP clonase reaction using standard Gateway cloning protocol (Invitrogen
293 #11789)

294 **2.1.2a.6** Transform the BP reaction mix into chemically competent *E coli* cells. Competent cells
295 can be made in house or purchased from commercial vendors (e.g. NEB #C2987H). Culture the
296 transformant overnight on an LB plate containing appropriate antibiotics for colony selection.
297 The next day select several colonies and grow them up in independent liquid cultures overnight.

298 **2.1.2a.7** Isolate DNA from the overnight cultures and perform diagnostic restriction digests.
299 Sanger sequence the positive clones to ensure that the cDNA is the correct sequence. We
300 recommend generating glycerol stocks from the cultures that were positive for your desired
301 sequence at this point.

302 **2.1.3.** Introduce the variant of interest into your Gateway plasmid with your reference human
303 cDNA. There are several ways to perform this mutagenesis step including methods described in
304 other JoVE protocols^{48,49}. We have been using the Q5 site-directed mutagenesis system in our
305 operation. Detailed protocol for this method can be found in the vendor's website (NEB
306 #E0554S). In addition to validating the presence of the variant in the mutated plasmid, perform
307 Sanger sequencing of the entire open reading frame (ORF) in order to make sure there are no
308 additional variants introduced through this mutagenesis step.

309 **Note:** If the mutagenesis is not successful, we typically re-design the primer and repeat the
310 experiment. If the mutagenesis fails again, we explore other methods such as QuikChange II Site
311 Directed Mutagenesis system (Aligent #200523).

312 **2.1.4** Subclone the reference and variant human cDNAs in the donor plasmid (Gateway plasmids
313 with *attL1* and *attL2* sites) into the transgenic plasmid (e.g. pGW-HA.attB with *attR1* and *attR2*
314 sites) via the LR clonase reaction (Thermo Fisher #11791100).

315 **Note:** There are UAS ϕ C31 vectors that are designed for conventional restriction enzyme based
316 subcloning (e.g. pUAST.attB⁵⁰) if one prefers to subclone human cDNAs via traditional methods.

317 **2.1.4** Inject the UAS-human cDNA constructs into flies expressing the ϕ C31 integrase in their
318 germline (e.g. *vas- ϕ C1*, *nos- ϕ C31*) (bdsc.indiana.edu/stocks/phic31/phic31_int.html).
319 Microinjection can be performed in house, or can be sent to core facilities or commercial entities
320 for transgenesis. Detailed protocol for generating transgenic flies can be found in the following
321 book chapter⁵¹.

322 **Note:** One needs to select a docking site for transgene insertion. Since it is convenient to have
323 the human transgene on a chromosome that does not contain the fly ortholog of the gene of

interest, we typically use a 2nd chromosome docking site [VK37 (BDSC stock #24872, flybase.org/reports/FBst0024872.html)] when the fly ortholog is on the X, 3rd or 4th chromosomes and a 3rd chromosome docking site [VK33 (BDSC stock #24871, flybase.org/reports/FBst0024871.html)] when the fly ortholog is on the 2nd chromosome. A number of additional docking sites have been generated by several laboratories and are publically available from stock centers (bdsc.indiana.edu/stocks/phic31/phic31_attp.html , kyotofly.kit.jp/stocks/documents/phiC31.html)^{50,52,53}.

2.1.5. Establish stable transgenic strains from the injected embryos. We typically inject ~100-200 embryos per construct. A representative crossing scheme for a transgene insertion into a 2nd chromosome docking site (VK37) is depicted in **Figure 1A**. For basic *Drosophila* genetics information, we refer the readers to the following books^{54,55}.

2.2 Obtain or generate a T2A-GAL4 line that facilitates rescue-based functional assays (see **Figure 2** and **Section 3.1**). This line will serve two purposes. First, most T2A-GAL4 lines tested behave as strong LOF alleles by functioning as a gene trap allele. Second, T2A-GAL4 lines function as a GAL4 driver that allows expression of UAS constructs (e.g. UAS-GFP, UAS-human cDNAs) in the pattern of the fly gene of interest^{56,57} (**Figure 2A-C**).

2.2.1 Search public stock collections for available T2A-GAL4 lines. Through the *Drosophila* Gene Disruption Project (GDP)⁵⁸, ~1,000 T2A-GAL4 lines have been generated⁵⁹. These strains are currently available from the Bloomington *Drosophila* Stock Center (BDSC) and are searchable through both the GDP (flypush.imgen.bcm.tmc.edu/pscreen) and BDSC (bdsc.indiana.edu) websites.

2.2.2. If a T2A-GAL4 line for your fly gene of interest is not available, check if a suitable coding intronic MiMIC (*Minos* mediated Integration Cassette) line is available for conversion into a T2A-GAL4 line using recombinase mediated cassette exchange (RMCE)⁶⁰ (**Figure 2A**). RMCE allows intronic MiMIC elements that are in between two coding exons to be converted into a T2A-GAL4 line through injection (an example of a crossing scheme is shown in **Figure 1B**) or series of crosses as described in detail in the following papers^{57,59}.

2.2.3. If a T2A-GAL4 is not available and an appropriate coding intronic MiMIC does not exist, explore the possibility of generating a T2A-GAL4 line via the CRIMIC (CRISPR-mediated Integration Cassette) system as described in Lee *et al.*⁵⁹. This methodology uses CRISPR-mediated DNA cleavage and homology directed repair (HDR) to integrate a MiMIC-like cassette into a coding intron in a gene of interest.

Note: Not all genes can be tagged using a T2A-GAL4 system. For genes that lack introns or only have small (<100bp) coding introns, one can attempt to knock-in a GAL4 transgene into the fly gene of using the CRISPR/Cas9 system using HDR as described in the following papers^{20,61,62}. Alternatively, if the gene of interest have previously characterized mutants, one can attempt to perform rescue experiments using these pre-existing alleles and ubiquitous or tissue-specific GAL4 drivers.

3. Perform Functional Analysis of the human variant of interest *in vivo* in *Drosophila*

Perform a rescue-based analysis (**Section 3.1**) as well as over-expression studies (**Section 3.2**) using the tools gathered or generated in **Section 2** to assess the functional consequence of the variant of interest *in vivo* in *Drosophila*. The two approaches are complementary to one another.

3.1. Functional analysis through rescue based experiments.

Heterologous rescue-based experiments in *Drosophila* using human proteins determine whether the molecular function of the two orthologous genes have been conserved over ~500 million years of evolution, and further assess the function of the variant in the context of the human protein⁶³. Although a systematic analysis studying hundreds of gene pairs has not been reported, several dozen human and mammalian (e.g. mouse) genes have been able to replace the function of *Drosophila* genes¹³.

3.1.1 In the rescue-based approach, we first determine whether there are obvious scorable and reproducible phenotype in LOF mutants in the fly ortholog before assessing the function of variants. Previous literature on the fly gene is the first place to mine for data and can be found using databases such as FlyBase (flybase.org/) and PubMed (www.ncbi.nlm.nih.gov/pubmed/). Additional databases such as MARRVEL (marvel.org), Monarch Initiative (monarchinitiative.org/), and Gene2Function (<http://www.gene2function.org>) are also useful in gathering this information. If the T2A-GAL4 allele is the first mutation to be characterized for a specific gene, one should perform a global survey of scorable phenotypes in homozygous and hemizygous (T2A-GAL4 allele over a molecularly defined chromosomal deficiency; e.g. bdsc.indiana.edu/stocks/df/index.html) animals. These include lethality, sterility, longevity, morphological (e.g. size and morphology of the eye) or behavioral phenotypes (e.g. courtship, flight, climbing and bang sensitivity defects). More sophisticated phenotypes such as neurological defects measured by electrophysiological recordings can also be used as long as they are highly reproducible and specific. Functional studies using electroretinogram (ERG) are described in **3.2.3**.

3.1.2 Once a scorable phenotype is identified in the fly LOF mutant, test whether the reference human cDNA can replace the function of the fly ortholog. If this “humanization” of the fly gene is successful, we now have a platform to compare the efficiency of the variant of interest compared to the reference counterpart. The rescue seen with reference human cDNA does not have to be perfect. Partial rescue of the fly mutant phenotype using a human cDNA still provides a reference point to perform comparative studies using the variant human cDNA strain.

3.1.3 Using the assay system selected in **3.1.2**, compare the rescue observed with the reference human cDNA to the rescue observed with the variant human cDNA to determine if the variant of interest has functional consequences.

3.1.4. If the variant is found to be a LOF allele, one can further compare the expression and intracellular localization of the reference and variant protein of interest *in vivo*, especially if the UAS-transgenes were generated from an ‘open’ clone and have a C-terminal tag that can be used

to image the proteins via western blot, immunofluorescence staining or other methods. If the transgenes were generated from a ‘closed’ clone, one can look into commercial antibodies raised against the human protein of interest and assess whether these reagents can be used to detect these proteins in a fly tissue.

3.2 Functional analysis through over-expression studies

Ubiquitous or tissue-specific over-expression of human cDNAs in otherwise wild-type flies can provide information that is complementary to the rescue-based experiments. While rescue-based assays are primarily designed to detect LOF variants (amorphic, hypomorphic), over-expression based assays may reveal gain-of-function (GOF) variants that may be missed (hypermorphic, antimorphic, neomorphic).

3.2.1 Select a set of GAL4 drivers to over-express the human cDNAs of interest. A number of ubiquitous, tissue and stage specific GAL4 drivers are available from public stock centers (e.g. bdsc.indiana.edu/stocks/gal4/index.html, kyotofly.kit.jp/stocks/documents/GAL4.html), some of which are more frequently used than others. A large collection of GAL4 lines and related resources are being constructed for which we refer the readers to the following papers^{56,57,59}. Upon obtaining these drivers, make sure to validate drivers with a reporter line (e.g. UAS-GFP) to confirm their expression pattern upon receiving the stocks from a stock center.

3.2.2 Express the reference and variant human cDNAs using the same driver under the same condition (e.g. temperature) and ascertain if there is a difference between them. If a phenotype is only seen in the reference but not in the variant line, the variant may be an amorphic or a strong hypomorphic allele. If the phenotype is seen in both genotypes, but the reference causes a stronger defect, the variant may be a mild to weak hypomorphic allele. If the reference does not show a phenotype, or only exhibits a weak phenotype, but the variant shows a strong defect, the variant may be a GOF allele. We recommend the readers first focus on ubiquitous drivers and easily scorable phenotypes (lethality, sterility, morphological phenotypes), and move on to tissue specific drivers and more specific phenotypes. We also recommend the readers to test the flies in different temperatures ranging between 18°C to 29°C because the UAS/GAL4 system is known to be temperature-dependent^{64,65}. Typically, the expression of UAS transgenes are higher at higher temperatures.

3.3 Perform additional functional studies related to the genes/protein of interest. In addition to examining general defects, one can select an assay system to probe into the molecular function of the gene and variant. In one of the example discussed under “Representative Results” section (*TBX2* case), we used ERG recordings to determine the effect of the variant on photoreceptor function since the fly gene of interest (*bifid*) had been studied extensively in the context of visual system development. Here, we describe a general outline of how to carry out such experiment in flies that over-express a reference or variant form of a human protein of interest in photoreceptors. Detailed protocol for ERG in *Drosophila* can be found in the following papers^{66–}

⁶⁸.

3.3.1 Set up crosses to generate flies to test for functional defects in the visual system. One can use Rh1-GAL4 to drive the reference or variant UAS-human cDNA transgenes in the R1-R6 photoreceptors by a single cross (Rh1-GAL4 virgin females x UAS-human cDNA males) to obtain the flies for ERG testing (Rh1-GAL4/+; UAS-human cDNA/+). We typically cross 3-5 virgin females to 3-5 male flies in a single vial and transfer the crosses every 2-3 days to have many animals eclosing from a single cross. The crosses are kept in an incubator set at the experimental temperature.

3.3.2 Once flies begin to eclose (at 25°C, ~10 days after setting the initial cross), separate them from the remaining pupae and place them in fresh vials. Place them back into the incubator set at the experimental temperature for an additional 3 days. We recommend recording ERGs on 3-5 day old flies since flies that are newly eclosed may still have large fluctuations in their ERG signal. If one wants to examine an age-dependent phenotype, these flies can be aged for several weeks as long as they are regularly (e.g. every ~5 days) transferred to a new vial to avoid the flies from drowning in wet food.

3.3.3 Prepare the flies for ERG recording by first anesthetizing the flies using CO₂ or placing them into a vial on ice. Gently glue one side of the fly onto a glass microscope slide to immobilize them. Multiple reference and variant flies can be glued on to a single slide. Place all flies in approximately the same orientation with one eye being accessible for the recording electrode. Be careful not to get glue on the eye and to leave the proboscis free.

