OrthoList 2: a new comparative genomic analysis of human and *C. elegans* genes

Woojin Kim[*], Ryan S. Underwood[†, 1], Iva Greenwald[†, ‡, **] and Daniel D. Shaye[§, **]

[*]Dept. of Data Science, Columbia University. New York, NY. USA

[†]Dept. of Biochemistry and Molecular Biophysics, Columbia University. New York, NY. USA

[‡]Dept. of Biological Sciences, Columbia University. New York, NY. USA

[§]Dept. of Physiology and Biophysics, University of Illinois-Chicago. Chicago, IL. USA.

Running Title: A new compendium of human-worm orthologs

Keywords: genome, homology, *C. elegans*, human

Corresponding Author:

Daniel D. Shaye.

University of Illinois at Chicago, College of Medicine.

Dept. of Physiology and Biophysics, MC/901.

835 S. Wolcott Ave., Room E202

Chicago, IL 60612

Tel: +1-312-413-0492

email: shaye@uic.edu

** co-senior authors

[1] Present address: Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093 USA

ABSTRACT

OrthoList, a compendium of *C. elegans* genes with human orthologs compiled in 2011 by a meta-analysis of four orthology-prediction methods, has been a popular tool for identifying conserved genes for research into biological and disease mechanisms. However, the efficacy of orthology prediction depends on the accuracy of gene model predictions, an ongoing process, and orthology-prediction algorithms have also been updated over time. Here we present OrthoList 2 (OL2), a new comparative genomic analysis between *C. elegans* and humans, and the first assessment of how changes over time affect the landscape of predicted orthologs between two species. Although we find that updates to the orthology-prediction methods significantly changed the landscape of *C. elegans*-human orthologs predicted by individual programs, and, unexpectedly, reduced agreement amongst them, we also show that our meta-analysis approach “buffered” against changes in gene content. We show that adding results from more programs did not lead to many additions to the list, and discuss reasons to avoid assigning "scores" based on support by individual orthology prediction programs, the treatment of "legacy" genes no longer predicted by these programs, and the practical difficulties of updating due to encountering deprecated, changed, or retired gene IDs. In addition, we consider what other criteria may support claims of orthology, and alternative approaches to find potential orthologs that elude identification by these programs. Finally, we created a new web-based tool that allows for rapid searches of OL2 by gene identifiers, protein domains (InterPro and SMART), or human-disease associations (OMIM), and also includes available RNAi resources to facilitate potential translational cross-species studies.

INTRODUCTION

Studies in *C. elegans* have illuminated many mechanisms relevant to human biology and disease. Forward genetic screens based on phenotype have identified genes homologous to human disease-associated genes, illuminating fundamental properties about their roles and mechanisms of action (e.g., Greenwald 2012; Sundaram 2013; Golden 2017; van der Bliek *et al.* 2017). Reverse genetic methods have expanded the repertoire of possible genetic approaches. These methods include the ability to phenocopy loss-of-function mutations by feeding worms bacteria expressing double-stranded RNA (Fire *et al.* 1998; Timmons and Fire 1998). The efficiency of RNAi in *C. elegans* has allowed for genome-wide screens (Fraser *et al.* 2000; Kamath *et al.* 2003; O'Reilly *et al.* 2016), or screens targeted to specific conserved genes, such as human disease genes (e.g., Sin *et al.* 2014; Vahdati Nia *et al.* 2017; Nordquist *et al.* 2018) or those involved in fundamental biological processes (e.g., Balklava *et al.* 2007; Dunn *et al.* 2010; Firnhaber and Hammarlund 2013; Allen *et al.* 2014; Du *et al.* 2015). Other efficient reverse genetic methods in *C. elegans* include the large-scale generation of deletion and point-mutations for functional genetic analysis (Moerman and Barstead 2008; Thompson *et al.* 2013), transgenesis to engineer models for gain-of-function mutations associated with disease (Markaki and Tavernarakis 2010; Tucci *et al.* 2011), and now CRISPR/Cas9-based genome engineering for manipulation of endogenous genes (Dickinson and Goldstein 2016).

To facilitate cross-platform studies, we created OrthoList, a compendium of *C. elegans* genes with human orthologs that was originally published in the form of an Excel spreadsheet (Shaye and Greenwald 2011). Subsequently, we created a minimal, unpublished, online tool distributed through informal *C. elegans* community channels to enhance its accessibility and utility. OrthoList has indeed both facilitated the identification of orthology (e.g., Firnhaber and

Hammarlund 2013; Du *et al.* 2015; Vahdati Nia *et al.* 2017) and has been used as the basis for streamlining RNAi screens (e.g., Gillard *et al.* 2015; Hernando-Rodriguez *et al.* 2018; Nordquist *et al.* 2018).

To generate OrthoList, we used a meta-analysis strategy in which we compiled the results of different orthology prediction programs. Because each sequence analysis method at the base of these programs has its strengths and weaknesses, each leading to a different trade-off between precision (true vs. false positive rate) and recall (true vs. false negative rate), we expected a meta-analysis to capture the greatest number of potential orthologs, with high precision and recall. Our expectation was subsequently supported by independent assessments (Pryszcz *et al.* 2011; Pereira *et al.* 2014), and now, by our new results below.

Genome annotation in both *C. elegans* and humans is an ongoing process, and the efficacy of genome-wide orthology prediction approaches depends on the accuracy of the gene models in the genomes under scrutiny. Thus, we have now performed a new meta-analysis using current information to generate OrthoList 2, an up-to-date compendium of genes with *C. elegans* and human orthologs. In addition, we have created an improved online tool associated with OrthoList 2 (found at ortholist.shaye-lab.org) with features that facilitate genetic analysis in *C. elegans* by containing links to the complete “feeding RNAi” clone set (Fraser *et al.* 2000; Kamath *et al.* 2003) as well as multiple data input options, links to other databases (SMART and InterPro for protein domains (Finn *et al.* 2017; Letunic and Bork 2018), OMIM for disease associations (McKusick 2007), and more flexibility in accessing results. We analyze the changes in content between OrthoList 1 and 2, and demonstrate the robustness of the meta-analysis strategy, including examples of the strengths and limitations of this approach that have emerged during this update. Our analysis highlights the importance of

assessing orthology by meta-analysis, rather than relying on a single “snapshot” in time or a single program to obtain a comprehensive list of genes conserved between *C. elegans* and humans.

MATERIALS AND METHODS

A detailed description, and accompanying source code, of how we obtained and compiled the data underlying OrthoList can be found at https://github.com/danshaye/OrthoList2, and a freeze of the underlying code is provided as File S8. Briefly, for all methods, except Ensembl Compara, we downloaded and analyzed results from the most current release available. For details on the source data underlying each of the orthology prediction methods queried see Supplemental Table S1. For Compara, which is updated every 2-3 months, we noticed a great deal of fluidity in results (see Fig. S1) within the 3 versions that were released as we compiled OrthoList 2 (Ensembl Compara v87, v88 and v89), so that only ~85% of the worm-human orthologs predicted were common between the three versions. For example, the update from v87 to v88 led to a loss of 294 worm genes, of which about half (158) were re-added upon update to v89 (Fig. S1). Similarly, the update from v88 to v89 led to a loss of 320 genes, of which 178 had been supported in both v87 and v88. Given these differences, and in order to ensure the most comprehensive results from Ensembl Compara, we decided to keep all genes found by the three versions released as we compiled and analyzed OrthoList 2.

Data comparisons and Venn diagrams were done with the web-based program Venny, found at http://bioinfogp.cnb.csic.es/tools/venny/index.html (Oliveros 2007-2015). Statistical analyses were conducted using resources from the Handbook of Biological Statistics

(McDonald 2014), found at http://www.biostathandbook.com, and with GraphPad Prism Software, v6.0.

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, tables, supplementary materials uploaded to figshare, and at the online repository located at https://github.com/danshaye/OrthoList2.

RESULTS

Addressing changes to gene predictions in the *C. elegans* genome

Each time a genome sequence database is updated, there are changes in gene predictions. Some of the changes are “correct” and will endure, while others may fluctuate as prediction algorithms continue to evolve and more sequencing data becomes available. We previously hypothesized that such changes to gene predictions would have a minor effect on the integrity of OrthoList (hereafter OL1), because conserved genes would be the most likely to be accurately represented in genome releases (Shaye and Greenwald 2011). The analysis in this and the next section supports this hypothesis, as only ~0.9% of *C. elegans* and ~0.1% of human genes in OL1 were removed, or "deprecated", due to updated gene predictions.

We analyzed alterations in *C. elegans* gene predictions by cross-checking the 7,663 genes in OL1, which was built using WormBase version WS210 (released in 2009), to WormBase WS257 (released in 2017). We found that only 151 worm genes changed due to updated predictions. Most (67/151) resulted from their re-classification as pseudogenes, ncRNA, being transposon-derived, or killed due to lack of evidence (Type I change, File S1). It is only this type of change, representing ~0.9% (67 of 7,663) of worm genes in OL1, that results in deprecation of a *C. elegans* gene previously believed to be conserved in humans.

A second type of change, seen with 43 worm genes, resulted from combining, or "merging", two or more genes that had each, separately, been found to have a human ortholog. This type of change (Type II, File S1) led to a net loss of 22 genes. Together, Type I and Type II changes led to a removal of 88 worm genes from OL1. Our analysis below addressing updates to human gene predictions led to removal of an additional 6 worm genes from OL1, leading to an updated final number of 7,569 worm genes predicted to have human orthologs in OL1 (Fig. 1 and File S2).

The final 41 worm genes that changed since OL1 were assigned new IDs, either because experimental evidence suggested that they should be merged to genes that were previously not in OL1 (16/41), or due to addition of previously unpredicted gene segments (25/41) leading to a new ID (Type III, File S1). This last type of change does not affect the total number of *C. elegans* genes in OL1.

Addressing changes to gene predictions in the human genome

One of the major challenges we encountered in our analysis was accommodating changes to the human gene annotation. We compiled OL1 using the Ensembl genome browser (Vilella *et al.* 2009) to obtain human genes, and their associated ENSG IDs, because Ensembl provides strong support for comparative genomic studies via its BioMart tool for large-scale data mining and analysis (Kasprzyk 2011). Based on Ensembl data, OL1 appeared to include 11,416 predicted human genes (ENSG IDs from Ensembl v57, 2010; File S3, tab A). However, we noticed that in some cases a single gene had multiple ENSG IDs associated with it (e.g., the gene *NOTCH4* has 7 associated IDs). These alternative IDs occur when new sequence differs from the primary assembly, due to new allelic sequences (haplotypes and

novel patches) or fix patches. Novel patches represent new allelic loci, but not necessarily haplotypes. Fix patches occur when the primary assembly was found to be incorrect, and the patch reflects the corrected sequence (for details see the Genome Reference Consortium page at https://www.ncbi.nlm.nih.gov/grc). Regardless of source, the fact that some genes have multiple IDs prevents us from making an accurate assessment of how many human genes were in OL1. Henceforth when discussing human genes, we use the number of ENSG IDs as an approximation for the number of human genes in OL, and consider the gene estimate further in the section describing the gene content of OrthoList 2.