3.3.4 Prepare the electrodes: Place a glass capillary into a needle puller (e.g. NARISHIGE Model PP-830) and switch on the filament. As soon as the weight drops, turn off the puller and detach the pulled capillary tube from the machine. This procedure will break the capillary tube to obtain two sharp tapered electrodes. The settings of the puller should be adjusted to obtain sharp tapered ends on the capillaries according to your system.

3.3.5 Fill the capillaries with saline solution (100 mM NaCl), making sure there are no air bubbles. Slide the glass capillaries over the silver wire electrodes (both the recording electrode and reference electrode, see **Figure 4**) and secure the capillaries in place.

3.3.6 Configure the stimulator and amplifier. Detailed set up can be found in Lauwers *et al.*⁶⁷ Our set up consists of the following equipment:

- Iso-Dam Isolated Biological Amplifier (World Precision Instruments, Sarasota, FL, USA): Set the amplifier to 0.1 Hz high pass filter, 300 Hz low pass filter, and 100 gain.
- S48 Stimulator (Astro-Med Inc. GRASS Instrument Division; West Warwick, RI, USA): Set the stimulator to 1 s period, 500 ms pulse width, 500 ms pulse delay, run mode, and 7 amplitude.
- Light source: We use a halogen light source (ACE Light Source, SCHOTT North America Inc., Southbridge, MA, USA) to stimulate the fly photoreceptors
- Axoscope 10.5 data acquisition software (Molecular Devices, San Jose, CA, USA): Create a stimulation protocol with acquisition model “fixed length events” and 20 s duration.

3.3.7 Acclimate the flies to complete darkness before initiating the ERG recordings. We typically place the flies into complete darkness for at least 10 minutes before beginning the experiment. Place the slide containing the flies onto the recording apparatus.

Note: Since flies cannot see red light, one can use a red light source during the period of dark habituation.

3.3.8 Move the micromanipulators carrying the reference and recording electrodes to a point that is close to the fly of interest on the slide. Watch the tip of the electrode and carefully place the reference electrode into the thorax of the fly. The exact position of this reference electrode does not have a major impact on the ERG signal. Then, place the recording electrode on the surface of the eye. ERG is a field recording so the recording electrode should be placed at the surface of the eye. The perfect amount of pressure will cause a small dimple, but should not penetrate the eye.

3.3.9 Turn off all lights for another 3 minutes to acclimate the flies again to the dark environment. In the Axoscope software, press play. If using a halogen light source with a manual shutter, turn on the light source at this point with the shutter closed (flies are still in dark). Next press record in the Axoscope and expose the fly eyes to light by opening and closing the shutter every 1 second for the 20 second duration of a single run. We control the on/off of the halogen light source manually but this can be programmed to have it automated using a white LED light source. In our experience, however, we have obtained much more robust and reliable ERG by using a halogen light source compared to a white light LED, likely due to the broader light spectrum emitted from the halogen light source.

3.3.10 Record ERGs from all of the flies that are mounted on the glass slide. Typically, we perform ERGs from 15 flies per genotype per condition. Parameters that can be altered to find a condition that shows robust differences between reference and variant cDNAs may include temperature, age, or environmental conditions (e.g. reared in light-dark cycle or constant light/darkness).

3.3.11 Perform data analysis: Compare the ERGs from the reference, variant, and controls to determine if there are differences. ERGs can be assessed for changes in on-transients, depolarization, off-transients, and repolarization⁶⁹ (**Figure 4B**). Depolarization and repolarization reflects the activation and inactivation of the phototransduction cascade within the photoreceptors, whereas the on- and off- transients are measures of the activities of post-synaptic cells that receive signals from the photoreceptors. Decreased amplitude and altered kinetics of repolarization are often associated in defects with photoreceptor function and health, whereas defect in on- an off-transients are found in mutants with defective synapse development, function or maintenance⁷⁰.

Note: Upon identification of differences in ERG phenotypes with over-expression of reference versus variant human cDNAs, one can further determine whether this electrophysiological phenotype is associated with structural and ultrastructural defects in photoreceptors and its synapses by performing histological analysis as well as transmission electron microscopy. Further

discussion on interpretation of ERG defects and structural/ultrastructural analysis can be found in the following article⁶⁹.

Representative Results:

1. Functional Study of a *de novo* missense variant in *EBF3* linked to neurodevelopmental phenotypes

In a 7-year-old male with neurodevelopmental phenotypes including hypotonia, ataxia, global developmental delay and expressive speech disorder, physicians and human geneticists at the National Institutes of Health Undiagnosed Diseases Project (UDP) identified a *de novo* missense variant (p.R163Q) in *EBF3* (*Early B-Cell Factor 3*)¹⁵, a gene that encodes a COE (Collier/Olfactory-1/Early B-Cell Factor) family transcription factor. This case was submitted to the UDN MOSC in March 2016 for functional studies of this variant. To assess whether this gene was a good candidate for this case, the MOSC gathered human genetic and genomic information from OMIM (www.omim.org/), ClinVar (www.ncbi.nlm.nih.gov/clinvar/), ExAC [exac.broadinstitute.org/] (now expanded to gnomAD, gnomad.broadinstitute.org/), Geno2MP (geno2mp.gs.washington.edu/Geno2MP/#/), DGV (dgv.tcag.ca/dgv/app/home), and DECIPHER (decipher.sanger.ac.uk/). In addition, we identified the orthologous genes in key MO species using the DIOPT tool (www.flyrnai.org/cgi-bin/DRSC_orthologs.pl), and further obtained gene expression and phenotypic information from individual MO databases [e.g. Wormbase (www.flyrnai.org/cgi-bin/DRSC_orthologs.pl), FlyBase (flybase.org/), ZFIN (zfin.org/) and MGI (www.informatics.jax.org/)]. Our gene variant interpretation methodology used for *EBF3* and other pioneering studies formed the basis for the later development of the MARRVEL resource (marrvel.org/) in 2017³⁰.

The information gathered from this methodology indicated *EBF3* was not associated with any known human genetic disorder at the time of analysis, and we concluded that the p.R163Q variant was a good candidate for this case based on the following information. (1) This variant has not been previously reported in control population databases (ExAC) and disease population database (Geno2PM), indicating that this is a very rare variant. (2) Based on ExAC, pLI (probability of LOF intolerance) score of this gene is 1.00 (pLI score ranges from 0.00 to 1.00). This indicates that there is a selective pressure against LOF variants in this gene in the general population and suggests that haploinsufficiency of this gene may cause disease. For more information on pLI score and its interpretation, please refer to the accompanying MARRVEL tutorial article in JoVE³¹ as well as related papers^{30,71}. (3) The p.R163Q variant is located in the evolutionarily conserved COE DNA binding domain of this protein, suggesting that it may affect DNA binding or other protein function. (4) The p.R163 residue is evolutionarily conserved from *C elegans* and *Drosophila* to human, suggesting that it may be critical for protein functional across species. (5) *EBF3* orthologs have been implicated in neuronal development in multiple MO⁷² including *C elegans*⁷³, *Drosophila*⁷⁴, *Xenopus*⁷⁵ and mice⁷⁶. (6) During brain development in mice, *Ebf3* was

shown to function downstream of *Arx* (*Aristaless-related homeobox*)⁷⁷, a gene known to be associated with several epilepsy and intellectual disability syndromes in human⁷⁸. Hence, these data together suggested that *EBF3* is highly likely to be crucial to human neurodevelopment and that the p.R163Q variant may have functional consequences.

To assess whether p.R163Q affects EBF3 function, a T2A-GAL4 line for *knot* (*kn*), the fly ortholog of human *EBF3*⁷⁹ was generated via generated via RMCE of a coding intronic MiMIC cassette¹⁵. The *kn*^{T2A-GAL4} line was recessive lethal and failed to complement the lethality of a classic *kn* allele (*kn*^{col-1}) as well as a molecularly defined deficiency that covers *kn* [*Df*(2R)*BSC429*]⁸⁰. Expression pattern of the GAL4 also reflected previously reported patterns of *kn* expression in the brain as well as in the wing imaginal disc¹⁵. UAS transgenic flies were generated to allow the expression of reference and variant human EBF3 cDNA as well as a wild-type fly *kn* cDNA. All three proteins were tagged with a C' 3xHA tag. Importantly, UAS-wild-type fly *kn* (*kn*⁺) or reference human *EBF3* (*EBF3*⁺) transgenes rescued the lethality of *kn*^{T2A-GAL4}/*Df*(2R)*BSC429* to a similar extent (**Figure 3C, left panel**)⁸¹. In contrast, UAS-human *EBF3* transgene with the p.R163Q variant (*EBF3*^{p.R163Q}) was not able to rescue this mutant, suggesting that the p.R163Q variant affects EBF3 function *in vivo*¹⁵. Interestingly, when assessed using an anti-HA antibody, the EBF3^{p.R163Q} protein was successfully expressed in the fly tissues and its levels and subcellular localization (primarily nuclear) was indistinguishable from that of EBF3⁺ and *Kn*⁺. This suggests that the variant is not causing a LOF phenotype due to protein instability or mis-localization. To further assess whether the p.R163Q variant affected the transcriptional activation function of *EBF3*, a luciferase based reporter assay was performed in HEK293 cells¹⁵. This experiment in cultured human cells revealed that the EBF3^{p.R163Q} variant failed to activate transcription of the reporter constructs, supporting the LOF model obtained from *Drosophila* experiments.

In parallel to the experimental studies, collaborations with physicians, human geneticists, and genetic counselors at BCM identified two additional individuals with similar symptoms. One patient carried the identical p.R163Q variant, and another patient varied a missense variant that affected the same residue (p.R163L). The p.R163L variant also failed to rescue the fly *kn* mutant⁹³ suggesting that this allele also affected *EBF3* function. Interestingly, this work was published back-to-back with two independent studies that reported additional individuals with *de novo* missense, nonsense, frameshift and splicing variants in *EBF3* linked to similar neurodevelopmental phenotypes^{82,83}. Subsequently, three additional papers were published reporting additional cases of *de novo* *EBF3* variants and copy number deletion^{84–86}. This novel neurodevelopmental syndrome is now known as the 'Hypotonia, Ataxia, and Delayed Development Syndrome (HADDs, OMIM #617330)' in the Online Mendelian Inheritance in Man (OMIM, www.omim.org), an authoritative database for genotype-phenotype relationships in human.

2. Functional Study of a dominantly inherited missense variant in *TBX2* linked to a syndromic cardiovascular and skeletal developmental disorder

In a small family with affected with overlapping spectrum of craniofacial dysmorphisms, cardiac anomalies, skeletal malformations, immune deficiency, endocrine abnormalities and developmental impairments, the UDN Duke Clinical Site identified a missense variant (p.R20Q) in *TBX2* that segregates with disease phenotypes⁸⁷. Three (son, daughter and mother) out of four family members are affected by this condition, and the son exhibited the most severe phenotype. Clinically, he met a diagnosis of ‘complete DiGeorge syndrome’, a condition that is often caused by haploinsufficiency of *TBX1*. While there were no mutation identified in *TBX1* in this family, the clinicians and human geneticists focused on a variant in *TBX2* since previous studies in mice showed that these genes have overlapping functions during development⁸⁸. *TBX1* and *TBX2* both belong to T-box (TBX) family of transcription factors that can act as transcriptional repressors as well as activators depending on the context. Previously, variants in 12 out of 17 members of the *TBX* family genes were linked to human diseases. The MOSC decided to experimentally pursue this variant based on the following information gathered through MARRVEL and other resources. (1) This variant was reported only once in a cohort of ~90,000 ‘control’ individuals in gnomAD (note that this variant was filtered out in a default view, likely due to low coverage reads). Considering the milder phenotypic presentation of the mother, this still can be considered as a very rare variant that may be responsible for the disease phenotypes. (2) The pLI score of *TBX2* in ExAC/gnomAD are 0.96/0.99 which is high (Max for pLI is 1.00). In addition, the o/e (observed/expected) LOF score in gnomAD is 0.05 (only 1/18.6 expected LOF variant is observed in gnomAD). These numbers suggest that LOF variants in this gene are selected against in the general population. (3) The p.R20 is evolutionarily conserved from *C elegans* and *Drosophila* to human, suggesting that this may be an important residue for *TBX2* function. (4) Multiple programs predict that the variant is likely damaging. Polyphen: Possibly/Probably Damaging, SIFT: Deleterious, CADD Score: 24.4, REVEL Score: 0.5. (5) MO mutants exhibit defects in tissues affected in patients (e.g. knockout mice exhibit defects in cardiovascular system, digestive/alimentary systems, craniofacial, limbs/digit). Hence, together with the biological links between *TBX1* and *TBX2* and the phenotypic links between our patients and DiGeorge Syndrome, we decided to perform functional studies of variants in this gene using *Drosophila*.

To begin to assess whether the p.R20Q variant affects *TBX2* function, we first generated a T2A-GAL4 line in *bifid* (*bi*), the *Drosophila* ortholog of human *TBX2*, via RMCE of a coding intronic MiMIC (**Figure 2**)⁸⁷. This allele, *bi*^{T2A-GAL4}, was recessive pupal lethal and behaved as a strong LOF mutant similar to previously reported *bi* LOF alleles (e.g. *bi*^{D2}, *bi*^{D4}) (**Figure 2E**). We were able to rescue the lethality of *bi*^{T2A-GAL4} as well as other *bi* alleles tested using an ~80kb genomic rescue construct carrying the entire *bi* locus, indicating that these are indeed clean LOF alleles. The expression pattern of GAL4 in the *bi*^{T2A-GAL4} line also matched well with previously reported patterns of *bi* expression in multiple tissues including in the wing imaginal disc (**Figure 2D**). In parallel, we generated UAS-transgenic lines for *TBX2* carrying the reference or variant (p.R20Q) sequences. Unfortunately, both transgenes were not able to rescue lethality of the *bi*^{T2A-GAL4} line. Importantly, we also found that a wild-type fly UAS-*bi* transgene also failed to rescue the *bi*^{T2A-}

GAL4 allele, likely due to the dosage-sensitivity of this gene. Indeed, over-expression of *UAS-bi*⁺ as well as *UAS-TBX2*⁺ and *UAS-TBX2*^{p.R20Q} caused some degree of lethality when overexpressed in a wild-type animal. We decided to use this toxic effect of *bi/TBX2* over-expression as a functional assay to assess whether the p.R20Q affects TBX2 function. Since the *Drosophila bi* gene has been extensively studied in the context of the visual system (gene is also known as [*optomotor blind* (*omb*)]), we decided to primarily focus on phenotypes related to the eye. When we expressed reference *TBX2* using an *ey-GAL4* driver that expresses UAS-transgenes in the eye as well as in parts of the brain relevant to the visual system, we observed ~85% lethality (**Figure 3C, right panel**) and significant reduction of eye size (**Figure 4B**). This phenotype was stronger than the phenotype observed when a wild-type fly *UAS-bi* transgene was expressed, suggesting that the human TBX2 causes is more detrimental to the fly when overexpressed. Interestingly, the p.R20Q TBX2 was less potent in causing lethality (**Figure 3C, right panel**) as well as inducing a small eye phenotype (**Figure 4B**) using the same driver under the identical condition⁸⁷, suggesting the variant affects protein function. Moreover, when we assessed the function of photoreceptors over-expressing reference and variant *TBX2* using a different GAL4 driver [*Rhodopsin 1 (Rh1)-GAL4* that specifically expresses UAS transgenes in R1-R6 photoreceptors), we also observed that variant TBX2 exhibited a much milder ERG phenotype compared to reference TBX2 (**Figure 4B**)⁸⁷. Interestingly, most of the p.R20Q TBX2 protein was still found in the nucleus similar to the reference protein, suggesting that the variant did not affect nuclear localization. When we performed a luciferase based transcription repression assay in HEK293T cells, we found that the p.R20Q was not able to effectively repress transcription of a reporter construct with palindromic T-box sites⁸⁷. In addition, we observed a decrease in protein levels of TBX2^{p.R20Q} compared to TBX2⁺, suggesting that the variant may affect translation or protein stability of TBX2, which in turn affects its abundance within a cell.

In parallel to these experimental studies, we attempted to identify additional patients with rare variants in *TBX2*. Through GeneMatcher, we identified an 8-year-old boy with a *de novo* missense (p.R305H) variant from an unrelated family who exhibited many of the features found in the first family⁸⁷. Additional functional studies in *Drosophila* and human cell line revealed that the p.R305H variant also affects TBX2 function and protein levels, strongly suggesting that defect in this gene is likely to underlie many of the phenotypes found in the two families. This disorder has been recently curated as 'Vertebral anomalies and variable Endocrine and T-cell Dysfunction (VETD, OMIM #618223)' in OMIM. Identification of additional individuals with functional variants in *TBX2* with overlapping phenotypes will be critical to establish the full spectrum of genotype-phenotype relationship for this gene in human disease.

Discussion:

Experimental studies using *Drosophila melanogaster* provides a robust assay system to assess functional consequences of disease associated human variants, thanks to the large body of knowledge and diverse genetic tools that have been generated by many researchers in the fly field over the past century⁸⁹. Just like any other experimental systems, however, it is important to acknowledge the caveats and limitation when using this system.

Caveats associated with data mining

Although the first step in this protocol is to mine databases for information pertaining to a gene of interest, it is important to use this information as a starting point and not as solid evidence. For example, although *in silico* prediction of variant function provides valuable insights, these data should always be interpreted with caution. There are some instances in which all major algorithms predict that a human variant is benign, yet functional studies in *Drosophila* clearly demonstrated the functionality of such variant²⁴. Similarly, although protein-protein interaction, co-expression and structural modeling data are all insightful pieces of information, there may be pseudo-positive and pseudo-negative information present in these large ‘omics’ data sets. For example, some of the previously identified or predicted protein-protein interactions may be artificial or only seen in certain cell or tissue types. In addition, there may be many false negative interactions that are not captured in these data sets since certain key protein-protein interactions are transient (e.g. enzyme-substrate interactions). Experimental validation is critical to demonstrate that certain genes or proteins genetically or physically interact *in vivo* and in the biological context of interest. Similarly, structures predicted based on homology modeling should only be treated as a ‘model’ rather than a concrete structure. Although this information could be useful if one finds that an amino acid of interest is present in a structurally important part of the protein, negative data does not rule out the possibility that the variant may be functional. Finally, some of the previously reported genotype-phenotype information may also need to be treated with caution since some information archived in public database may not be accurate. For example, some information in MO databases are based on experiments that have been well controlled and performed rigorously, whereas others may have been one of many hits that are described in a large screen paper without additional follow-up studies with stringent controls.

‘Humanization’ experiments using T2A-GAL4 strategy may not always be successful

While rescue and over-expression based functional studies using human cDNAs allows assessments of variants in the context of the human protein, this approach is not always successful. If a reference human cDNA cannot rescue the fly mutant phenotype, there are two possible explanations. The first possibility is that the human protein is nonfunctional or has significantly reduced activity in the context of a fly cell. This could be due to reduced protein expression, stability, activity and/or localization, or could be due to the lack of compatibility with fly proteins that work in a multi-protein complex. Since the UAS/GAL4 system is temperature sensitive, one can raise the flies at a relatively high temperature (e.g. 29°C) to see if one may be able to see a rescue in this condition. In addition, one can also generate a UAS-fly cDNA construct and transgene as a positive control. If the variant of interest affects a conserved amino acid, the analogous variant can be introduced into the fly cDNA for functional study of the variant in the

context of the fly ortholog. Although this is not absolutely necessary, it greatly helps the study in case the experiments using human cDNA transgenic line gives negative or inconclusive results (**Figure 3**). The second possibility is that the expression of the human protein causes some sort of cellular or organism level toxicity. This could be due to an antimorphic effect (e.g. acting as a dominant negative protein), hypermorphic effect (e.g. too much activity), or neomorphic effect (e.g. gain of toxic function such as protein aggregation). In this case, keeping the flies in a low temperature (e.g. 18°C) may alleviate some of these problems. Importantly, if the human cDNA causes a gain of toxic function phenotype, we can take advantage of this and use this specific phenotype to assess the variant function as described in Section 3. Finally, there are some scenarios in which the over-expression of a fly cDNA may not rescue the fly T2A-GAL4 line as we have seen in the TBX2 example, likely due to the strict dosage dependence of the gene product. To avoid the over-expression of a protein of interest, one can modify the fly gene of interest via CRISPR or engineer a genomic rescue construct that contains the variant of interest and perform rescue experiments using a LOF allele²¹. For small genes, one can also consider ‘humanizing’ the fly genomic rescue construct to test human variants that affects non-conserved amino acids²⁴.

Things to note when interpreting negative and positive results

If both the reference and variant human cDNAs rescues the fly mutant phenotypes to a similar degree, and there is no difference observed in all conditions tested, we conclude that the variant is functionally indistinguishable in *Drosophila in vivo*. It is important to note that this information is not sufficient to rule out that the variant of interest is non-pathogenic since the *Drosophila* assay may not be sensitive enough or may not capture all potential functions of the gene/protein of interest that matter in humans. Positive data, on the other hand, is a strong indication that the variant has functional consequences, but is not sufficient to claim pathogenicity. American College of Medical Genetics and Genomics (ACMG) has published a set of standards and guidelines to classify variants in human disease associated genes into “benign”, “likely benign”, “variant of unknown significance (VUS)”, “likely pathogenic” and “pathogenic”⁹⁰. Although this classification only applies to established disease-associated genes and not directly applicable to variants in ‘genes of uncertain significance (GUS)’, we strongly encourage all individuals who are involved in human variant functional studies to read and adhere to this guideline when reporting variant function.

Extracting useful biological information when MO phenotypes do not ‘model’ the human disease condition

It is important to keep in mind that over-expression based functional assays have their own limitations, especially since some of the phenotypes being scored may have little relevance to the disease condition of interest. Similarly, the phenotypes that are being assessed through rescue experiments may not have any direct relevance to the disease of interest. Since these experiments are conducted outside the endogenous contexts in an invertebrate system, they should not be considered as a disease models but rather as a gene functional test using a ‘living test tube’. Scorable phenotypes used in rescue experiments can often provide biological insights into the disease conditions. The concept of ‘phenologs (non-obvious homologous phenotypes)’

771 (www.phenologs.org)⁹¹ can be used to further determine the underlying molecular connection
772 between the *Drosophila* and human phenotypes. For example, morphological phenotypes in the
773 fly wing are excellent phenotypic readouts for defects in Notch signaling pathway, an
774 evolutionarily conserved pathway linked to many congenital disorders including cardiovascular
775 defects in humans⁶². By understanding the molecular logic behind certain phenotypes in
776 *Drosophila*, one may identify hidden biological links between genes and phenotypes in humans
777 that have yet to be understood.

778

Acknowledgements:

We thank Jose Salazar and Julia Wang for critical reading of the manuscript. Undiagnosed Diseases Network Model Organisms Screening Center was supported through the National Institutes of Health (NIH) Commonfund (U54 NS093793). HTC was further supported by the CNCDP-K12 and NINDS (1K12 NS098482), American Academy of Neurology (Neuroscience Research grant), Burroughs Wellcome Fund (Career Award for Medical Scientists), Child Neurology Society and Child Neurology Foundation (PERF Elterman grant), and the NIH Director's Early Independence Award (DP5 OD026426). MFW was further supported by Simons Foundation (SFARI Award: 368479). SY was further supported by the NIH (R01 DC014932), the Simons Foundation (SFARI Award: 368479), the Alzheimer's Association (New Investigator Research Grant: 15-364099), Naman Family Fund for Basic Research and Caroline Wiess Law Fund for Research in Molecular Medicine. Confocal microscopy at BCM is supported in part by NIH Grant U54HD083092 to the Intellectual and Developmental Disabilities Research Center (IDDRC) Neurovisualization Core.

Disclosures:

The authors have nothing to disclose.