To begin addressing changes to gene predictions in the human genome, we cross-checked the 11,416 ENSG IDs from OL1 with a recent release of Ensembl (v89, 2017) and found that 574 IDs appeared to be lost (File S3, tab A). Although this is a small fraction of the IDs in OL1 (~5%), this number seemed high in light of our hypothesis that conserved genes should be stable. Unfortunately, Ensembl does not provide details of ENSG ID curation. Instead they make available a “version history” that describes changes and indicates when an ID was “retired” (File S3, tab B). However, as discussed below, our manual curation suggests that most of the ENSG ID marked as retired represent genes that still exist in the human genome assembly with a different ID.

To ask whether the 574 retired ENSG IDs represented genes that were truly deprecated, we undertook a cross-species comparison. We extracted the 624 worm orthologs of these apparently-deprecated human genes from OL1. Based on our analysis discussed above, six of these worm genes had changed: two were themselves deprecated (Type I change, File S1), so it is likely that the two human genes that matched to these were themselves also truly deprecated (File S3, tab D). The remaining four worm genes were

updated (Type II and Type III changes, File S1), and these are considered further with respect to their relationship to apparently-deprecated human genes.

Of the 622 current worm genes that matched apparently-deprecated human genes almost all (616/622 or ~99%) continue to have human orthologs with current ENSG IDs. Manual inspection of a randomly selected subset (n=20) of these human-worm pairs showed that, in all cases, the underlying human gene that appeared to be deprecated because its ENSG ID had been "retired" actually has another current ENSG ID assigned to it, and in almost all cases (19 of 20 in the sampled set) the current ENSG ID is not linked to the retired one (File S3, tab C. POLDIP2, the only gene within this set for which its retired ENSG ID is linked to its current one, is shown in bold). Therefore, it appears that in most, if not all, cases where a worm gene matched an apparently-deprecated human gene in OL1, the human gene actually still exists with a new ID that is not linked to the retired one. An alternative, not mutually-exclusive, possibility is that worm genes that matched apparently deprecated human genes remain matched to one, or more, paralogs of the original human gene. However, since Ensembl does not make available a detailed history of ID changes, we are unable to address this possibility. Regardless, based on the continued extensive orthology between *C. elegans* genes and erroneously-deprecated human genes, we are only able to confirm deprecation of 16 ENSG IDs from OL1 (see below).

The last 6 of 622, worm genes that matched apparently-deprecated human genes had as sole orthologs 14 human genes that appear to be truly lost, as these worm genes do not match any current ENSG ID (File S3, tab D). Moreover, these 6 worm genes do not pick up any human sequences even by simple BLAST searches (File S3, tab D). Therefore, these 6 worm genes no longer have human orthologs and were thus removed from OL1 (resulting in

the final number of 7,569 worm genes in OL1. Fig. 1 and File S2), and their 14 cognate human genes are truly deprecated. If we add the 2 human ENSG IDs that matched deprecated *C. elegans* genes (see above) to the 14 ENSG IDs discussed here, the total number of confirmed deprecated IDs is 16, or just ~0.1% of the ENSG IDs in OL1 (File S3, tab D and Fig. 1).

This analysis supports our hypothesis that conserved genes are stable, and demonstrates there are some difficulties with human gene annotations that need to be taken into account when performing genome-wide homology analyses. Given these deficiencies in annotation, we are unable to reliably address the changes in gene content of the human portion of OrthoList. Therefore, to avoid confounding effects that arise from differences in the quality of genome annotation, hereafter our analysis will focus on the *C. elegans* content of OrthoList.

Updates to the individual orthology-prediction methods used in OL1 change the landscape of *C. elegans*-human orthologs in the absence of meta-analysis

Orthology prediction methods can be classified into three general categories: graph-based, tree-based or hybrid strategies, although recent analysis suggest there is no obvious systematic difference in performance between these strategies per se, even while there are differences in performance of individual programs (Altenhoff *et al.* 2016; Sutphin *et al.* 2016). Graph-based programs begin with pairwise alignments between all protein sequences from two species to identify the most-likely orthologous pair, followed by different clustering criteria. Tree-based strategies take advantage of the evolutionary relationships between species, simultaneously aligning sequences from multiple species to build phylogenetic trees for each protein. Hybrid strategies combine aspects of both graph and tree-based approaches, applying

graph-based clustering methods at the nodes of phylogenetic trees to generate ortholog predictions. To generate OL1 we combined data from four programs: 1) InParanoid (Remm *et al.* 2001), a graph-based approach that clusters orthologs between two species, and defines paralogs, based on reciprocal-best BLAST hit (RBH) scores, 2) OrthoMCL (Li *et al.* 2003), a graph-based approach that generates a similarity matrix using RBH scores within species, and between species, followed by Markov clustering, to produce inter-species ortholog groups, 3) Ensembl Compara (Vilella *et al.* 2009), a tree-based approach, and 4) HomoloGene (Wheeler *et al.* 2007), a hybrid approach.

We wanted to assess the effects that updates to the orthology-prediction methods used to generate OL1 would have on the landscape of worm-human orthologs. The previously-used programs have been updated with varying regularity since OL1 was compiled (Table 1): InParanoid and OrthoMCL have been updated once, HomoloGene has been updated four times (the latest version, which we use here, released in 2014), and Ensembl Compara is updated every 2-3 months. As discussed in "Methods", here we use combined data from three recent Ensembl releases (v87, 88 and 89, Dec 2016-May 2017).

As shown in Tables 1 and 2, at first glance the net number of worm genes with human orthologs predicted by each program did not appear to change greatly between versions of the orthology-prediction methods: the mean change in worm genes with predicted human orthologs was -0.5% (±3.0% s.e.m). However, closer examination showed that the change in gene content, i.e., the actual genes in the results, is larger than reflected by the change in net numbers (Fig. 2 and Table 2).

The average decrease in *C. elegans* genes with predicted human orthologs resulting from updates to orthology-prediction methods was 7.9% (±3.6% s.e.m. Table 2),

corresponding to a predicted loss of 598 (±272) worm genes from OL1. As discussed above, updates in gene predictions resulted in a loss of only 95 worm genes from OL1; therefore, it appears that updates to orthology-prediction methods causes ~6x more losses, suggesting that changes in orthology-prediction algorithms over time have a greater effect on the landscape of worm-human orthologs than do changes in underlying gene models. Updates also appear to increase sensitivity, because there was an average increase of 7.4% (±2.6% s.e.m.) *C. elegans* genes with predicted human ortholog (Table 2), corresponding to a predicted gain of 568 (±197) worm genes.

Taken together, our analysis in this section suggests that updates to individual orthology-prediction methods over time have a drastic effect on the landscape of orthologs between worms and humans, on the order of ~16% change in total gene content. However, as shown below, the meta-analysis approach of combining results from the different orthology-methods appears to buffer some of this change, in particular when it comes to apparent loss of orthology. The documentation associated with updates to the four previously-used orthology-prediction programs does not provide details of the changes to their algorithms that might have led to the large changes in gene content despite the minor changes in gene structure predictions that we found in both species (see sections above). We speculate that one possible reason behind the larger change in the landscape of orthologs after updates may be related to the inclusion of more sequenced genomes when orthology-prediction methods were updated. For example, in updating InParanoid from v7 (which was used for OL1) to v8 (analyzed here) the number of species included to generate ortholog groups increased from 100 to 273, leading to an increase in the number of ortholog groups of 423% (from 1.5 to 8.0 million) and orthologous proteins by 141%, from 1.2 to 3.0 million (Sonnhammer and Ostlund

2015). Such large-scale changes in orthology assignments seem likely to be the cause of the large shift in the landscape of orthologs predicted by the four previously-used methods.

Updates to orthology-prediction methods do not lead to greater agreement between them

Less than half of the worm genes in OL1 were supported by all four programs queried, suggesting a low degree of agreement between individual prediction methods. Moreover, ~20% of worm genes in OL1 were found by a single orthology-prediction method, and hence we term these genes “uniques” (Shaye and Greenwald 2011. See also Fig. S2). If updates to the orthology prediction programs generally resulted in improved prediction power, we reasoned that there should be greater agreement among them (i.e., an increase in worm genes found by all programs and/or a reduction in "uniques"). To this end, we performed the same meta-analysis on the results from updated versions of the previously-used orthology-prediction methods to generate OL1.1, which contains 7,812 worm genes (Fig. 3A-C and File S4).

Surprisingly, we found that updates to orthology-prediction methods actually resulted in less convergence among their results (Figs. 3A, B and S2): the proportion of *C. elegans* genes scored as having human orthologs by all four methods declined from 44.7% to 41.7% ($p=6.5x10^{-8}$. Statistical analysis here, and below, were done via two-tailed chi-square goodness-of-fit tests with Yates correction). Conversely, the proportion of "uniques" increased from 21.8% to 23.8% ($p=1.2x10^{-5}$).

Not only were the results from these programs less convergent after updating, but updates did not seem to provide stronger support for predictions. The majority of OL1 genes

(5,487 of 7,569, or 72.5%) remained in the same "class" (i.e., "unique", "found by two programs", "found by three programs", or "found by all") after updating to OL1.1 (Table 3 and Fig. 3D), suggesting the same level of support. However, among the genes that changed class, the number that lost support (e.g., went from being supported by all, to being supported by three, two or one, or those that went from unique to not being supported at all, etc.) outnumbered those that gained it: 1,285 genes (17.0%) lost support, while 797 genes (10.5%) gained it (Table 3). This difference is statistically significant ($p<0.001$), consistent with the decreased convergence in results from the different methods sampled.

We also note that the class a gene belonged to in OL1 does not appear to be a predictor of increased or decreased support after updates (Table 3 and Fig. 3D). Among genes that did not change support after updates, the most represented type (~52% of this class) were those predicted by all four methods before and after updates, however, the next most numerous class were those that remained unique (~21% of this class). By this metric, genes supported by two or three programs seem to be less stable. Among genes that lost support, the vast majority (~93%) only changed by one “level” (i.e., unique to lost, two to unique, three to two or all to three. Fig. 3D and Table 3). Somewhat surprisingly, the largest contributing set of genes to the class that lost support was the subset that was predicted by all four methods in OL1, suggesting that genes predicted by all methods are not necessarily the most likely to retain the highest level of support after updates.