Online Resources:

Variant function prediction algorithms

PolyPhen-2: <http://genetics.bwh.harvard.edu/pph2>

SIFT: <https://sift.bii.a-star.edu.sg>

CADD: <https://cadd.gs.washington.edu>

PROVEAN: <http://provean.jcvi.org/index.php>

MutationTaster: <http://www.mutationtaster.org>

REVEL <https://sites.google.com/site/revelgenomics>

Rare and undiagnosed disease research consortiums

UDN: <https://undiagnosed.hms.harvard.edu>

RDMM: <http://www.rare-diseases-catalyst-network.ca>

IRUD: <https://irudbeyond.nig.ac.jp/en/index.html>

SOLVE-RD: <http://solve-rd.eu>

Australian Functional Genomics Network: <https://www.functionalgenomics.org.au>

Integrative database for human and model organism Information

MARRVEL: <http://marrvel.org>

Monarch Initiative: <https://monarchinitiative.org>

Gene2Function: <http://www.gene2function.org>

Phenologs: <http://www.phenologs.org>

Human Genetic and Genomics Databases

OMIM: <https://www.omim.org/>

820 ClinVar: <https://www.ncbi.nlm.nih.gov/clinvar/>)
 821 ExAC: <http://exac.broadinstitute.org/>
 822 gnomAD: <http://gnomad.broadinstitute.org/>
 823 Geno2MP: <http://geno2mp.gs.washington.edu/Geno2MP/#/>
 824 DGV: <http://dgv.tcag.ca/dgv/app/home>
 825 DECIPHER: <https://decipher.sanger.ac.uk/>
 826
 827 Ortholog Identification Tool
 828 DIOPT: https://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl
 829
 830 Model Organism Databases and Pubmed
 831 Wormbase (*C elegans*): https://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl
 832 FlyBase (*Drosophila*): <http://flybase.org>
 833 ZFIN (Zebrafish): <https://zfin.org>
 834 MGI (Mouse): <http://www.informatics.jax.org>
 835 Pubmed: <https://www.ncbi.nlm.nih.gov/pubmed/>
 836
 837 Genetic and protein interaction databases
 838 STRING: <https://string-db.org>
 839 MIST: <http://fgertools.hms.harvard.edu/MIST/>
 840
 841 Protein structure databases and modeling tools
 842 WWPBD: <http://www.wwpdb.org>
 843 SWISS-MODEL: <https://swissmodel.expasy.org/>
 844 Modeller: <https://salilab.org/modeller/>
 845 Phyre²: <http://www.sbg.bio.ic.ac.uk/phyre2>
 846
 847 Patient matchmaking platforms
 848 Matchmaker Exchange: <http://www.matchmakerexchange.org>
 849 GeneMatcher: <https://www.genematcher.org>
 850 AGHA Archive <https://mme.australiangenomics.org.au/#/home>
 851 matchbox: <https://seqr.broadinstitute.org/matchmaker/matchbox>
 852 DECIPHER: <https://decipher.sanger.ac.uk>
 853 MyGene²: <https://www.mygene2.org/MyGene2>
 854 Phenome Central: <https://phenomecentral.org>
 855
 856 Human transcript annotation and cDNA clone information
 857 Mammalian Gene Collection: <https://genecollections.nci.nih.gov/MGC>
 858 Ensembl: <http://useast.ensembl.org>
 859 Refseq: <http://www.ncbi.nlm.nih.gov/refseq>
 860
 861

862 **Table of Materials:**

<i>Drosophila</i> Stocks for UAS-human cDNA transgenesis		
VK33 (3 rd chromosome) injection line	BDSC	#24871
VK37 (2 nd chromosome injection) line	BDSC	#24872
Plasmid DNA		
pDONR221	Thermo Fisher	#12536-017
pGW-HA. attB	Gift from Drs. Johannes Bischof and Konrad Basler (Bischof et al., 2013 PNAS)	
Molecular biology kits and reagents		
Q5 Polymerase kit	NEB	#M0491
BP Clonase kit	Thermo Fisher	#11789020
LR Clonase II Enzyme kit	Thermo Fisher	#11791100
PureLink Gel Extraction Kit	Thermo Fisher	#K210012
Quick Change II Mutagenesis kit	Agilent	#200523
Agarose (molecular biology grade)	Sigma-Aldrich	#A2790
QIAprep Spin Miniprep Kit	Qiagen	#27104
DH5α	Thermo Fisher	#18265017
Electroretinogram Rig related equipment		
ISO-DAM Isolated Biologic Amplifier	LabX	#R150358
Square Pulse Stimulator	Astro-Med	#S48
Axon pCLAMP 10 Data Software Package	Molecular Devices	N/A

863

Figure Legends:

Figure 1: Injection and crossing scheme to generate UAS-human cDNA and T2A-GAL4 lines. **(A)** Generation of UAS-human cDNA transgenes through microinjections and crosses. Crossing scheme to integrate the transgenes into a 2nd chromosome docking site (VK37) using male flies in the 1st and 2nd generation are shown as an example. Upon injection of the human cDNA ϕ C31 transgenic construct (pGW-HA.attB) into early embryos that contain a germline source of ϕ C31 integrase (labeled with both 3xP3-GFP and 3xP3-RFP) and VK37 docking site [labeled with a *yellow*⁺ (*y*⁺) marker], one can follow the transgenic events with the *white*⁺ (*w*⁺) minigene that is present in the transgenic vector. We recommend the readers to cross out the ϕ C31 integrase by selecting against flies with GFP and RFP. The final stable stock can be kept as homozygotes or as a balanced stock if the chromosome carries a 2nd site lethal/sterile hit mutation. Presence of 2nd site lethal/sterile mutations on a transgenic constructs usually does not affect the outcome of functional studies as long as these transgenes are used in a heterozygous state (see **Figure 3**). **(B)** Generation of T2A-GAL4 lines through microinjection and crosses. Crossing scheme to convert a 2nd chromosome MiMIC insertion into a T2A-GAL4 element is shown here as an example. By microinjecting an expression vector for ϕ C31 integrase and a RMCE vector for T2A-GAL4 (pBS-KS-attB2-SA-T2A-Gal4-Hsp70, must select appropriate reading frame for the MiMIC of interest. See the following papers for details^{57,59}) into embryos carrying a MiMIC in a coding intron in gene of interest, one can convert the original MiMIC into a T2A-GAL4 line. See **Figure 2A** for a schematic diagram of the RMCE conversion. The conversion event can be selected by screening against the *y*⁺ marker in the original MiMIC cassette⁶⁰. Since RMCE can happen in two directions, only 50% of the successful conversion event will lead to successful production of GAL4, which can be detected by a UAS-GFP reporter transgene in the next generation. The final stable stock can be kept as homozygotes or as a balanced stock if the LOF of the gene is lethal/sterile.

Figure 2: Conversion of MiMIC elements into T2A-GAL lines via RMCE. **(A)** ϕ C31 integrase facilitates the recombination between the two *attP* sites in the fly **(A-top)** and the two *attB* sites flanking a T2A-GAL4 cassette shown as a circular vector **(A-bottom)**. **(B)** Successful RMCE event leads to a loss of a selectable marker (*yellow*⁺), and insertion of the T2A-GAL4 cassette in the same orientation of the gene of interest. Since the RMCE event can happen in two orientations, only 50% of the RMCE reaction will give a desired product. RMCE product inserted in the opposite orientation will not function as a gene-trap allele and will not express GAL4. The directionality of the construct must be confirmed via Sanger sequencing. **(C)** Transcription **(C-top)** and translation **(C-bottom)** of the gene of interest leads to generation of a truncated mRNA and protein due to the polyA signal that present at the 3' end of the T2A-GAL4 cassette. The T2A is a ribosome skipping signal, which allows the ribosome to halt and reinitiate translation after this signal. This is used to generate a GAL4 element that is not covalently attached to the truncated gene product of interest. The GAL4 will enter the nucleus and will facilitate the transcription of transgenes that are under the control of UAS elements. UAS-GFP can be used as a gene expression reporter, and

UAS-human cDNA can be used for rescue experiments via gene ‘humanization’. **(D)** Example of a T2A-GAL4 element in *bi* driving expression of UAS-GFP shown on the top. This expression pattern resembles a previously generated enhancer trap line for the same gene (*bi^{omb-GAL4}*) shown on the bottom. **(E)** Comparison of T2A-GAL4 allele of *bi* with previously reported LOF *bi* alleles. This figure has been adopted and modified from ^{57,87}.

Figure 3: Functional Analysis of human variants using rescue-based (left) and over-expression (right) studies. **(A-left panel)** The function of *EBF3* variants was assessed with a rescue-based analysis of the fly *knot (kn)* LOF allele focusing on lethality/viability. **(A-right panel)** The function of variants in *TBX2* was assessed by performing over-expression of human *TBX2* transgenes in wild-type flies, focusing on lethality/viability as well as eye morphology and electrophysiology phenotypes (see **Figure 4**). **(B)** Crossing schemes to obtain the flies that would be tested in the functional studies. One should always use a neutral UAS element (e.g. *UAS-lacZ*, *UAS-GFP*) as a control experiment. **(C)** Representative results from functional studies of *EBF3^{p.R163Q}* and *TBX2^{p.R20Q}* variants, respectively, along with appropriate control experiments that are necessary to interpret the results. Both the rescue-based analysis and over-expression studies reveal that the variants behave as amorphic or hypomorphic alleles. The lethality/viability data shown here are based on the experimental data presented in ^{15,87}.

Figure 4: Functional analysis of a rare missense variant in human *TBX2* based on eye morphology and electroretinogram in *Drosophila*. **(A)** A schematic image showing the typical placement of recording and reference electrodes on the fly eye along with a representative electroretinogram recording with four major components (on-transient, depolarization, off-transient, repolarization). **(B)** *TBX2* variant (p.R20Q) functions as a partial LOF allele based on over-expression studies in the fly eye using GAL4 drivers specific to the visual system (*ey-GAL4* and *Rh1-GAL4*) showed that the reference *TBX2* caused a strong morphological and electrophysiological phenotype compared to the variant protein. **(B-top panels)** A severe reduction in eye size is seen upon over-expression of *UAS-TBX2⁺* with *ey-GAL4*. *UAS-TBX2^{p.R20Q}* driven with *ey-GAL4* also causes a smaller eye but the phenotype is much milder. **(B-bottom panels)** When *UAS-TBX2⁺* is expressed in core R1-R6 photoreceptors using *Rh1-GAL4*, there is a loss of the on-transient and off-transient, reduced depolarization, and a large abnormal prolonged depolarization after potential (PDA) phenotype that is not seen in control flies. These phenotypes are not as severe when *UAS-TBX2^{p.R20Q}* is expressed using the same *Rh1-GAL4*. This figure has been adopted and modified from ^{69,87}.

References Cited

1. Boycott, K. M., Rath, A., *et al.* International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *The American Journal of Human Genetics* **100**, (5)695–705 (2017).
2. Lupski, J. R., Reid, J. G., *et al.* Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy. *New England Journal of Medicine* **362**, (13)1181–1191 (2010).
3. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics* **14**, (10)681–691 (2013).
4. Yang, Y., Muzny, D. M., *et al.* Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. *JAMA* **312**, (18)1870 (2014).
5. Lee, H., Deignan, J. L., *et al.* Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. *JAMA* **312**, (18)1880 (2014).
6. Coban-Akdemir, Z., White, J. J., *et al.* Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. *The American Journal of Human Genetics* **103**, (2)171–187 (2018).
7. Muller, H. J. Further studies on the nature and causes of gene mutations. *Proceedings of the Sixth International Congress of Genetics* 213–255 (1932).
8. Ghosh, R., Oak, N. & Plon, S. E. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biology* **18**, (1)225 (2017).
9. Adzhubei, I. A., Schmidt, S., *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, (4)248–249 (2010).
10. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nature Protocols* **11**, (1)1–9 (2016).
11. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* (2018).doi:10.1093/nar/gky1016
12. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PloS one* **7**, (10)e46688 (2012).
13. Wangler, M. F., Yamamoto, S., *et al.* Model Organisms Facilitate Rare Disease Diagnosis and Therapeutic Research. *Genetics* **207**, (1)9–27 (2017).
14. Oriel, C. & Lasko, P. Recent Developments in Using Drosophila as a Model for Human Genetic Disease. *International Journal of Molecular Sciences* **19**, (7)2041 (2018).
15. Chao, H.-T., Davids, M., *et al.* A Syndromic Neurodevelopmental Disorder Caused by De Novo Variants in EBF3. *American journal of human genetics* **100**, (1)128–137 (2017).
16. Oláhová, M., Yoon, W. H., *et al.* Biallelic Mutations in ATP5F1D, which Encodes a Subunit

- 973 of ATP Synthase, Cause a Metabolic Disorder. *American journal of human genetics* **102**,
974 (3)494–504 (2018).
- 975 17. Liu, N., Schoch, K., *et al.* Functional variants in TBX2 are associated with a syndromic
976 cardiovascular and skeletal developmental disorder. *Human molecular genetics* **27**,
977 (14)2454–2465 (2018).
- 978 18. Marcogliese, P. C., Shashi, V., *et al.* IRF2BPL Is Associated with Neurological Phenotypes.
979 *American journal of human genetics* **103**, (2)245–260 (2018).
- 980 19. Ferreira, C. R., Xia, Z.-J., *et al.* A Recurrent De Novo Heterozygous COG4 Substitution
981 Leads to Saul-Wilson Syndrome, Disrupted Vesicular Trafficking, and Altered
982 Proteoglycan Glycosylation. *The American Journal of Human Genetics* **103**, (4)553–567
983 (2018).
- 984 20. Kanca, O., Andrews, J., *et al.* De novo variants in WDR37 are associated with epilepsy,
985 colobomas and cerebellar hypoplasia. *Americal Journal of Human Genetics* **Submitted**,
986 (2019).
- 987 21. Luo, X., Rosenfeld, J. A., *et al.* Clinically severe CACNA1A alleles affect synaptic function
988 and neurodegeneration differentially. *PLOS Genetics* **13**, (7)e1006905 (2017).
- 989 22. Chung, H., Wangler, M., *et al.* ACOX1 induces autoimmunity whereas a *de novo* gain of
990 function variant induces elevated ROS and glial loss in humans and flies. *Cell Metabolism*
991 **Submitted**, (2019).
- 992 23. Yamamoto, S., Jaiswal, M., *et al.* A Drosophila Genetic Resource of Mutants to Study
993 Mechanisms Underlying Human Genetic Diseases. *Cell* **159**, (1)200–214 (2014).
- 994 24. Jakobsdottir, J., van der Lee, S. J., *et al.* Rare Functional Variant in TM2D3 is Associated
995 with Late-Onset Alzheimer’s Disease. *PLoS genetics* **12**, (10)e1006327 (2016).
- 996 25. Yoon, W. H., Sandoval, H., *et al.* Loss of Nardilysin, a Mitochondrial Co-chaperone for α -
997 Ketoglutarate Dehydrogenase, Promotes mTORC1 Activation and Neurodegeneration.
998 *Neuron* **93**, (1)115–131 (2017).
- 999 26. Harel, T., Yoon, W. H., *et al.* Recurrent De Novo and Biallelic Variation of ATAD3A,
1000 Encoding a Mitochondrial Membrane Protein, Results in Distinct Neurological
1001 Syndromes. *American journal of human genetics* **99**, (4)831–845 (2016).
- 1002 27. Tan, K. L., Haelterman, N. A., *et al.* Ari-1 Regulates Myonuclear Organization Together
1003 with Parkin and Is Associated with Aortic Aneurysms. *Developmental Cell* **45**, (2)226–
1004 244.e8 (2018).
- 1005 28. Ansar, M., Chung, H., *et al.* Visual impairment and progressive phthisis bulbi caused by
1006 recessive pathogenic variant in MARK3. *Human Molecular Genetics* **27**, (15)2703–2711
1007 (2018).
- 1008 29. Ansar, M., Chung, H., *et al.* Bi-allelic Loss-of-Function Variants in DNMBP Cause Infantile
1009 Cataracts. *The American Journal of Human Genetics* **103**, (4)568–578 (2018).