In sum, our analysis shows that updates to orthology-prediction methods do not necessarily lead to greater agreement among them, nor do these updates unambiguously or consistently provide stronger support for specific predictions. These observations demonstrate the difficulty of assessing *a priori* which orthology prediction method is the most accurate, a

question that continues to be debated in the field of orthology prediction (Altenhoff *et al.* 2016). Thus, favoring one method over another, and relying on results from a single version in time of an orthology-prediction method, can introduce unintended bias and increase false-negative rates when compiling a comprehensive list of orthologs between species. A corollary that we discuss further below is that using the number of programs that support a prediction as a proxy for how good the prediction is, as several meta-analysis-based methods do (Hu *et al.* 2011; Pryszcz *et al.* 2011; Sutphin *et al.* 2016), is an uncertain metric, since the degree of support appears to be fluid. As we show in the next section, the meta-analysis approach also appears to guard against these potential problems.

Meta-analysis "buffers" against losses resulting from updates to individual orthology-prediction methods

When we compiled OL1 there was no ''gold standard'' for identifying a set of orthologs between two species. We argued that a meta-analysis would insure high recall and precision, resulting in the most accurate picture of *C. elegans* and human orthologs (Shaye and Greenwald 2011). Other studies (Pryszcz *et al.* 2011; Pereira *et al.* 2014) supported this inference, and show that the meta-analysis approach results in a higher level of accurately predicted ortholog groups than individual methods. Here, we provide additional support for this view by demonstrating that a meta-analysis effectively buffers against losses resulting from changes over time in individual prediction methods.

The meta-analysis used to generate OL1.1 led to a gain of 530 worm genes when compared to OL1 (Fig. 1 and 3C). As mentioned above, the mean gain in gene content when analyzing individual programs was 7.3%, corresponding to a predicted gain of ~568 worm

genes (Fig. 1 and File S3). Therefore, gains obtained with the meta-analysis are close (within the s.e.m) of the expected. This shows that, with respect to gene gains, the meta-analysis does not differ greatly from the variability seen within individual programs.

On the other hand, the meta-analysis resulted in a loss of just 287 genes (Figs. 1 and 2C). This contrasts with the mean loss in gene content seen with individual programs, which was 7.9%, corresponding to a predicted loss of ~598 genes (Fig. 1 and File S3). Thus, the number of genes lost using the meta-analysis is much less than what would be expected due to losses in individual programs. This suggests that the meta-analysis approach provides a “buffer” against loss in gene content due to changes in orthology-prediction methods over time.

The majority of worm genes lost after updating to OL1.1 (260 of 287, or ~90%) were "uniques" in OL1 (File S2, tab C and File S4, tabs D, E), suggesting that this class is the most likely to lose orthology after updates to prediction methods. However, two other considerations indicate that genes predicted by a single method should be included in OrthoList to ensure the most accurate representation of orthology: (1) it is important to note that the 260 lost genes represent just a small fraction (~16%) of the 1,650 "uniques" in OL1 (Fig. S2, File S2, tab B, and File S4, tab E), and (2) we found that a similar fraction of OL1 "uniques" (222 genes, or ~13%) are now supported by two, or more, programs used to compile OL1.1 (File S4, tab E.)

Adding more orthology-prediction methods has only a low impact on the landscape of human-worm orthologs identified in OL1.1

In choosing the prediction programs to generate OL1 we focused on those that, at the time, were rated highly by publications that analyzed the performance of orthology-prediction methods (Hulsen *et al.* 2006; Chen *et al.* 2007; Altenhoff and Dessimoz 2009) and were

amenable to extraction of genome-scale data. A more recent assessment of 15 orthology-prediction methods (Altenhoff *et al.* 2016), which did not include OrthoMCL or HomoloGene, continues to support InParanoid as a solid performer (i.e., generating results that balance precision with recall), while Ensembl Compara performed less well. We note that, in regard to the *C. elegans*-human set of orthologs, this assessment fits our observations: InParanoid seemed to be more stable over time, showing fewer changes in total gene number and content, when compared to Ensembl Compara (Table 2).

Our finding that OL1.1 displayed a gain of 530 and a loss of 287 genes when compared to OL1 led us to test if including results from additional orthology-prediction methods would support these changes, or reveal shortcomings in the methods used previously. We chose two additional orthology-prediction methods, the Orthologous Matrix (OMA) project (Roth *et al.* 2008) and OrthoInspector (Linard *et al.* 2011; Linard *et al.* 2015) (see Table 1) for their ease when it came to obtaining genome-wide data, and for their accuracy when compared to other orthology-prediction methods. In terms of recall and precision, among the 15 programs assessed by Altenhoff *et al.* 2016 OMA appears to be the most stringent, exhibiting the highest precision but with low recall (few false positives, but may miss true hits), while OrthoInspector typically exhibited the most well-balanced set of results with respect to precision and recall, being most similar in these respects to InParanoid.

OMA defines orthologs using a three-step process: first it analyzes full proteome sequences using all-against-all Smith-Waterman alignments. Second, to identify orthologous pairs from within significant alignment matches, closest homologs are identified based on evolutionary distance, taking into account an estimation of uncertainty, the possibility for differential gene losses, and identifying paralogs based on third-party proteome sequences as

"witnesses of non-orthology". Finally, ortholog groups are built using a maximum-weight clique algorithm. For our analysis, we downloaded the Humans-*C. elegans* “Genome Pair View” dataset from the OMA website.

The OrthoInspector algorithm is also divided into three main steps. First, the results of a BLAST all-versus-all alignment are parsed to find all the BLAST best hits for each protein within an organism, which is used to create groups of inparalogs. Second, the inparalog groups of each organism are compared in a pairwise fashion to define potential orthologs and inparalogs. Third, best hits that contradict the potential orthology between entities are detected and annotated. Unlike InParanoid and OrthoMCL, OrthoInspector does not consider reciprocal-best BLAST hits as a preliminary condition to detect potential inparalogs. Instead, inparalog groups are inferred directly in each organism, and these groups are then compared between organisms. This approach allows for exploration of a larger search space to discover potential orthologs.

When we compared results from OMA and OrthoInspector to OL1.1, we found that the addition of these two programs did not greatly change the landscape of human-worm orthologs predicted by the four previously-used methods (Fig. 4A, File S5). Of the 3,881 worm genes with human orthologs predicted by OMA, 3,768 (97.0%) were already present in OL1.1. Similarly, of the 5,361 worm genes predicted to have human orthologs by OrthoInspector, 5,343 (99.7%) were already present in OL1.1. Therefore, these two programs at first glance appear to have added 131 more predicted orthologs to OrthoList. However, we note that 31 of the 131 genes added by OMA and OrthoInspector were actually in OL1, but had been lost after the updates to the original orthology prediction methods that yielded OL1.1 (Fig. 1, 4B).

Therefore, the new content added by OMA and OrthoInspector is actually only 100 genes, or +1.3% of what was already present in OL1.1.

The final gene content of OrthoList 2

To generate OL2, we summed the content of OL1.1 and the 101 additional genes identified by OMA and OrthoInspector. As in our original meta-analysis, we included genes found by even a single program as a conservative approach to maximize the inclusion of genes with potential conservation, especially with the view of using OL2 as a guide for RNAi screens. Taken together, OL2 includes a total of 7,943 *C. elegans* genes, or ~41% of the protein-coding genome (Fig. 4C and File S5, Tab C).

After compiling OL2 we were left with 256 *C. elegans* genes that were previously predicted to have human orthologs, and thus were in OL1, but are not supported by current versions of orthology-prediction programs (Fig.4B, C, and File S5, Tab C). Below we discuss this gene set, which we term “legacy”, and why we chose to retain them in our searchable database even though they no longer score as orthologs in analysis programs.

As we noted above, there is some redundancy in Ensembl human gene entries. In the version used to compile OL2 (v89) Ensembl contained 20,310 protein coding genes and 2,751 alternative sequences (which are the ones that give rise to the extra IDs for genes like *NOTCH4*, as described above). Thus, there were a total of 23,061 human ENSG IDs, of which ~13.5% were alternative sequences. OL2 has 12,345 ENSG IDs, which, given the numbers above, we estimate corresponds to ~10,678 bona fide protein-coding genes and ~1,667 alternative sequences. These considerations indicate that ~52.6% (10,678/20,310) of the human protein-coding genome has recognizable worm orthologs supported by current versions

of orthology-prediction methods.

The "legacy gene" set

We found that 256 *C. elegans* genes that were present in OL1 were not identified either in OL1.1, using updates of the four original programs, or by OMA or OrthoInspector (File S5, tab C, and S6 tab A). Thus, they would not be considered "orthologs" as conventionally defined. Many (205/256, or ~80%) of these genes have functional domains recognized by programs such as SMART (Letunic and Bork 2018) and InterPro (Finn *et al.* 2017), while others have been placed in protein families based on other criteria; e.g., the C/EBP protein homolog *cebp-1* (Yan *et al.* 2009; Bounoutas *et al.* 2011; Kim *et al.* 2016; McEwan *et al.* 2016), or several hedgehog-related genes, called "groundhog" or *grd* in *C. elegans*, (Burglin and Kuwabara 2006) (see File S6, tab A). In addition, some of these genes have been discussed as orthologs in the literature, based on their inclusion in OL1 or by independent analyses using the underlying prediction programs or other methods. We therefore needed to consider how to deal with such genes in our new meta-analysis here, and as will be described below, we concluded that we needed a special designation for such "legacy genes" that would recognize their history without considering them current orthologs.

We give here four examples of *C. elegans* genes that illustrate properties of these "legacy" genes and complications of orthology prediction. All four were included in OL1 based on Ensembl Compara, a program that performed less well in the assessment of Altenhoff *et al.* (2016), and which was also the least congruent with the others, and thus provided the most unique hits in OL1 (Shaye and Greenwald 2011. See also Fig. S2).

(i) *C. elegans cdk-2* is not predicted by any of the six programs used here. Nevertheless, *cdk-2* is functionally related to human CDK2 in that it regulates cell cycle progression from G1-S phase (Fox *et al.* 2011; Korzelius *et al.* 2011). BLAST analysis indicates that *C. elegans* CDK-2 is 52% identical to human CDK2 and has a low "e-value", but CDK-2 would not be predicted as an "ortholog" by Reciprocal Best Hits (RBH), a simple assessment of orthology (Altenhoff *et al.* 2016) because if *C. elegans* CDK-2 is used as the query in a BLAST search of the human database, CDK3 and CDK1 have higher e-values, whereas if human CDK2 is used as a query of *C. elegans*, CDK-1 and CDK-5 have higher e-values. This situation may be relatively rare, but underscores the complexity of ascertaining phylogenetic relationships of individual genes of gene families.