- 1010 30. Wang, J., Al-Ouran, R., *et al.* MARRVEL: Integration of Human and Model Organism
1011 Genetic Resources to Facilitate Functional Annotation of the Human Genome. *The*
1012 *American Journal of Human Genetics* **100**, (6)843–853 (2017).
- 1013 31. Wang, J., Liu, Z., Bellen, H. & Yamamoto, S. MARRVEL, a web-based tool that integrates
1014 human and model organism genomics information. *Journal of Visualized Experiments*
1015 **Submitted**, (2019).
- 1016 32. Mungall, C. J., McMurtry, J. A., *et al.* The Monarch Initiative: an integrative data and
1017 analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids*
1018 *Research* **45**, (D1)D712–D722 (2017).
- 1019 33. Hu, Y., Comjean, A., Mohr, S. E., Perrimon, N. & Perrimon, N. Gene2Function: An
1020 Integrated Online Resource for Gene Function Discovery. *G3:*
1021 *Genes/Genomes/Genetics* **7**, (8)2855–2858 (2017).
- 1022 34. Ioannidis, N. M., Rothstein, J. H., *et al.* REVEL: An Ensemble Method for Predicting the
1023 Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics* **99**,
1024 (4)877–885 (2016).
- 1025 35. Szklarczyk, D., Morris, J. H., *et al.* The STRING database in 2017: quality-controlled
1026 protein–protein association networks, made broadly accessible. *Nucleic Acids Research*
1027 **45**, (D1)D362–D368 (2017).
- 1028 36. Hu, Y., Vinayagam, A., *et al.* Molecular Interaction Search Tool (MIST): an integrated
1029 resource for mining gene and protein interaction data. *Nucleic Acids Research* **46**,
1030 (D1)D567–D574 (2018).
- 1031 37. Lawson, C. L., Patwardhan, A., *et al.* EMDatBank unified data resource for 3DEM. *Nucleic*
1032 *Acids Research* **44**, (D1)D396–D403 (2016).
- 1033 38. Bienert, S., Waterhouse, A., *et al.* The SWISS-MODEL Repository—new features and
1034 functionality. *Nucleic Acids Research* **45**, (D1)D313–D319 (2017).
- 1035 39. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Current*
1036 *Protocols in Bioinformatics* **54**, 5.6.1–5.6.37 (2016).
- 1037 40. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web
1038 portal for protein modeling, prediction and analysis. *Nature Protocols* **10**, (6)845–858
1039 (2015).
- 1040 41. Bamshad, M. J., Shendure, J. A., *et al.* The Centers for Mendelian Genomics: A new large-
1041 scale initiative to identify the genes underlying rare Mendelian conditions. *American*
1042 *Journal of Medical Genetics Part A* **158A**, (7)1523–1525 (2012).
- 1043 42. Sobreira, N. L. M., Arachchi, H., *et al.* Matchmaker Exchange. *Current Protocols in Human*
1044 *Genetics* **95**, 9.31.1–9.31.15 (2017).
- 1045 43. Temple, G., Gerhard, D. S., *et al.* The completion of the Mammalian Gene Collection
1046 (MGC). *Genome Research* **19**, (12)2324–2333 (2009).

- 1047 44. Katzen, F. Gateway[®] recombinational cloning: a biological operating system. *Expert*
1048 *Opinion on Drug Discovery* **2**, (4)571–589 (2007).
- 1049 45. Venken, K. J. T., He, Y., Hoskins, R. A. & Bellen, H. J. P[acman]: A BAC Transgenic Platform
1050 for Targeted Insertion of Large DNA Fragments in *D. melanogaster*. *Science* **314**,
1051 (5806)1747–1751 (2006).
- 1052 46. Bischof, J., Björklund, M., Furger, E., Schertel, C., Taipale, J. & Basler, K. A versatile
1053 platform for creating a comprehensive UAS-ORFeome library in *Drosophila*. *Development*
1054 (*Cambridge, England*) **140**, (11)2434–42 (2013).
- 1055 47. Bischof, J., Sheils, E. M., Björklund, M. & Basler, K. Generation of a transgenic ORFeome
1056 library in *Drosophila*. *Nature Protocols* **9**, (7)1607–1620 (2014).
- 1057 48. Laible, M. & Boonrod, K. Homemade site directed mutagenesis of whole plasmids.
1058 *Journal of visualized experiments : JoVE* (27) (2009).doi:10.3791/1135
- 1059 49. Balana, B., Taylor, N. & Slesinger, P. A. Mutagenesis and Functional Analysis of Ion
1060 Channels Heterologously Expressed in Mammalian Cells. *Journal of Visualized*
1061 *Experiments* (44) (2010).doi:10.3791/2189
- 1062 50. Bischof, J., Maeda, R. K., Hediger, M., Karch, F. & Basler, K. An optimized transgenesis
1063 system for *Drosophila* using germ-line-specific C31 integrases. *Proceedings of the*
1064 *National Academy of Sciences* **104**, (9)3312–3317 (2007).
- 1065 51. Ringrose, L. Transgenesis in *Drosophila melanogaster*. *Methods in molecular biology*
1066 (*Clifton, N.J.*) **561**, 3–19 (2009).
- 1067 52. Venken, K. J. T., He, Y., Hoskins, R. A. & Bellen, H. J. P[acman]: A BAC Transgenic Platform
1068 for Targeted Insertion of Large DNA Fragments in *D. melanogaster*. *Science* **314**,
1069 (5806)1747–1751 (2006).
- 1070 53. Groth, A. C., Fish, M., Nusse, R. & Calos, M. P. Construction of transgenic *Drosophila* by
1071 using the site-specific integrase from phage phiC31. *Genetics* **166**, (4)1775–82 (2004).
- 1072 54. Greenspan, R. *Fly Pushing: The Theory and Practice of Drosophila Genetics*. (Cold Spring
1073 Harbor Laboratory Press: Cold Spring Harbor, New York, 2004).
- 1074 55. Ashburner, M., Golic, K. & Hawley, R. S. *Drosophila: A Laboratory Handbook*. (Cold Spring
1075 Harbor Laboratory Press: Cold Spring Harbor, New York, 2005).
- 1076 56. Diao, F. & White, B. H. A Novel Approach for Directing Transgene Expression in
1077 *Drosophila*: T2A-Gal4 In-Frame Fusion. *Genetics* **190**, (3)1139–1144 (2012).
- 1078 57. Diao, F., Ironfield, H., *et al.* Plug-and-Play Genetic Access to *Drosophila* Cell Types using
1079 Exchangeable Exon Cassettes. *Cell Reports* **10**, (8)1410–1421 (2015).
- 1080 58. Bellen, H. J., Levis, R. W., *et al.* The *Drosophila* Gene Disruption Project: Progress Using
1081 Transposons With Distinctive Site Specificities. *Genetics* **188**, (3)731–743 (2011).
- 1082 59. Lee, P.-T., Zirin, J., *et al.* A gene-specific T2A-GAL4 library for *Drosophila*. *eLife* **7**, (2018).

- 1083 60. Venken, K. J. T., Schulze, K. L., *et al.* MiMIC: a highly versatile transposon insertion
1084 resource for engineering *Drosophila melanogaster* genes. *Nature methods* **8**, (9)737–43
1085 (2011).
- 1086 61. Li-Kroeger, D., Kanca, O., *et al.* An expanded toolkit for gene tagging based on MiMIC and
1087 scarless CRISPR tagging in *Drosophila*. *eLife* **7**, (2018).
- 1088 62. Salazar, J. L. & Yamamoto, S. Integration of *Drosophila* and Human Genetics to
1089 Understand Notch Signaling Related Diseases. *Advances in experimental medicine and*
1090 *biology* **1066**, 141–185 (2018).
- 1091 63. Wangler, M. F., Yamamoto, S. & Bellen, H. J. Fruit Flies in Biomedical Research. *Genetics*
1092 **199**, (3)639–653 (2015).
- 1093 64. Duffy, J. B. GAL4 system in *Drosophila*: A fly geneticist's swiss army knife. *genesis* **34**, (1–
1094 2)1–15 (2002).
- 1095 65. Nagarkar-Jaiswal, S., Lee, P.-T., *et al.* A library of MiMICs allows tagging of genes and
1096 reversible, spatial and temporal knockdown of proteins in *Drosophila*. *eLife* **4**, (2015).
- 1097 66. Dolph, P., Nair, A. & Raghu, P. Electroretinogram Recordings of *Drosophila*. *Cold Spring*
1098 *Harbor Protocols* **2011**, (1)pdb.prot5549-pdb.prot5549 (2011).
- 1099 67. Lauwers, E. & Verstreken, P. Assaying Mutants of Clathrin-Mediated Endocytosis in the
1100 Fly Eye. *Methods in molecular biology (Clifton, N.J.)* **1847**, 109–119 (2018).
- 1101 68. Rhodes-Mordov, E., Samra, H. & Minke, B. Electroretinogram (ERG) Recordings from
1102 *Drosophila*. *BIO-PROTOCOL* **5**, (21) (2015).
- 1103 69. Deal, S. & Yamamoto, S. Unraveling novel mechanisms of neurodegeneration through a
1104 large-scale forward genetic screen in *Drosophila*. *Frontiers in Genetics* **In press**, (2019).
- 1105 70. Chouhan, A. K., Guo, C., *et al.* Uncoupling neuronal death and dysfunction in *Drosophila*
1106 models of neurodegenerative disease. *Acta Neuropathologica Communications* **4**, (1)62
1107 (2016).
- 1108 71. Lek, M., Karczewski, K. J., *et al.* Analysis of protein-coding genetic variation in 60,706
1109 humans. *Nature* **536**, (7616)285–291 (2016).
- 1110 72. Liberg, D., Sigvardsson, M. & Akerblad, P. The EBF/Olf/Collier family of transcription
1111 factors: regulators of differentiation in cells originating from all three embryonal germ
1112 layers. *Molecular and cellular biology* **22**, (24)8389–97 (2002).
- 1113 73. Prasad, B. C., Ye, B., Zackhary, R., Schrader, K., Seydoux, G. & Reed, R. R. *unc-3*, a gene
1114 required for axonal guidance in *Caenorhabditis elegans*, encodes a member of the O/E
1115 family of transcription factors. *Development (Cambridge, England)* **125**, (8)1561–8
1116 (1998).
- 1117 74. Jinushi-Nakao, S., Arvind, R., Amikura, R., Kinameri, E., Liu, A. W. & Moore, A. W.
1118 Knot/Collier and Cut Control Different Aspects of Dendrite Cytoskeleton and Synergize to

1119 Define Final Arbor Shape. *Neuron* **56**, (6)963–978 (2007).

1120 75. Pozzoli, O., Bosetti, A., Croci, L., Consalez, G. G. & Vetter, M. L. Xebf3 is a regulator of
1121 neuronal differentiation during primary neurogenesis in *Xenopus*. *Developmental biology*
1122 **233**, (2)495–512 (2001).

1123 76. Wang, S. S., Lewcock, J. W., Feinstein, P., Mombaerts, P. & Reed, R. R. Genetic
1124 disruptions of O/E2 and O/E3 genes reveal involvement in olfactory receptor neuron
1125 projection. *Development* **131**, (6)1377–1388 (2004).

1126 77. Fulp, C. T., Cho, G., Marsh, E. D., Nasrallah, I. M., Labosky, P. A. & Golden, J. A.
1127 Identification of Arx transcriptional targets in the developing basal forebrain. *Human*
1128 *Molecular Genetics* **17**, (23)3740–3760 (2008).

1129 78. Gécz, J., Cloosterman, D. & Partington, M. ARX: a gene for all seasons. *Current Opinion in*
1130 *Genetics & Development* **16**, (3)308–316 (2006).

1131 79. Dubois, L. & Vincent, A. The COE--Collier/Olf1/EBF--transcription factors: structural
1132 conservation and diversity of developmental functions. *Mechanisms of development* **108**,
1133 (1–2)3–12 (2001).

1134 80. Cook, R. K., Christensen, S. J., *et al.* The generation of chromosomal deletions to provide
1135 extensive coverage and subdivision of the *Drosophila melanogaster* genome. *Genome*
1136 *Biology* **13**, (3)R21 (2012).

1137 81. Chao, H.-T., Davids, M., *et al.* A Syndromic Neurodevelopmental Disorder Caused by De
1138 Novo Variants in EBF3. *The American Journal of Human Genetics* **100**, (1)128–137 (2017).

1139 82. Sleven, H., Welsh, S. J., *et al.* De Novo Mutations in EBF3 Cause a Neurodevelopmental
1140 Syndrome. *The American Journal of Human Genetics* **100**, (1)138–150 (2017).

1141 83. Harms, F. L., Girisha, K. M., *et al.* Mutations in EBF3 Disturb Transcriptional Profiles and
1142 Cause Intellectual Disability, Ataxia, and Facial Dysmorphism. *The American Journal of*
1143 *Human Genetics* **100**, (1)117–127 (2017).

1144 84. Tanaka, A. J., Cho, M. T., *et al.* De novo variants in *EBF3* are associated with hypotonia,
1145 developmental delay, intellectual disability, and autism. *Molecular Case Studies* **3**,
1146 (6)a002097 (2017).

1147 85. Blackburn, P. R., Barnett, S. S., *et al.* Novel de novo variant in *EBF3* is likely to impact DNA
1148 binding in a patient with a neurodevelopmental disorder and expanded phenotypes:
1149 patient report, in silico functional assessment, and review of published cases. *Molecular*
1150 *Case Studies* **3**, (3)a001743 (2017).

1151 86. Lopes, F., Soares, G., Gonçalves-Rocha, M., Pinto-Basto, J. & Maciel, P. Whole Gene
1152 Deletion of EBF3 Supporting Haploinsufficiency of This Gene as a Mechanism of
1153 Neurodevelopmental Disease. *Frontiers in Genetics* **8**, 143 (2017).

1154 87. Liu, N., Schoch, K., *et al.* Functional variants in TBX2 are associated with a syndromic
1155 cardiovascular and skeletal developmental disorder. *Human molecular genetics* **27**,

1156 (14)2454–2465 (2018).

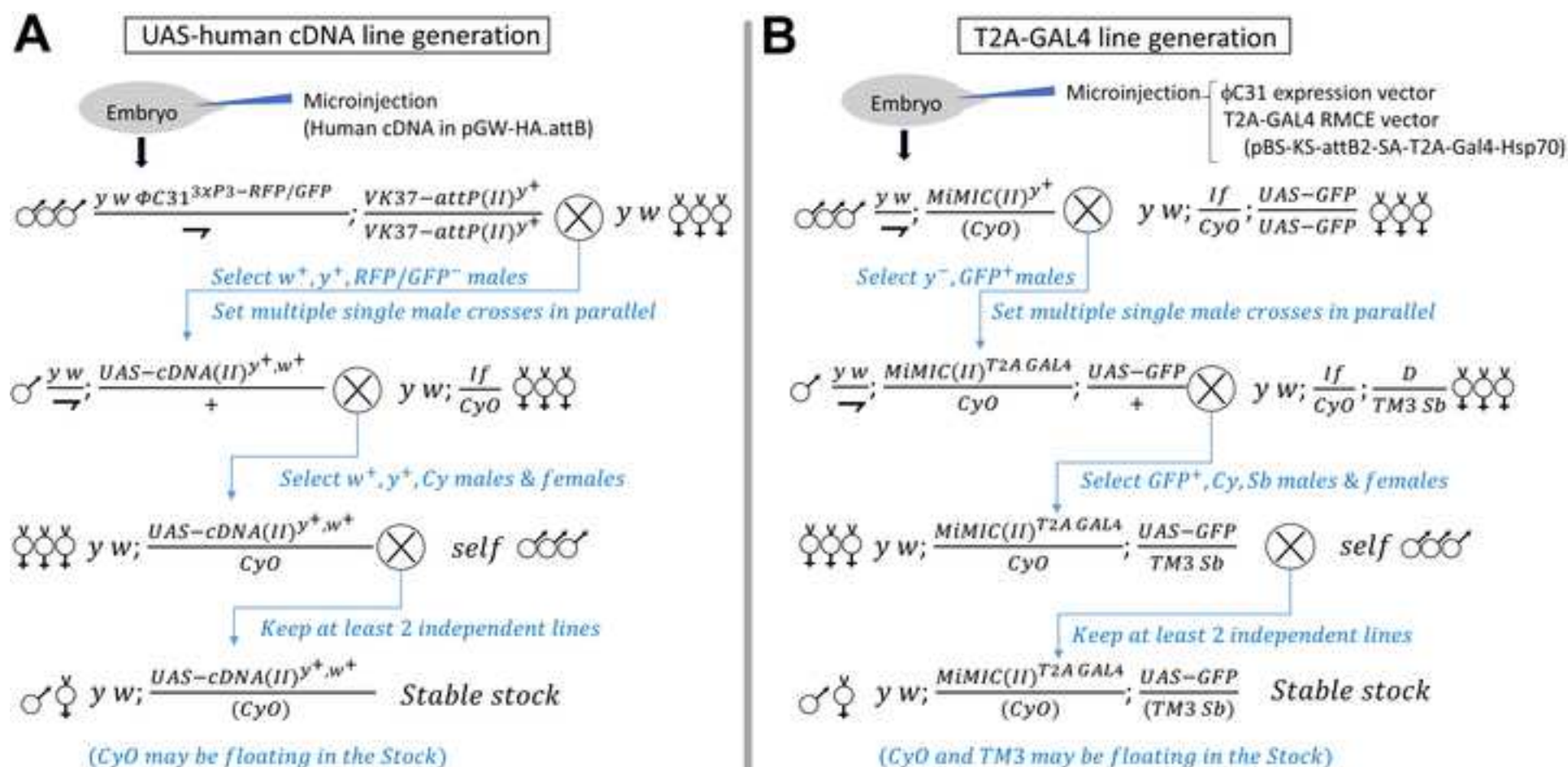
1157 88. Mesbah, K., Rana, M. S., *et al.* Identification of a Tbx1/Tbx2/Tbx3 genetic pathway
 1158 governing pharyngeal and arterial pole morphogenesis. *Human Molecular Genetics* **21**,
 1159 (6)1217–1229 (2012).

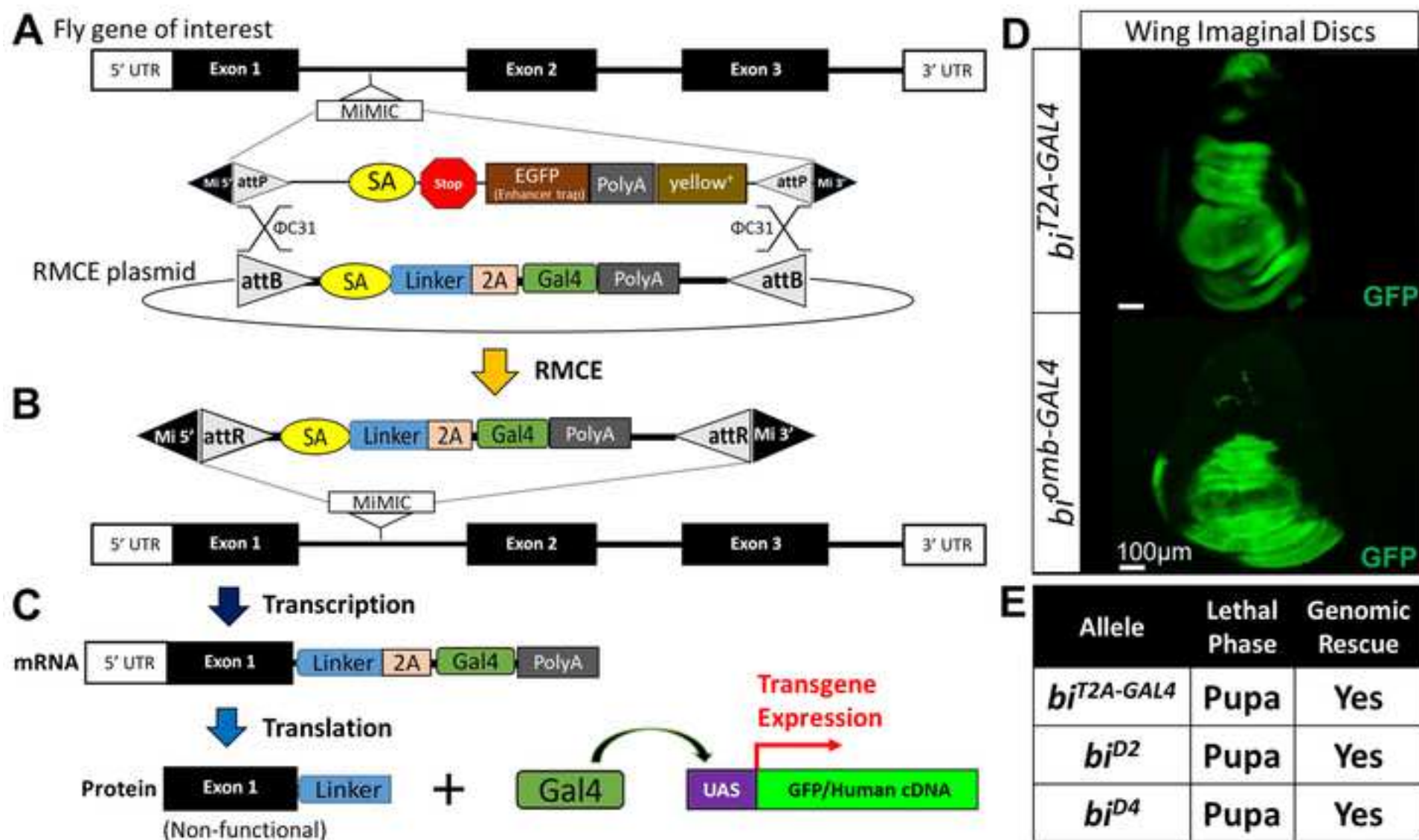
1160 89. Bellen, H. J. & Yamamoto, S. Morgan’s legacy: fruit flies and the functional annotation of
 1161 conserved genes. *Cell* **163**, (1)12–4 (2015).