(ii) *C. elegans ceh-51* encodes a homeodomain-containing transcription factor that functions in mesoderm (Broitman-Maduro *et al.* 2009). In OL1 it was called as the ortholog of *VENTX*, a homeodomain transcription factor that functions in the human mesodermal derivatives of the myeloid lineage (Rawat *et al.* 2010; Wu *et al.* 2011; Gao *et al.* 2012; Wu *et al.* 2014). In OL2, four other *C. elegans* homeodomain (*ceh*) genes are now called as *VENTX* orthologs, underscoring how adjustments to the prediction programs may lead to shifts in which possible paralogs in *C. elegans* are called as orthologs of human genes.

(iii) *C. elegans* FOS-1 is a transcription factor required for the gonadal anchor cell to breach a basement membrane during vulval development (Sherwood *et al.* 2005). In OL1, ENSEMBL Compara predicted a total of six genes as potential orthologs: c-FOS and five additional FOS-related genes, all bZIP proteins containing a "BRLZ" domain according to SMART (Letunic and Bork 2018). In contrast to *ceh-51*, where there seemed to be a shift in

the orthology call, here, none of the paralogs or other proteins with BRLZ domains in humans were called as *fos-1* orthologs in OL2.

(iv) *C. elegans* SEL-8, a core component of the Notch signaling system, is a glutamine-rich protein that appears to be homologous to the glutamine-rich human MAML proteins based on its equivalent role in a ternary complex with the Notch intracellular domain and the LAG-1/CSL DNA binding protein, even though there is no primary sequence similarity or any recognizable domains (Doyle *et al.* 2000; Petcherski and Kimble 2000; Wu *et al.* 2000). However, in OL1, Compara predicted SEL-8 to be homologous to MED15, a component of the Mediator complex (Allen and Taatjes 2015), while InParanoid uniquely predicted *C. elegans* MDT-15 as the ortholog of human MED15, a relationship that is also consistent with the SMART protein domain prediction.

The 256 worm genes that comprise the legacy set were previously found to be orthologous to 382 human ENSG IDs. Of these, 217 (~57%) continue to have worm orthologs, and thus are included in OL2. The remaining ENSG IDs, corresponding to 165 individual human genes, do not have currently-supported worm orthologs, and thus represent the human legacy set of genes (File S6, tab B).

Given that one of the incentives for compiling OrthoList was to obtain the most comprehensive set of functionally similar human-worm homologs for cross-species studies, and to acknowledge the publication history of these genes as orthologs if questions arose in the future, we have retained these worm and human genes as a "legacy set" (File S6), clearly indicating that they were not found as orthologs per se by current programs. Their change of status underscores the difficulty of identifying orthologs between *C. elegans* and humans, which have such a distant evolutionary relationship.

An OrthoList 2 online tool with enhanced search capabilities and links to external databases

OL1 was originally published in the form of a set of Excel spreadsheets (Shaye and Greenwald 2011). However, this form limited its utility, and may have led to some confusion when searching for worm genes with human orthologs, as evidenced by publications that referenced OL1, but missed genes that were in the spreadsheet and thus reported a lower degree of reliability for this list [e.g. (Roy *et al.* 2014)]. To facilitate access, we subsequently developed a basic online tool, which was never formally published but instead publicized through a reader comment at the original journal website and announcements in *C. elegans* venues. This simple tool allowed *C. elegans* genes to be input (through their gene or locus name, or WormBase ID), and human genes to be input via ENSG ID, and outputs were similarly displayed.

To access OL2, we have developed a significantly improved online tool (http://ortholist.shaye-lab.org) with several features (Fig. 5A) not present in the original version made available informally to the community. As before, searches may be conducted using *C. elegans* or human gene identifiers, but importantly, this feature is now augmented by the ability to search using HGNC names (Yates et al., 2017) and the ability to permit partial matches to facilitate searches when there are multiple paralogs, such as "*NOTCH*," for the four paralogs *NOTCH1-4* when the "partial match allowed" option is selected. Additionally, we now include the ability to query the database based on InterPro (Finn *et al.* 2017) and SMART (Letunic and Bork 2018) protein domain annotations, and human-disease associations provided by the OMIM database (McKusick 2007). We also provide the option to restrict searches based on a

given number of programs that predict an orthologous relationship, but as we discuss below, we believe that “unique” hits in OL2 should be viewed as orthologs since they fit the criteria used by the most recent version of a validated program. The “legacy” genes described above are also found in this online tool, and can be included in searches by selecting “no minimum” in the “No. of programs” field. Finally, we include an “Instructions, Tips and Feedback” section, which we can update in response to user feedback.

In the results page (Fig. 5B), users will find the number of programs that call a particular *C. elegans*-human ortholog prediction (Fig. 5B). If the result displays a “0” in this column, the genes returned are from the “legacy” set, and are not considered “orthologs” at this time (see "Discussion" of legacy genes below). If a result displays 1 or more programs, hovering over the "?" symbol shows which program(s) called a particular orthology relationship. The results may be sorted by clicking at the top of the columns for any of the names (WormBase ID, Common Name Locus ID, Ensembl ID, or HGNC Symbol) or the number of programs. Finally, we include links to SMART and InterPro protein domain descriptions, as well as to OMIM entries for human disease associations. Clicking on “toggle” displays links for the entire column; clicking on “view” displays the links for a given gene.

One of the rationales for creating OrthoList was to facilitate RNAi screens by pre-selecting genes with human orthologs (Shaye and Greenwald 2011). To this end we had incorporated identifiers for the most utilized and extensive set of “feeding RNAi” clones (Fraser *et al.* 2000; Kamath *et al.* 2003) to our informally-released online tool. When initially produced, the feeding library targeted ~72% of *C. elegans* genes. More recently a collection of new bacterial strains was produced to supplement and enhance this library, which now targets ~87% of currently-annotated genes (https://www.sourcebioscience.com/products/life-sciences-

research/clones/rnai-resources/c-elegans-rnai-collection-ahringer/). We have now added clone identifiers for this newly-released supplemental RNAi set to our database in order to provide the most up-to-date resource for finding RNAi clones that target genes conserved in humans.

DISCUSSION

C. elegans is a powerful experimental system for using genetic approaches to address biological problems of relevance to human development, physiology, and disease. Harnessing the full power of the system is enhanced by the knowledge of evolutionarily related genes (homologs) between *C. elegans* and humans. Homologs across species are often divided into those that originated through speciation (orthologs) and those that originated through duplication (paralogs). Although orthology is an evolutionary, and not necessarily a functional, definition, the "ortholog conjecture" proposes that orthologs tend to maintain function, whereas paralogs, are more diverged. However, recent work suggests that even paralogs retain significant functional similarity (Altenhoff *et al.* 2012; Gabaldon and Koonin 2013; Dunn *et al.* 2018). Therefore, as a proxy for functional conservation, establishing the orthology relationship among genes in different species has served as a useful tool to identify candidates for cross-species and translational studies. However, identifying homologs is not a simple undertaking, and a wide variety of methods exist, with different balances between precision (positive predictive value) and recall (true positive rate) (Altenhoff *et al.* 2016). Furthermore, there are different versions of genome sequence databases and curation of predicted genes, and different versions of prediction programs.

We had previously used a meta-analysis approach to compile OrthoList, a compendium of *C. elegans* genes with human orthologs (Shaye and Greenwald 2011). Initially compiled for

the practical purpose of streamlining RNAi screens, it also had value as a study of the relationship between the two genomes. Here, we have created OrthoList 2 (OL2), a new meta-analysis, which has similar value as both a practical tool and for insights into the genomes. We consider three main topics in this Discussion. First, we discuss how our longitudinal analysis here reveals that the meta-analysis approach is not just more accurate as a snapshot view of the relationship between the genomes, but also means that OL2 will remain a practical tool for facilitating cross-platform studies for many years to come. Next, we discuss how our results suggest that assigning reliability scores in meta-analysis approaches, a common component of studies that followed OrthoList, may be misleading. Finally, we provide a practicum on what to do when a gene of interest is, or is not, found in OL2.

The meta-analysis approach results in a stable landscape of orthologs

The initial rationale for performing a meta-analysis to generate a compendium of human-worm orthologs was based on the fact that, at the time we compiled OL1, there was no reliable benchmark that defined which orthology-prediction method was the “best”. Another publication that used meta-analysis to study genome-wide orthology, published at the same time as OL1 (Pryszcz *et al.* 2011), generated the "Meta-Phylogeny-Based Orthologs (MetaPhOrs)" database (now offline), and a subsequent study (Pereira *et al.* 2014) that developed a "Meta-Approach Requiring Intersections for Ortholog predictions (MARIO)" further supported the idea that a meta-analysis results in a higher level of accurately predicted ortholog groups than individual methods.

Our work here not only shows that the meta-analysis approach provides more accurate predictions, but also generates a robust set of orthologs that withstand the test of time. Indeed,

to our knowledge, this study is the first to assess how changes in gene structure and orthology prediction methods over time (a longitudinal analysis) affects the landscape of orthologs between two species, and the effect that the meta-analysis approach has on these changes. Although very few of the worm-human orthologs predicted in OL1 (<1%) were lost due to changes in underlying gene predictions over the last ~7 years, we find that there have been significant changes in gene content within individual orthology-prediction methods over time, indicating that genome-wide orthology inference based on a single version of any individual orthology-prediction method will miss orthology relationships. Furthermore, these changes did not lead to greater agreement between methods. However, our meta-analysis approach buffered against ortholog losses that led to this divergence between methods, demonstrating a further, unexpected advantage of this approach.

This stability means that OL2 will remain a practical tool for facilitating cross-platform studies for many more years. This observation is important because there is a large labor cost to the manual curation and quality control steps required to ensure that results from new methods are appropriately vetted. For example, we found that a bottleneck of manual curation was required to ensure that gene IDs for *C. elegans* and humans were not deprecated, changed or retired. We also needed to take manual curation steps in order to confirm that no errors were introduced upon large-scale conversion of gene IDs (which tend to be different for each program) to forms that can be directly compared. We note that it is not clear from the published reports if these steps were taken for other published meta-analysis approaches.