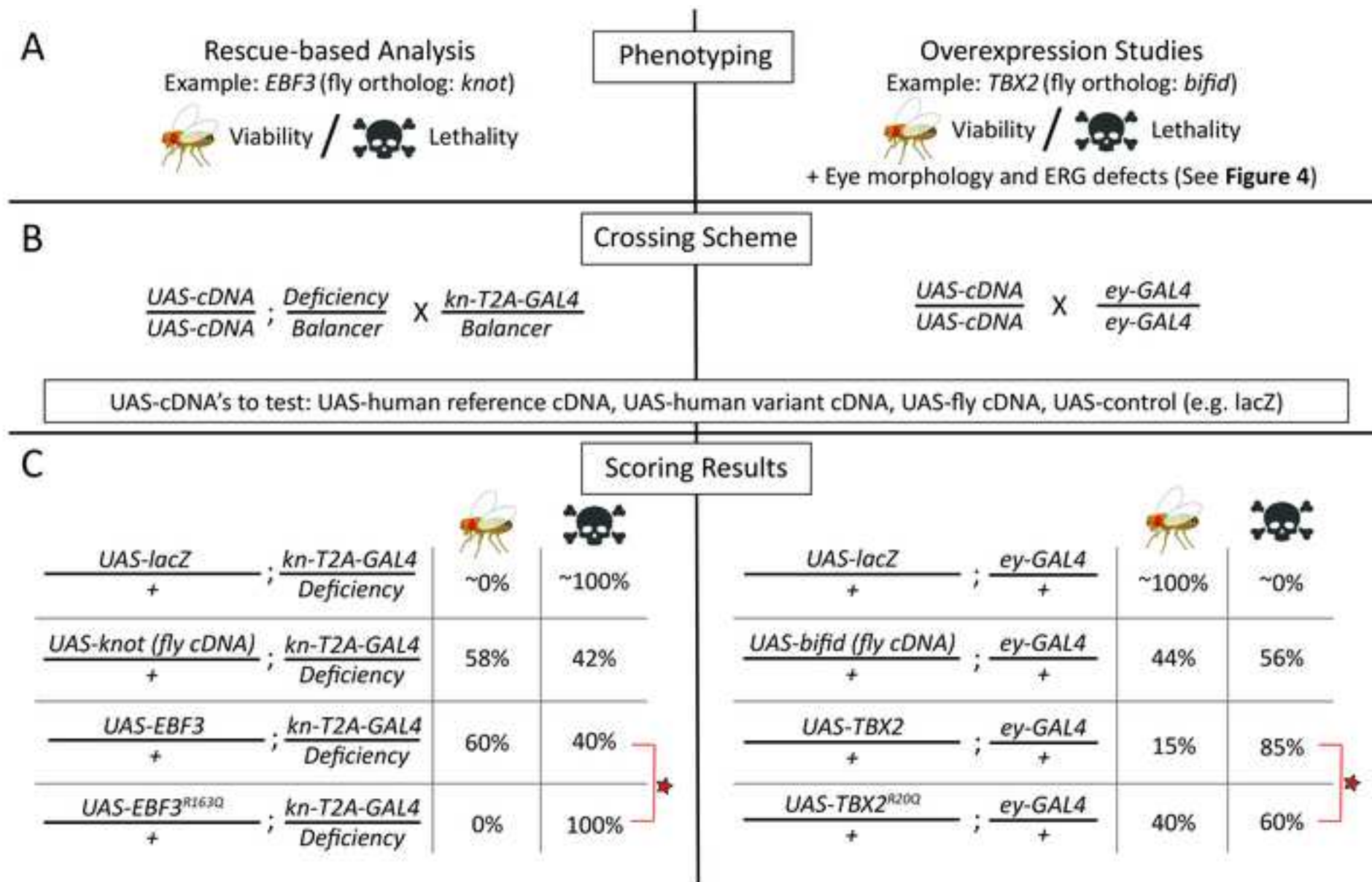
1162 90. Richards, S., Aziz, N., *et al.* Standards and guidelines for the interpretation of sequence
 1163 variants: a joint consensus recommendation of the American College of Medical Genetics
 1164 and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* **17**,
 1165 (5)405–423 (2015).

1166 91. McGary, K. L., Park, T. J., Woods, J. O., Cha, H. J., Wallingford, J. B. & Marcotte, E. M.
 1167 Systematic discovery of nonobvious human disease models through orthologous
 1168 phenotypes. *Proceedings of the National Academy of Sciences of the United States of*
 1169 *America* **107**, (14)6544–9 (2010).

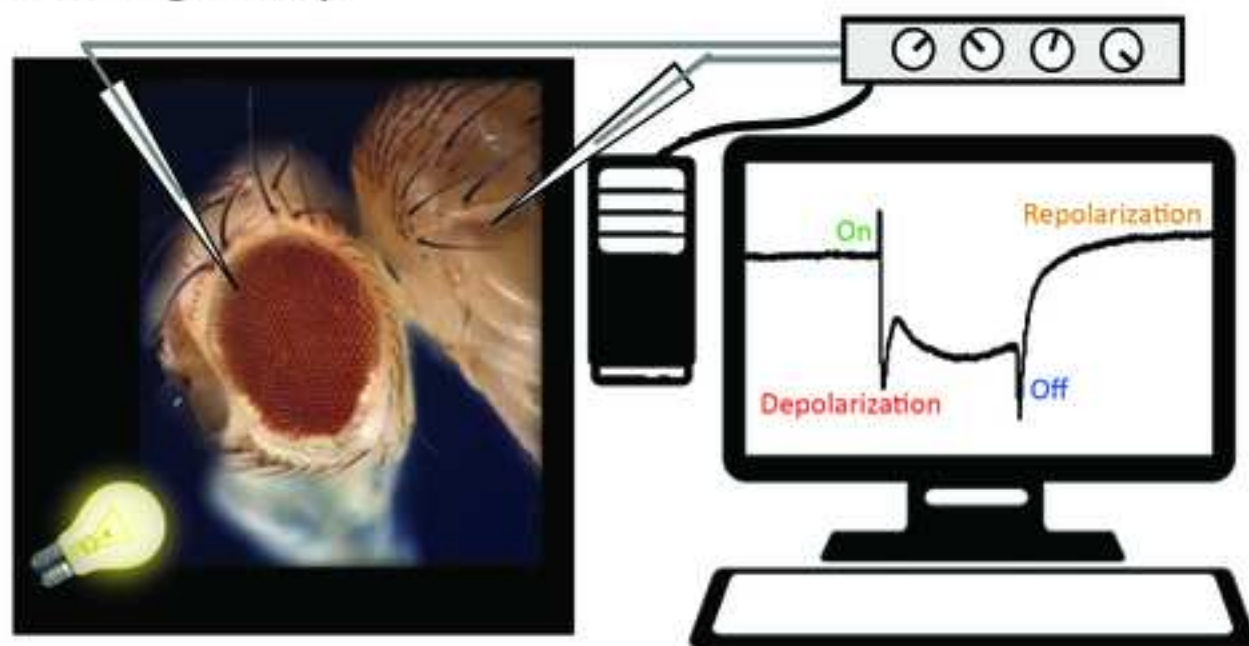
1170







A ERG Rig Setup



B Eye Developmental and ERG Phenotypes

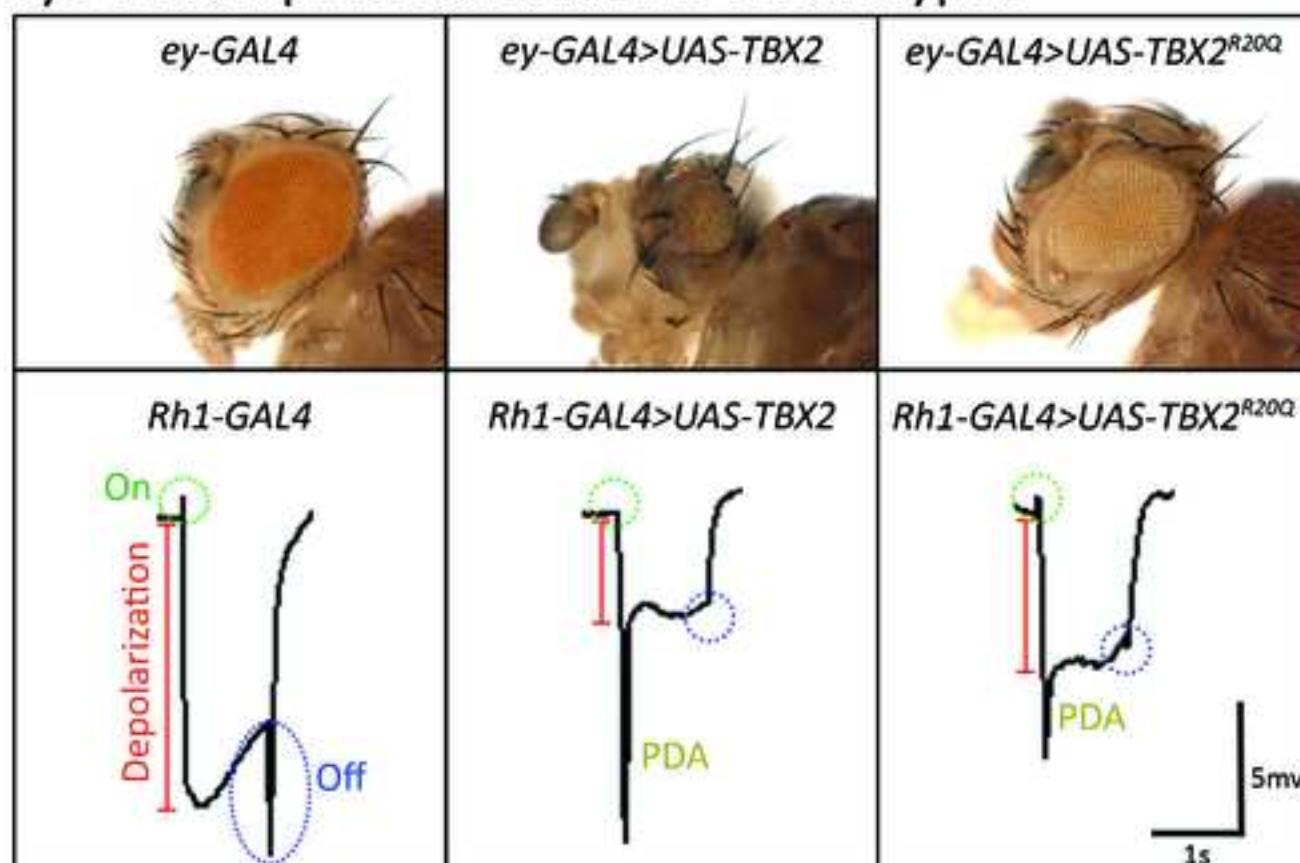


Table of Materials:

<i>Drosophila</i> Stocks for UAS-human cDNA transgenesis	
VK33 (3 rd chromosome) injection line	BDSC
VK37 (2 nd chromosome injection) line	BDSC
Plasmid DNA	
pDONR221	Thermo Fisher
pGW-HA. attB	Gift from Drs. Johannes Bischof and Konrad Basler (Bischof et al., 2013 PNAS)
Molecular biology kits and reagents	
Q5 Polymerase kit	NEB
BP Clonase kit	Thermo Fisher
LR Clonase II Enzyme kit	Thermo Fisher
PureLink Gel Extraction Kit	Thermo Fisher
Quick Change II Mutagenesis kit	Agilent
Agarose (molecular biology grade)	Sigma-Aldrich
QIAprep Spin Miniprep Kit	Qiagen
DH5α	Thermo Fisher
Electroretinogram Rig related equipment	
ISO-DAM Isolated Biologic Amplifier	LabX
Square Pulse Stimulator	Astro-Med
Axon pCLAMP 10 Data Software Package	Molecular Devices

#24871
#24872
#12536-017
#M0491
#11789020
#11791100
#K210012
#200523
#A2790
#27104
#18265017
#R150358
#S48
N/A



1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

ARTICLE AND VIDEO LICENSE AGREEMENT

Title of Article:

In vivo functional study of disease-associated human variants using Prosephila

Author(s):

J. Michael Hannish, Samantha L. Deal, VDN, Hsiao-Tuan Chao, Michael F. Wanger, Shinya Kumamoto

Item 1 (check one box): The Author elects to have the Materials be made available (as described at

<http://www.jove.com/author>) via: ☒ Standard Access ☐ Open Access

Item 2 (check one box):



The Author is NOT a United States government employee.



The Author is a United States government employee and the Materials were prepared in the course of his or her duties as a United States government employee.



The Author is a United States government employee but the Materials were NOT prepared in the course of his or her duties as a United States government employee.

ARTICLE AND VIDEO LICENSE AGREEMENT

1. **Defined Terms.** As used in this Article and Video License Agreement, the following terms shall have the following meanings: **"Agreement"** means this Article and Video License Agreement; **"Article"** means the article specified on the last page of this Agreement, including any associated materials such as texts, figures, tables, artwork, abstracts, or summaries contained therein; **"Author"** means the author who is a signatory to this Agreement; **"Collective Work"** means a work, such as a periodical issue, anthology or encyclopedia, in which the Materials in their entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole; **"CRC License"** means the Creative Commons Attribution-Non Commercial-No Derivs 3.0 Unported Agreement, the terms and conditions of which can be found at: <http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode>; **"Derivative Work"** means a work based upon the Materials or upon the Materials and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Materials may be recast, transformed, or adapted; **"Institution"** means the institution, listed on the last page of this Agreement, by which the Author was employed at the time of the creation of the Materials; **"JoVE"** means MyJoVE Corporation, a Massachusetts corporation and the publisher of *The Journal of Visualized Experiments*; **"Materials"** means the Article and / or the Video; **"Parties"** means the Author and JoVE; **"Video"** means any video(s) made by the Author, alone or in conjunction with any other parties, or by JoVE or its affiliates or agents, individually or in collaboration with the Author or any other parties, incorporating all or any portion of the Article, and in which the Author may or may not appear.

2. **Background.** The Author, who is the author of the Article, in order to ensure the dissemination and protection of the Article, desires to have the JoVE publish the Article and create and transmit videos based on the Article. In furtherance of such goals, the Parties desire to memorialize in this Agreement the respective rights of each Party in and to the Article and the Video.

3. **Grant of Rights in Article.** In consideration of JoVE agreeing to publish the Article, the Author hereby grants to JoVE, subject to **Sections 4 and 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Article in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Article into other languages, create adaptations, summaries or extracts of the Article or other Derivative Works (including, without limitation, the Video) or Collective Works based on all or any portion of the Article and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. If the "Open Access" box has been checked in **Item 1** above, JoVE and the Author hereby grant to the public all such rights in the Article as provided in, but subject to all limitations and requirements set forth in, the CRC License.