Evaluating the utility of reliability scores in meta-analysis approaches

Two different approaches have been used to infer reliability of predictions in meta-analyses. One is to use the number of methods that support an orthology prediction as a "simple score" for the reliability of the prediction. The other is to use different “weighting” approaches to emphasize predictions of some methods over others. However, our results here raise doubts as to whether either of these approaches is an appropriate scoring methodology, because the level of support is not only dependent on which programs are used, but also on when these programs were sampled. Furthermore, our work, and that of Pryszcz *et al.*, demonstrates that increasing the number of orthology-prediction methods does not have a major impact on the performance of a meta-analysis. The study that generated the MetaPhOrs database (Pryszcz *et al.* 2011) noted a significant increase in recall (fewer false negatives) when results from two orthology-prediction programs were combined, compared to when individual programs were sampled. However, there was little difference in recall, or precision, metrics when results from a third program were added to the combinations of two. Our results here support this observation, as addition of two more programs (OMA and OrthoInspector) to the four already used for OL1 did not greatly increase recall, leading to addition of only ~100 worm genes to OrthoList. Given the lack of correlation between having more programs in the meta-analysis, and increased recall or precision, we caution researchers against discarding hits with lower simple scores, for example “"uniques"”, as it would lead to a higher false-negative rate when performing large-scale studies using meta-analysis-derived databases.

Two other meta-analyses, DIOPT (Hu *et al.* 2011), which samples 15 different orthology-prediction methods, and WORMHOLE (Sutphin *et al.* 2016), which samples 14 methods, use alternative, weighted, approaches to score reliability. DIOPT assigns a different weight to each underlying orthology-prediction program based on how well each performs in a

"functional" assessment; namely, the degree of semantic similarity between high quality GO molecular function annotations of fly-human ortholog pairs predicted by each method sampled. Unfortunately, several reports have shown that GO annotation congruence as a proxy for functional similarity is a problematic metric (Chen and Zhang 2012; Thomas *et al.* 2012). Moreover, it is not clear how GO semantic similarity applied to fly-human ortholog pairs translates to other species, particularly *C. elegans* and humans. Therefore, it is not clear that this weighing approach is better than the "simple scoring" approach, and, as discussed above, even the “simple scoring” approach can introduce a higher level of false negative calls.

WORMHOLE developed a "scaled" confidence score based on a supervised learning model that analyzes data for classification purposes called a support vector machine (SVM) classifier system. An SVM uses a set of training examples, each marked as belonging to one or another of two categories (in the case of WORMHOLE, the categories were: being a least-diverged ortholog (LDO) vs. not), then the SVM training algorithm builds a model to assign new examples (i.e. putative ortholog pairs) to one category or the other. WORMHOLE used the PANTHER LDO dataset (Mi *et al.* 2013) as reference for training their SVM. This training set includes all one-to-one orthologs, as well as the single least divergent gene pair in one-to-many and many-to-many ortholog groups within the broader PANTHER ortholog dataset. PANTHER LDOs perform well in orthology benchmarking assessments, however this set tends to be very conservative (Altenhoff *et al.* 2016): consistently showing high precision, but low recall (i.e., missing a lot of possible orthologs compared to other programs). Therefore, using the PANTHER LDO set as the "training" algorithm to generate a confidence score has the potential of missing bona fide orthologs.

We have included the number and identity of programs for each gene in OL2 for reference, but given the various difficulties of current scoring systems we consider here, we believe that the best approach is to avoid using scoring criteria to support, or contradict, orthology assignments achieved via meta-analysis, and to consider any gene identified by at least one program as an ortholog for all practical purposes.

A gene is, or is not, in OL2: what does that mean?

OrthoList has proven to be a useful way to streamline RNAi screens and to ask questions about the genome, particularly as a first step to ask if a gene of interest in one system has an ortholog in the other. However, the vast evolutionary distance between *C. elegans* and humans has allowed for extensive sequence divergence, as well as for larger-scale genomic alterations, such as domain shuffling and local, or genome-scale, duplications (Babushok *et al.* 2007). Given the existence of such mechanisms for genome divergence, which can impact the ways that phylogenetic relationships are inferred by orthology-prediction programs, the presence or absence of a gene in OL2 should not be the only consideration when deciding about homology. We consider here some common scenarios we have observed when using a worm gene to query OL2, other tests and extensions to support claims of orthology, and other approaches to find potential orthologs that elude identification by the programs used here, even though, as described above, they are generally high-performing and use different criteria in assessing orthology relationships. The same scenarios could apply in principle when a human gene is used to identify the worm ortholog(s).

(i) using a worm gene as the query returns a set of human paralogs. E.g. *wnk-1* elicits the four paralogs, *WNK1, WNK2, WNK3* and *WNK4.* The *C. elegans* gene is the ortholog of all four of these paralogous human genes, not just the eponymous *WNK1*. Thus, functional

information about *C. elegans wnk-1* may be applied to any of the four human genes, and vice versa.

(ii) using a worm gene as the query returns a set of non-paralogous human genes. This may occur when proteins share a domain but differ otherwise. For example, entering *C. elegans lin-12* identifies the four human *NOTCH* genes, as expected. However, two programs also call the gene *EYS*, and two single programs (Compara or OrthoMCL) call ten additional non-paralogous human genes. These additional genes encode proteins with EGF-like motifs, which are also found in *bona fide* NOTCH proteins, but lack the other hallmark domains of NOTCH. The real NOTCH proteins, including LIN-12, have a similar domain architecture with several identifiable domains in a similar arrangement, and therefore can easily be distinguished from the proteins that contain EGF-like motifs, but are otherwise dissimilar, by using a domain architecture program such as SMART. However, for proteins with single identifiable domains, domain architecture will not resolve which of the set of non-paralogous genes is the "ortholog."

(iii) using a worm gene as the query only identifies “legacy” relationships. Because the longitudinal analysis presented here has not been performed before, we devised the concept of “legacy genes” as a category for genes that were called orthologs in OL1 but are no longer called as such in OL2. When a gene is no longer called as an ortholog by contemporary programs, it cannot be considered an ortholog in the phylogenetic sense presented at the outset of this Discussion. Nevertheless, we retained "legacy” genes in the searchable database because many have recognizable functional domains (File S6, Tab A), and, in some cases, additional work has established conserved function (e.g. *cdk-2* and *sel-8* discussed above), suggesting that additional work on other "legacy" genes may yet support “orthology”.

Thus, if a gene of interest only exists in the legacy set, it will likely have a domain that gives some clue as to its function, or it may be that future work will establish conserved function even in the absence of strict phylogenetic orthology.

(iv) using a worm gene as the query does not identify any potential human orthologs. If there are identifiable domains, domain architecture searches may yield potential functional orthologs.

An important key to resolving these questions comes from the ability to use genetic analysis in *C. elegans* for functional assessment. The most straightforward approach is to use functional, trans-species rescue of a *C. elegans* mutant by expression of a human protein to bolster an inference of orthology. Indeed, the question of "orthology" vs. analogy/convergence becomes moot for practical purposes if the human protein can replace the *C. elegans* protein. Similarly, the conservation of biochemical/molecular function of different human paralogs can be assessed by a rescue assay. Eventually, similarities at the level of higher-order structure may be another way to identify worm-human orthologs that have diverged at the primary amino acid sequence level.

Finally, as noted previously (Shaye and Greenwald 2011), some components of pathways or complexes have diverged to the point that they are not identified by primary sequence and hence are not in our compendium. In such cases, the presence of some components of conserved pathways or complexes will essentially compensate for the absence of others when performing RNAi screens streamlined by OL2. To illustrate this point, we consider the conserved Notch pathway (Greenwald and Kovall 2013). Notch is essentially a membrane-tethered transcriptional coactivator regulated by ligand. When ligand binds, the intracellular domain is released by proteolytic cleavage to join a nuclear complex to activate

target genes. The *C. elegans* Notch orthologs, LIN-12 and GLP-1, the protease components that cleave the transmembrane form to release the intracellular domain, and the associated DNA binding protein LAG-1 are all present in OL2; the canonical DSL transmembrane ligands, LAG-2, APX-1 and ARG-1, and the SEL-8 Mastermind-like protein are not. Thus, if the Notch pathway is involved in a phenotype of interest, then enough components would be present in a streamlined, but otherwise unbiased, RNAi screen based on OL2.

OrthoList has already been used to design streamlined RNAi screens that yielded important discoveries (e.g., Gillard *et al.* 2015; Hernando-Rodriguez *et al.* 2018; Nordquist *et al.* 2018). To further facilitate the design of such screens, our new web-based tool not only includes the most up-to-date version of the widely-used *C. elegans* feeding RNAi library, but it also allows users to focus their screens even further by generating lists based on protein domains and/or human-disease associations. Therefore, our work here not only updates the genome-wide orthology between humans and *C. elegans*, it offers insight into how to evaluate results from orthology-prediction methods, and provides an easily accessible tool that will aid in streamlining functional studies and analyzing results with translational potential.

ACKNOWLEDGMENTS

We thank Claire de la Cova and Hana Littleford for helpful comments on the manuscript; Eashan Bhattacharyya, James Chen and Amrapali Patil for assistance; and Jan Kitajewski for support and encouragement. This work was supported by grants GM114140 and GM115718 (to I.G.) and HL119043-02S1 (to D.D.S).

FIGURE LEGENDS

Figure 1: Workflow for genome analysis and generation of OrthoList 2. The workflow proceeded in four steps. Step 1) we addressed changes to gene models in the worm genome that have occurred since OL1 was published (File S1) to yield an updated OL1 (File S2). We also addressed changes to human gene predictions (File S3). Step 2) we queried updated versions of the orthology-prediction methods used in OL1 (See Table 1) to generate OL1.1 (File S4), and found that the number of worm genes added was within the parameters predicted by changes in individual programs (Table 2), whereas gene loss appeared to be "buffered" by combining results from the different methods (i.e. the meta-analysis approach). Step 3) we next added results from two additional orthology-prediction methods (See Table 1) and found that this had a low impact on the landscape of human-worm orthologs identified in OL1.1 (File S5). Finally, in step 4) we combined the genes identified by these two additional programs with OL1.1, to generate OrthoList 2 (Files S5, S7). We note that genes that did not continue to be supported by orthology-prediction methods were retained as a "legacy" set present in the searchable database (Files S6, S7). Both OL2 and the legacy set of genes were cross-referenced to the *C. elegans* “feeding” RNAi library, protein-domain prediction databases (InterPro and SMART), and to a human-disease association database (OMIM) to generate a final master list (File S7) which can be queried via the new web-based tool found at ortholist.shaye-lab.org.

Figure 2: Changes in the landscape of *C. elegans* genes with human orthologs due to updates in methods used to generate the original OrthoList. Venn diagrams shown here compare the worm gene content of the original and updated versions of **(A)** Ensembl Compara, **(B)** HomoloGene, **(C)** InParanoid and **(D)** OrthoMCL. See also Table 2.