ARTICLE AND VIDEO LICENSE AGREEMENT

4. **Retention of Rights in Article.** Notwithstanding the exclusive license granted to JoVE in **Section 3** above, the Author shall, with respect to the Article, retain the non-exclusive right to use all or part of the Article for the non-commercial purpose of giving lectures, presentations or teaching classes, and to post a copy of the Article on the Institution's website or the Author's personal website, in each case provided that a link to the Article on the JoVE website is provided and notice of JoVE's copyright in the Article is included. All non-copyright intellectual property rights in and to the Article, such as patent rights, shall remain with the Author.

5. **Grant of Rights in Video – Standard Access.** This **Section 5** applies if the "Standard Access" box has been checked in **Item 1** above or if no box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby acknowledges and agrees that, Subject to **Section 7** below, JoVE is and shall be the sole and exclusive owner of all rights of any nature, including, without limitation, all copyrights, in and to the Video. To the extent that, by law, the Author is deemed, now or at any time in the future, to have any rights of any nature in or to the Video, the Author hereby disclaims all such rights and transfers all such rights to JoVE.

6. **Grant of Rights in Video – Open Access.** This **Section 6** applies only if the "Open Access" box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby grants to JoVE, subject to **Section 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Video in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Video into other languages, create adaptations, summaries or extracts of the Video or other Derivative Works or Collective Works based on all or any portion of the Video and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. For any Video to which this Section 6 is applicable, JoVE and the Author hereby grant to the public all such rights in the Video as provided in, but subject to all limitations and requirements set forth in, the CRC License.

7. **Government Employees.** If the Author is a United States government employee and the Article was prepared in the course of his or her duties as a United States government employee, as indicated in **Item 2** above, and any of the licenses or grants granted by the Author hereunder exceed the scope of the 17 U.S.C. 403, then the rights granted hereunder shall be limited to the maximum rights permitted under such

statute. In such case, all provisions contained herein that are not in conflict with such statute shall remain in full force and effect, and all provisions contained herein that do so conflict shall be deemed to be amended so as to provide to JoVE the maximum rights permissible within such statute.

8. **Likeness, Privacy, Personality.** The Author hereby grants JoVE the right to use the Author's name, voice, likeness, picture, photograph, image, biography and performance in any way, commercial or otherwise, in connection with the Materials and the sale, promotion and distribution thereof. The Author hereby waives any and all rights he or she may have, relating to his or her appearance in the Video or otherwise relating to the Materials, under all applicable privacy, likeness, personality or similar laws.

9. **Author Warranties.** The Author represents and warrants that the Article is original, that it has not been published, that the copyright interest is owned by the Author (or, if more than one author is listed at the beginning of this Agreement, by such authors collectively) and has not been assigned, licensed, or otherwise transferred to any other party. The Author represents and warrants that the author(s) listed at the top of this Agreement are the only authors of the Materials. If more than one author is listed at the top of this Agreement and if any such author has not entered into a separate Article and Video License Agreement with JoVE relating to the Materials, the Author represents and warrants that the Author has been authorized by each of the other such authors to execute this Agreement on his or her behalf and to bind him or her with respect to the terms of this Agreement as if each of them had been a party hereto as an Author. The Author warrants that the use, reproduction, distribution, public or private performance or display, and/or modification of all or any portion of the Materials does not and will not violate, infringe and/or misappropriate the patent, trademark, intellectual property or other rights of any third party. The Author represents and warrants that it has and will continue to comply with all government, institutional and other regulations, including, without limitation all institutional, laboratory, hospital, ethical, human and animal treatment, privacy, and all other rules, regulations, laws, procedures or guidelines, applicable to the Materials, and that all research involving human and animal subjects has been approved by the Author's relevant institutional review board.

10. **JoVE Discretion.** If the Author requests the assistance of JoVE in producing the Video in the Author's facility, the Author shall ensure that the presence of JoVE employees, agents or independent contractors is in accordance with the relevant regulations of the Author's institution. If more than one author is listed at the beginning of this Agreement, JoVE may, in its sole discretion, elect not take any action with respect to the Article until such time as it has received complete, executed Article and Video License Agreements from each such author. JoVE reserves the right, in its absolute and sole discretion and without giving any reason therefore, to accept or decline any work submitted to JoVE. JoVE and its employees, agents and independent contractors shall have

ARTICLE AND VIDEO LICENSE AGREEMENT

full, unfettered access to the facilities of the Author or of the Author's institution as necessary to make the Video, whether actually published or not. JoVE has sole discretion as to the method of making and publishing the Materials, including, without limitation, to all decisions regarding editing, lighting, filming, timing of publication, if any, length, quality, content and the like.

11. **Indemnification.** The Author agrees to indemnify JoVE and/or its successors and assigns from and against any and all claims, costs, and expenses, including attorney's fees, arising out of any breach of any warranty or other representations contained herein. The Author further agrees to indemnify and hold harmless JoVE from and against any and all claims, costs, and expenses, including attorney's fees, resulting from the breach by the Author of any representation or warranty contained herein or from allegations or instances of violation of intellectual property rights, damage to the Author's or the Author's institution's facilities, fraud, libel, defamation, research, equipment, experiments, property damage, personal injury, violations of institutional, laboratory, hospital, ethical, human and animal treatment, privacy or other rules, regulations, laws, procedures or guidelines, liabilities and other losses or damages related in any way to the submission of work to JoVE, making of videos by JoVE, or publication in JoVE or elsewhere by JoVE. The Author shall be responsible for, and shall hold JoVE harmless from, damages caused by lack of sterilization, lack of cleanliness or by contamination due to the making of a video by JoVE its employees, agents or independent contractors. All sterilization, cleanliness or decontamination procedures shall be solely the responsibility of the Author and shall be undertaken at the Author's

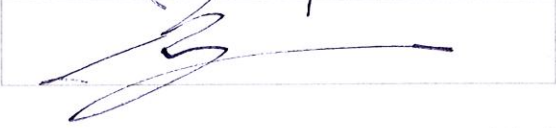
expense. All indemnifications provided herein shall include JoVE's attorney's fees and costs related to said losses or damages. Such indemnification and holding harmless shall include such losses or damages incurred by, or in connection with, acts or omissions of JoVE, its employees, agents or independent contractors.

12. **Fees.** To cover the cost incurred for publication, JoVE must receive payment before production and publication the Materials. Payment is due in 21 days of invoice. Should the Materials not be published due to an editorial or production decision, these funds will be returned to the Author. Withdrawal by the Author of any submitted Materials after final peer review approval will result in a US\$1,200 fee to cover pre-production expenses incurred by JoVE. If payment is not received by the completion of filming, production and publication of the Materials will be suspended until payment is received.

13. **Transfer, Governing Law.** This Agreement may be assigned by JoVE and shall inure to the benefits of any of JoVE's successors and assignees. This Agreement shall be governed and construed by the internal laws of the Commonwealth of Massachusetts without giving effect to any conflict of law provision thereunder. This Agreement may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to be one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

A signed copy of this document must be sent with all new submissions. Only one Agreement required per submission.

CORRESPONDING AUTHOR:

Name: SHINYA YAMAMOTO
Department: MOLECULAR & HUMAN GENETICS
Institution: BAYLOR COLLEGE OF MEDICINE
Article Title: In vivo functional study of disease-associated rare human variants using Drosophila
Signature:  Date: 1/4/2019

Please submit a signed and dated copy of this license by one of the following three methods:

- 1) Upload a scanned copy of the document as a pdf on the JoVE submission site;
- 2) Fax the document to +1.866.381.2236;
- 3) Mail the document to JoVE / Attn: JoVE Editorial / 1 Alewife Center #200 / Cambridge, MA 02139

For questions, please email submissions@jove.com or call +1.617.945.9051

Members of the Undiagnosed Diseases Network

Maria T. Acosta
David R. Adams
Aaron Aday
Mercedes E. Alejandro
Patrick Allard
Euan A. Ashley
Mahshid S. Azamian
Carlos A. Bacino
Eva Baker
Ashok Balasubramanyam
Dustin Baldrige
Gabriel F. Batzli
Alan H. Beggs
Hugo J. Bellen
Jonathan A. Bernstein
Gerard T. Berry
Anna Bican
David P. Bick
Camille L. Birch
Carsten Bonnenmann
Devon Bonner
Braden E. Boone
Bret L. Bostwick
Lauren C. Briere
Elly Brokamp
Donna M. Brown
Matthew Brush
Elizabeth A. Burke
Lindsay C. Burrage
Manish J. Butte
Hsiao-Tuan Chao
Gary D. Clark
Terra R. Coakley
Joy D. Cogan
F. Sessions Cole
Heather A. Colley
Cynthia M. Cooper
Heidi Cope
William J. Craigen
Precilla D'Souza
Mariska Davids
Jean M. Davidson
Jyoti G. Dayal
Esteban C. Dell'Angelica
Shweta U. Dhar
Laurel A. Donnell-Fink
Naghmeh Dorrani
Daniel C. Dorset
Emilie D. Douine

David D. Draper
Annika M. Dries
Laura Duncan
David J. Eckstein
Lisa T. Emrick
Christine M. Eng
Gregory M. Enns
Cecilia Esteves
Tyra Estwick
Liliana Fernandez
Carlos Ferreira
Elizabeth L. Fieg
Paul G. Fisher
Brent L. Fogel
Noah D. Friedman
William A. Gahl
Rena A. Godfrey
Alica M. Goldman
David B. Goldstein
Jean-Philippe F. Gourdine
Catherine A. Groden
Andrea L. Gropman
Melissa Haendel
Rizwan Hamid
Neil A. Hanchard
Frances High
Ingrid A. Holm
Jason Hom
Alden Huang
Yong Huang
Fariha Jamal
Yong-hui Jiang
Jean M. Johnston
Angela L. Jones
Lefkothea Karaviti
Emily G. Kelley
David M. Koeller
Isaac S. Kohane
Jennefer N. Kohler
Donna M. Krasnewich
Susan Korrick
Mary Koziura
Joel B. Krier
Jennifer E. Kyle
Seema R. Lalani
C. Christopher Lau
Jozef Lazar
Kimberly LeBlanc
Brendan H. Lee
Hane Lee
Shawn E. Levy

Richard A. Lewis
Sharyn A. Lincoln
Pengfei Liu
Sandra K. Loo
Joseph Loscalzo
Richard L. Maas
Ellen F. Macnamara
Calum A. MacRae
Valerie V. Maduro
Marta M. Majcherska
May Christine V. Malicdan
Laura A. Mamounas
Teri A. Manolio
Thomas C. Markello
Ronit Marom
Martin G. Martin
Julian A. Martínez-Agosto
Shruti Marwaha
Thomas May
Allyn McConkie-Rosell
Colleen E. McCormack
Alexa T. McCray
Jason D. Merker
Thomas O. Metz
Matthew Might
Paolo M. Moretti
Marie Morimoto
John J. Mulvihill
David R. Murdock
Avi Nath
Stan F. Nelson
J. Scott Newberry
John H. Newman
Sarah K. Nicholas
Donna Novacic
James P. Orengo
Stephen Pak
J. Carl Pallais
Christina GS. Palmer
Jeanette C. Papp
Neil H. Parker
Loren DM. Pena
John A. Phillips III
Jennifer E. Posey
John H. Postlethwait
Lorraine Potocki
Barbara N. Pusey
Genecee Renteria
Chloe M. Reuter
Lynette Rives
Amy K. Robertson

Lance H. Rodan
Jill A. Rosenfeld
Robb K. Rowley
Jacinda B. Sampson
Susan L. Samson
Timothy Schedl
Kelly Schoch
Daryl A. Scott
Lisa Shakachite
Prashant Sharma
Vandana Shashi
Kathleen Shields
Jimann Shin
Rebecca Signer
Catherine H. Sillari
Edwin K. Silverman
Janet S. Sinsheimer
Kevin S. Smith
Lilianna Solnica-Krezel
Rebecca C. Spillmann
Joan M. Stoler
Nicholas Stong
Jennifer A. Sullivan
David A. Sweetser
Cecelia P. Tamburro
Queenie K.-G. Tan
Cynthia J. Tift
Camilo Toro
Alyssa A. Tran
Tiina K. Urv
Tiphonie P. Vogel
Daryl M. Waggott
Colleen E. Wahl
Nicole M. Walley
Chris A. Walsh
Melissa Walker
Jennifer Wambach
Jijun Wan
Lee-kai Wang
Michael F. Wangler
Patricia A. Ward
Katrina M. Waters
Bobbie-Jo M. Webb-Robertson
Daniel Wegner
Monte Westerfield
Matthew T. Wheeler
Anastasia L. Wise
Lynne A. Wolfe
Jeremy D. Woods
Elizabeth A. Worthey
Shinya Yamamoto

John Yang
Amanda J. Yoon
Guoyun Yu
Diane B. Zastrow
Chunli Zhao