Figure 3: OL1.1 and longitudinal analysis of changes in the landscape of worm-human orthologs. To generate OL1.1 we combined results from updated versions of the four previously-used orthology prediction methods. The Venn diagram **(A)** shows overlap in gene content between the four programs, while the table **(B)** gives an overall measure of how many genes were found by one or more programs (regardless of which one(s) found them). The Venn diagram in **(C)** shows the change in gene content between OL1 (Fig. S2, File S2) and OL1.1 (File S4), indicating a loss of 287 genes and a gain of 530 genes after updates to orthology prediction methods. Bar graph in **(D)** illustrates the changes in orthology support after updates, also described in Table 3, demonstrating that most genes maintained the same level of support, but among those that changed support level, there was no obvious trend towards gaining more support with updates to prediction methods, nor was there more stability among genes that had higher support in OL1.

Figure 4: Adding more orthology-prediction methods has a low impact on the landscape of human-worm orthologs identified in OL1.1. We queried two additional programs, OMA and OrthoInspector, for worm-human orthologs, and compared their gene content to OL1.1. The Venn diagram in **(A)** shows that the vast majority of orthologs called by OMA (3,768/3,881 or ~97%) and OrthoInspector (5,343/5,361 or ~99%) were present in OL1.1. The Venn Diagram in **(B)** shows that among the “new” orthologs called by OMA and OrthoInspector ~24% (31/131) were in OL1, but had been lost due to updates to previously-used methods. Therefore, only 100 new orthologs were added after including results from two more orthology-prediction methods. Diagram in **(C)** shows how the gene content of OrthoList 2, which was

compiled by combining the results shown in panel A (see File S5, tab C), compares to the gene content from OrthoList 1 (see File S1, tab C).

Figure 5: OrthoList 2 query interface. Panel **(A)** shows input page at ortholist.shaye-lab.org. Users can select which fields to search (human and worm identifiers, SMART or InterPro protein domains, and disease phenotypes described in OMIM), whether to set a threshold for orthology support (see main text) and whether partial matches should be allowed, which is useful when users want to find all members of a similarly-named gene family (e.g., input "Notch" to find all human Notch family members). Panel **(B)** shows a sample results page for the gene *let-60*, with a search conducted using the default settings, returning a set of Ras orthologs consistent with its sequence and genetic validation in a canonical Ras pathway (Han and Sternberg 1990; Sundaram 2013). The results page contains links for viewing additional information about results and for exporting results to a comma-separated value (CSV) spreadsheet.

SUPPLEMENTAL MATERIAL

Supplemental Figure S1: Changes in gene content in Ensembl Compara v87, v88 and v89. The Ensembl Compara database was updated three times while we were compiling OL2. In this time-span we noticed that the landscape of worm genes with predicted human orthologs changed after each update, so that each version had ~2% of genes unique to it, while another ~2-4% of genes were found in only two of the three versions (see also Materials and Methods).

Supplemental Figure S2: Updated OL1. We updated OL1 by addressing changes in worm gene structure, classification and nomenclature for the genes present in our original compendium. We then combined results from the corrected OL1 programs. The Venn diagram **(A)** shows overlap in corrected gene content between the four programs, while the table **(B)** gives an overall measure of how many genes were found by one or more programs (regardless of which one(s) found them).

Supplemental Table S1: Data sources for orthology-prediction programs used to compile OL2. The source data for each program is found at each program's website.

Supplemental File S1: Changed OL1 worm genes. This file lists genes whose classification, or ID, changed since the release of OL1. Type I changes correspond to genes that were re-classified as pseudogenes, ncRNA, being transposon-derived, or killed due to lack of evidence. Type II changes results from combining, or "merging" two or more genes that had each, separately, been found to have a human ortholog in OL1. Type III changes represent genes that were assigned new IDs, either because experimental evidence suggested that they should be merged with genes previously not in OL1 (marked red), or due to addition of previously unpredicted gene segments (denoted as a red "?")

Supplemental File S2: Corrected *C. elegans* genes in OL1. All corrected worm genes found by each OL1-era version of orthology prediction methods are shown in tab **(A)**. Tab **(B)** shows the distribution of results between OL1-era orthology-prediction methods, while tab **(C)** shows

the corrected OL1 as well as the distribution of genes by support class (supported by one, two, three or all methods).

Supplemental File S3: Changed OL1 human genes. Human ENSG gene IDs from OL1 are listed for each orthology-prediction method in tab **(A)**. This tab also shows the 574 ENSG IDs that are no longer found in current versions of the Ensembl genome browser. Tab **(B)** shows the Ensembl-provided history for the 574 lost ENSG IDs, showing that most are now just classified as “retired”. Tab **(C)** shows a randomly selected subset of 20 IDs that were “retired”. Note that the gene name (HGNC-approved symbol) associated with the “retired” ENSG ID is always associated with current ENSG IDs, demonstrating that curation of ENSG IDs rarely links “retired” IDs with their current counterparts. Tab **(D)** lists the sixteen human ENSG IDs that we could confirm were deprecated.

Supplemental File S4: *C. elegans* genes in OL1.1. Tab **(A)** shows the worm genes found to have human orthologs by updated versions of prediction methods used in OL1. Tab **(B)** shows the distribution of results between orthology-prediction methods. Tab **(C)** shows the final OL1.1, as well as the distribution of genes by support class (supported by one, two, three or all methods). Tab **(D**) lists those genes found only in OL1 (termed “lost”), and those added upon update to OL1.1.

Supplemental File S5: *C. elegans* OMA and OrthoInspector results, their relationship to OL1.1 and genes not supported by current versions of orthology-prediction methods. Tab **(A)** shows the worm genes found to have human orthologs by OMA, OrthoInspector and

those already in OL1.1. Tab **(B)** shows the distribution of results amongst these three sets. Tab **(C)** lists all worm genes with human orthologs supported by current orthology-prediction methods (OL2) as well as those no longer supported (the "legacy" set).

Supplemental File S6: the "legacy" set. Tab **(A)** lists the 256 *C. elegans* genes previously-predicted to have human orthologs, but not supported by current versions of orthology-prediction methods, and their predicted protein domains determined by SMART and InterPro. Tab **(B)** lists the human "legacy" set: 165 human genes that were previously predicted to have worm orthologs, but for whom orthology is no longer supported.

Supplemental File S7: OL2 and legacy master list. This file, which underlies the database hosted at ortholist.shaye-lab.org, contains all orthology predictions (current and legacy), with *C. elegans* and human gene identifiers, as well as associated protein domain (SMART and InterPro) and human disease (OMIM) information.

Supplemental File S8: Freeze of code used to compile OrthoList 2. The code was downloaded from https://github.com/danshaye/OrthoList2 at the time of submission.

LITERATURE CITED

Allen, A. K., J. E. Nesmith and A. Golden, 2014 An RNAi-based suppressor screen identifies interactors of the Myt1 ortholog of *Caenorhabditis elegans*. G3 (Bethesda) 4**:** 2329-2343.

Allen, B. L., and D. J. Taatjes, 2015 The Mediator complex: a central integrator of transcription. Nat Rev Mol Cell Biol 16**:** 155-166.

Altenhoff, A. M., B. Boeckmann, S. Capella-Gutierrez, D. A. Dalquen, T. DeLuca *et al.*, 2016 Standardized benchmarking in the quest for orthologs. Nat Methods 13**:** 425-430.

Altenhoff, A. M., and C. Dessimoz, 2009 Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol 5**:** e1000262.

Altenhoff, A. M., R. A. Studer, M. Robinson-Rechavi and C. Dessimoz, 2012 Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol 8**:** e1002514.

Babushok, D. V., E. M. Ostertag and H. H. Kazazian, Jr., 2007 Current topics in genome evolution: molecular mechanisms of new gene formation. Cell Mol Life Sci 64**:** 542-554.

Balklava, Z., S. Pant, H. Fares and B. D. Grant, 2007 Genome-wide analysis identifies a general requirement for polarity proteins in endocytic traffic. Nat Cell Biol 9**:** 1066-1073.

Bounoutas, A., J. Kratz, L. Emtage, C. Ma, K. C. Nguyen *et al.*, 2011 Microtubule depolymerization in *Caenorhabditis elegans* touch receptor neurons reduces gene expression through a p38 MAPK pathway. Proc Natl Acad Sci U S A 108**:** 3982-3987.

Broitman-Maduro, G., M. Owraghi, W. W. Hung, S. Kuntz, P. W. Sternberg *et al.*, 2009 The NK-2 class homeodomain factor CEH-51 and the T-box factor TBX-35 have overlapping function in *C. elegans* mesoderm development. Development 136**:** 2735-2746.

Burglin, T. R., and P. E. Kuwabara, 2006 Homologs of the Hh signalling network in *C. elegans*. WormBook**:** 1-14.

Chen, F., A. J. Mackey, J. K. Vermunt and D. S. Roos, 2007 Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One 2**:** e383.

Chen, X., and J. Zhang, 2012 The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. PLoS Comput Biol 8**:** e1002784.

Dickinson, D. J., and B. Goldstein, 2016 CRISPR-Based Methods for *Caenorhabditis elegans* Genome Engineering. Genetics 202**:** 885-901.

Doyle, T. G., C. Wen and I. Greenwald, 2000 SEL-8, a nuclear protein required for LIN-12 and GLP-1 signaling in *Caenorhabditis elegans*. Proc Natl Acad Sci U S A 97**:** 7877-7881.

Du, Z., A. Santella, F. He, P. K. Shah, Y. Kamikawa *et al.*, 2015 The Regulatory Landscape of Lineage Differentiation in a Metazoan Embryo. Dev Cell 34**:** 592-607.

Dunn, C. D., M. L. Sulis, A. A. Ferrando and I. Greenwald, 2010 A conserved tetraspanin subfamily promotes Notch signaling in *Caenorhabditis elegans* and in human cells. Proc Natl Acad Sci U S A 107**:** 5907-5912.

Dunn, C. W., F. Zapata, C. Munro, S. Siebert and A. Hejnol, 2018 Pairwise comparisons across species are problematic when analyzing functional genomic data. Proc Natl Acad Sci U S A 115**:** E409-E417.

Finn, R. D., T. K. Attwood, P. C. Babbitt, A. Bateman, P. Bork *et al.*, 2017 InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res 45**:** D190-D199.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver *et al.*, 1998 Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391**:** 806-811.

Firnhaber, C., and M. Hammarlund, 2013 Neuron-specific feeding RNAi in *C. elegans* and its use in a screen for essential genes required for GABA neuron function. PLoS Genet 9**:** e1003921.

Fox, P. M., V. E. Vought, M. Hanazawa, M. H. Lee, E. M. Maine *et al.*, 2011 Cyclin E and CDK-2 regulate proliferative cell fate and cell cycle progression in the *C. elegans* germline. Development 138**:** 2223-2234.

Fraser, A. G., R. S. Kamath, P. Zipperlen, M. Martinez-Campos, M. Sohrmann *et al.*, 2000 Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. Nature 408**:** 325-330.

Gabaldon, T., and E. V. Koonin, 2013 Functional and evolutionary implications of gene orthology. Nat Rev Genet 14**:** 360-366.

Gao, H., X. Wu, Y. Sun, S. Zhou, L. E. Silberstein *et al.*, 2012 Suppression of homeobox transcription factor VentX promotes expansion of human hematopoietic stem/multipotent progenitor cells. J Biol Chem 287**:** 29979-29987.

Gillard, G., M. Shafaq-Zadah, O. Nicolle, R. Damaj, J. Pecreaux *et al.*, 2015 Control of E-cadherin apical localisation and morphogenesis by a SOAP-1/AP-1/clathrin pathway in *C. elegans* epidermal cells. Development 142**:** 1684-1694.

Golden, A., 2017 From phenologs to silent suppressors: Identifying potential therapeutic targets for human disease. Mol Reprod Dev 84**:** 1118-1132.

Greenwald, I., 2012 Notch and the Awesome Power of Genetics. Genetics 191**:** 655-669.

Greenwald, I., and R. Kovall, 2013 Notch signaling: genetics and structure. WormBook**:** 1-28.

Han, M., and P. W. Sternberg, 1990 let-60, a gene that specifies cell fates during *C. elegans* vulval induction, encodes a ras protein. Cell 63**:** 921-931.

Hernando-Rodriguez, B., A. P. Erinjeri, M. J. Rodriguez-Palero, V. Millar, S. Gonzalez-Hernandez *et al.*, 2018 Combined flow cytometry and high-throughput image analysis for the study of essential genes in *Caenorhabditis elegans*. BMC Biol 16**:** 36.

Hu, Y., I. Flockhart, A. Vinayagam, C. Bergwitz, B. Berger *et al.*, 2011 An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics 12**:** 357.

Hulsen, T., M. A. Huynen, J. de Vlieg and P. M. Groenen, 2006 Benchmarking ortholog identification methods using functional genomics data. Genome Biol 7**:** R31.

Kamath, R. S., A. G. Fraser, Y. Dong, G. Poulin, R. Durbin *et al.*, 2003 Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature 421**:** 231-237.

Kasprzyk, A., 2011 BioMart: driving a paradigm change in biological data management. Database (Oxford) 2011**:** bar049.

Kim, K. W., N. Thakur, C. A. Piggott, S. Omi, J. Polanowska *et al.*, 2016 Coordinated inhibition of C/EBP by Tribbles in multiple tissues is essential for *Caenorhabditis elegans* development. BMC Biol 14**:** 104.

Korzelius, J., I. The, S. Ruijtenberg, M. B. Prinsen, V. Portegijs *et al.*, 2011 *Caenorhabditis elegans* cyclin D/CDK4 and cyclin E/CDK2 induce distinct cell cycle re-entry programs in differentiated muscle cells. PLoS Genet 7**:** e1002362.

Letunic, I., and P. Bork, 2018 20 years of the SMART protein domain annotation resource. Nucleic Acids Res 46**:** D493-D496.

Li, L., C. J. Stoeckert, Jr. and D. S. Roos, 2003 OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13**:** 2178-2189.

Linard, B., A. Allot, R. Schneider, C. Morel, R. Ripp *et al.*, 2015 OrthoInspector 2.0: Software and database updates. Bioinformatics 31**:** 447-448.

Linard, B., J. D. Thompson, O. Poch and O. Lecompte, 2011 OrthoInspector: comprehensive orthology analysis and visual exploration. BMC Bioinformatics 12**:** 11.

Markaki, M., and N. Tavernarakis, 2010 Modeling human diseases in *Caenorhabditis elegans*. Biotechnol J 5**:** 1261-1276.

McDonald, J. H., 2014 *Handbook of Biological Statistics*. Sparky House Publishing, Baltimore, MD. USA.

McEwan, D. L., R. L. Feinbaum, N. Stroustrup, W. Haas, A. L. Conery *et al.*, 2016 Tribbles ortholog NIPI-3 and bZIP transcription factor CEBP-1 regulate a *Caenorhabditis elegans* intestinal immune surveillance pathway. BMC Biol 14**:** 105.

McKusick, V. A., 2007 Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 80**:** 588-604.

Mi, H., A. Muruganujan and P. D. Thomas, 2013 PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 41**:** D377-386.

Moerman, D. G., and R. J. Barstead, 2008 Towards a mutation in every gene in *Caenorhabditis elegans*. Brief Funct Genomic Proteomic 7**:** 195-204.

Nordquist, S. K., S. R. Smith and J. T. Pierce, 2018 Systematic Functional Characterization of Human 21st Chromosome Orthologs in *Caenorhabditis elegans*. G3 (Bethesda) 8**:** 967-979.

O'Reilly, L. P., R. R. Knoerdel, G. A. Silverman and S. C. Pak, 2016 High-Throughput, Liquid-Based Genome-Wide RNAi Screening in *C. elegans*. Methods Mol Biol 1470**:** 151-162.

Oliveros, J. C., 2007-2015 Venny. An interactive tool for comparing lists with Venn's diagrams.

Pereira, C., A. Denise and O. Lespinet, 2014 A meta-approach for improving the prediction and the functional annotation of ortholog groups. BMC Genomics 15 Suppl 6**:** S16.

Petcherski, A. G., and J. Kimble, 2000 LAG-3 is a putative transcriptional activator in the *C. elegans* Notch pathway. Nature 405**:** 364-368.

Pryszcz, L. P., J. Huerta-Cepas and T. Gabaldon, 2011 MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res 39**:** e32.

Rawat, V. P., N. Arseni, F. Ahmed, M. A. Mulaw, S. Thoene *et al.*, 2010 The vent-like homeobox gene VENTX promotes human myeloid differentiation and is highly expressed in acute myeloid leukemia. Proc Natl Acad Sci U S A 107**:** 16946-16951.

Remm, M., C. E. Storm and E. L. Sonnhammer, 2001 Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314**:** 1041-1052.

Roth, A. C., G. H. Gonnet and C. Dessimoz, 2008 Algorithm of OMA for large-scale orthology inference. BMC Bioinformatics 9**:** 518.

Roy, S. H., D. V. Tobin, N. Memar, E. Beltz, J. Holmen *et al.*, 2014 A complex regulatory network coordinating cell cycles during *C. elegans* development is revealed by a genome-wide RNAi screen. G3 (Bethesda) 4**:** 795-804.

Shaye, D. D., and I. Greenwald, 2011 OrthoList: a compendium of *C. elegans* genes with human orthologs. PLoS ONE 6**:** e20085.

Sherwood, D. R., J. A. Butler, J. M. Kramer and P. W. Sternberg, 2005 FOS-1 promotes basement-membrane removal during anchor-cell invasion in *C. elegans*. Cell 121**:** 951-962.

Sin, O., H. Michels and E. A. Nollen, 2014 Genetic screens in *Caenorhabditis elegans* models for neurodegenerative diseases. Biochim Biophys Acta 1842**:** 1951-1959.

Sonnhammer, E. L., and G. Ostlund, 2015 InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res 43**:** D234-239.

Sundaram, M. V., 2013 Canonical RTK-Ras-ERK signaling and related alternative pathways. WormBook**:** 1-38.

Sutphin, G. L., J. M. Mahoney, K. Sheppard, D. O. Walton and R. Korstanje, 2016 WORMHOLE: Novel Least Diverged Ortholog Prediction through Machine Learning. PLoS Comput Biol 12**:** e1005182.

Thomas, P. D., V. Wood, C. J. Mungall, S. E. Lewis, J. A. Blake *et al.*, 2012 On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. PLoS Comput Biol 8**:** e1002386.

Thompson, O., M. Edgley, P. Strasbourger, S. Flibotte, B. Ewing *et al.*, 2013 The million mutation project: a new approach to genetics in *Caenorhabditis elegans*. Genome Res 23**:** 1749-1762.

Timmons, L., and A. Fire, 1998 Specific interference by ingested dsRNA. Nature 395**:** 854.

Tucci, M. L., A. J. Harrington, G. A. Caldwell and K. A. Caldwell, 2011 Modeling dopamine neuron degeneration in *Caenorhabditis elegans*. Methods Mol Biol 793**:** 129-148.

Vahdati Nia, B., C. Kang, M. G. Tran, D. Lee and S. Murakami, 2017 Meta Analysis of Human AlzGene Database: Benefits and Limitations of Using *C. elegans* for the Study of Alzheimer's Disease and Co-morbid Conditions. Front Genet 8**:** 55.

van der Bliek, A. M., M. M. Sedensky and P. G. Morgan, 2017 Cell Biology of the Mitochondrion. Genetics 207**:** 843-871.

Vilella, A. J., J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin *et al.*, 2009 EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19**:** 327-335.

Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese *et al.*, 2007 Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35**:** D5-12.

Wu, L., J. C. Aster, S. C. Blacklow, R. Lake, S. Artavanis-Tsakonas *et al.*, 2000 MAML1, a human homologue of Drosophila mastermind, is a transcriptional co-activator for NOTCH receptors. Nat Genet 26**:** 484-489.

Wu, X., H. Gao, R. Bleday and Z. Zhu, 2014 Homeobox transcription factor VentX regulates differentiation and maturation of human dendritic cells. J Biol Chem 289**:** 14633-14643.

Wu, X., H. Gao, W. Ke, R. W. Giese and Z. Zhu, 2011 The homeobox transcription factor VentX controls human macrophage terminal differentiation and proinflammatory activation. J Clin Invest 121**:** 2599-2613.

Yan, D., Z. Wu, A. D. Chisholm and Y. Jin, 2009 The DLK-1 kinase promotes mRNA stability and local translation in *C. elegans* synapses and axon regeneration. Cell 138**:** 1005-1018.

Tables:

Program	Version in OL1 (date)	Version(s) in OL2 (dates)	# *C. elegans* genes in OL1	# *C. elegans* genes in OL2 (% change)	# Human ENSG IDs in OL1	# Human ENSG IDs in OL2 (% change)*
Ensembl Compara	v57 (2010)	v87-89 (2016-2017)	6,404	6,801 (+6.2%)	8,642	9,186 (+6.3%)
HomoloGene	v64 (2009)	v68 (2014)	4,127	3,778 (-8.5%)	2,956	3,205 (+8.4%)
InParanoid	v7 (2009)	v8 (2013)	5,591	5,581 (-0.2%)	7,527	8,949 (+18.9%)
OrthoMCL	v4 (2010)	v5 (2011)	5,663	5,699 (+0.6%)	7,417	7,588 (+2.3%)
OMA	NA	1 (2016)	NA	3,882	NA	4,558
OrthoInspector	NA	2 (2015)	NA	5,361	NA	7,771

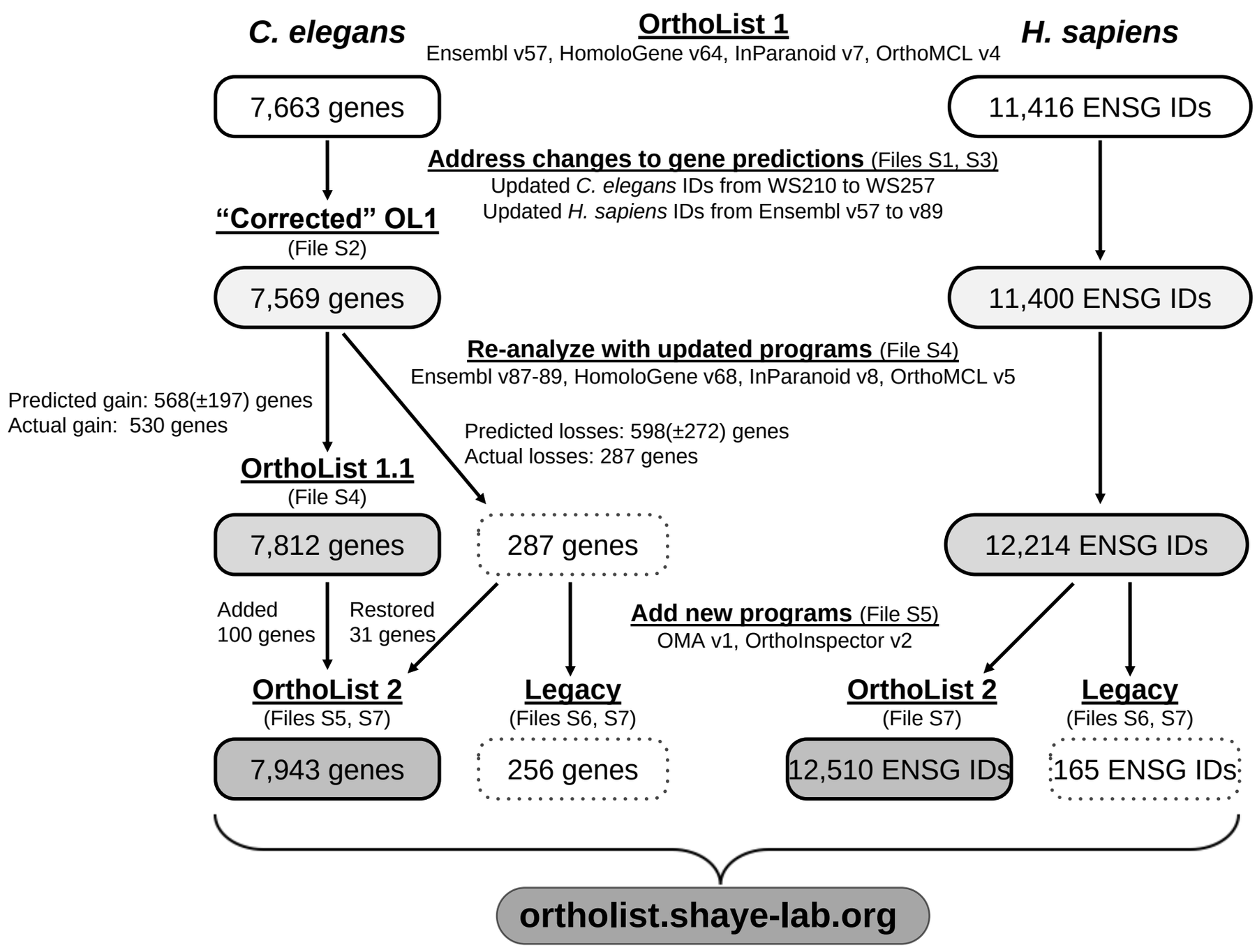
Table 1: Databases used to build OrthoList 2. The programs used here all scored highly in a recent assessment of orthology-prediction methods (ALTENHOFF *et al.* 2016). For the four previously-used programs, we report the net change (%) in *C. elegans* and human genes predicted to be orthologs between versions. (For the other two these measurements are not applicable, NA.). *The change in human ENSG ID numbers upon updates includes those whose original IDs were retired, but which still exist in the Ensembl database with a new, unlinked, ID. This deficiency in annotation makes it impossible to assess the true extent of gains and losses in the human gene set (see main text).

	Gene Numbers (net change)			Gene Content (actual genes in results)			
	Original	**Updated**	**% Change**	**# Lost**	**# Gained**	**% Lost**	**% Gained**
Ensembl Compara	6,404	6,801	+6.2%	467	864	-7.3%	+13.5%
HomoloGene	4,127	3,776	-8.5%	747	396	-18.1%	+9.6%
InParanoid	5,591	5,581	-0.2%	290	280	-5.2%	+5.0%
OrthoMCL	5,663	5,699	+0.6%	57	93	-1.0%	+1.6%
		mean	-0.5%		mean	-7.9%	+7.4%
		s.e.m	±3.0%		s.e.m	±3.6%	±2.6%

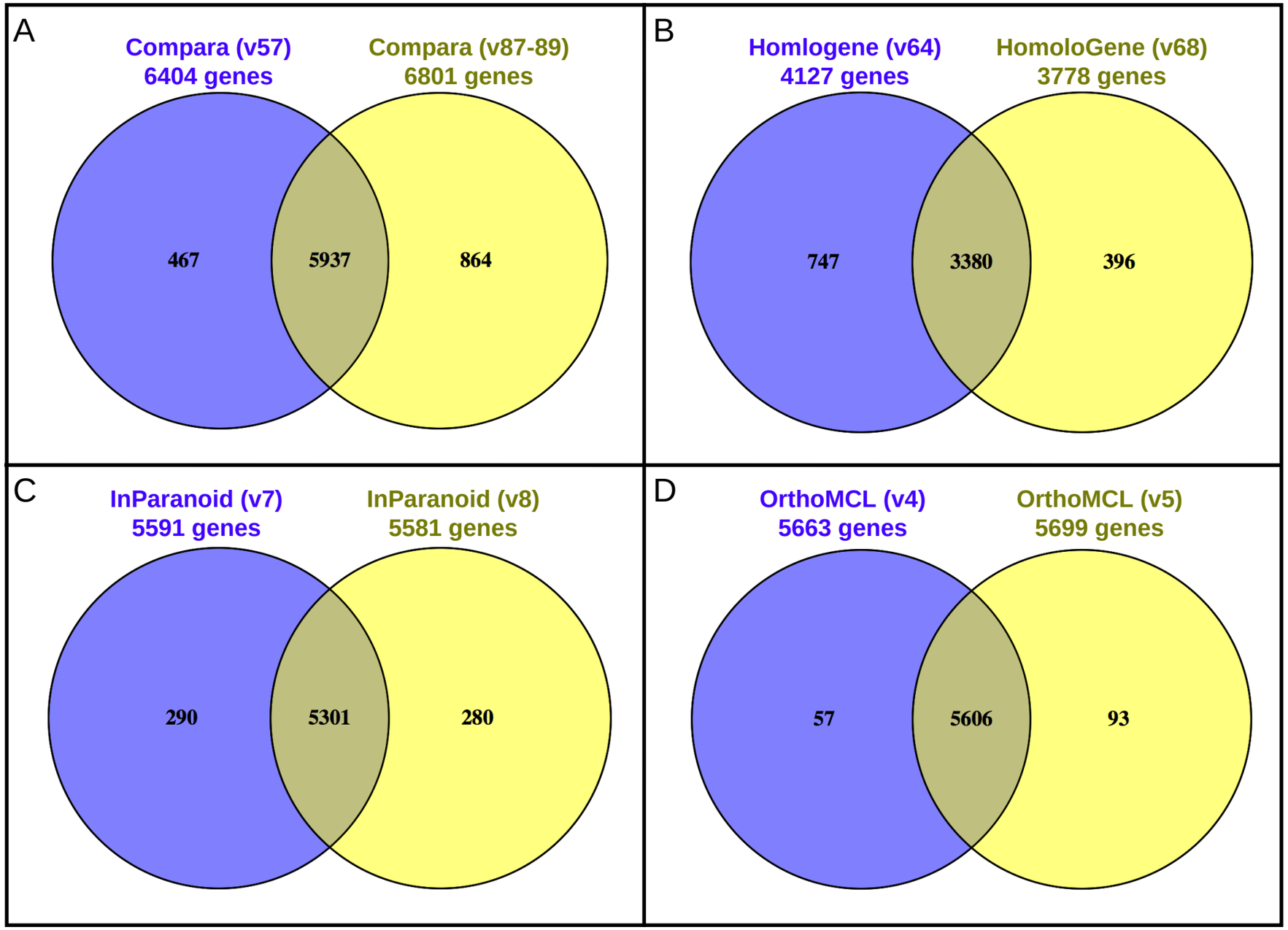
Table 2: Changes in gene number and content after updates to orthology-prediction methods. The mean change in total number of worm genes with human orthologs predicted by each individual program was quite low (-0.5±3.0%) after updates, although each program showed distinct patterns of change, with Ensembl Compara adding more genes vs. all the other programs losing genes. However, when considering the change in actual gene content, each program appears to have larger changes than what is apparent by just looking at the net change in numbers.

Class	Type of support	# of Genes	% of class	Representation with respect to proportion in OL1 (significance)	Total genes in class	% of OL1
Stayed the same	Unique	1164	21.2%	unchanged (p=0.4340)	5487	72.5%
	Two	589	10.7%	underrepresented (p<0.001)		
	Three	882	16.1%	underrepresented (p<0.001)		
	Four	2852	52.0%	overrepresented (p<0.001)		
Lost support	Unique to Lost	260	20.2%	unchanged (p=0.2205)	1285	17.0%
	Two to Unique	184	14.3%	overrepresented (p=0.0034)		
	Two to Lost	27	2.1%			
	Three to Two	253	19.7%	unchanged (p=0.2034)		
	Three to Unique	26	2.0%			
	Three to Lost	0	0.0%			
	Four to Three	492	38.3%	underrepresented (p=0.0406)		
	Four to Two	38	3.0%			
	Four to Unique	5	0.4%			
	Four to Lost	0	0.0%			
Gained support	Unique to Two	200	25.1%	overrepresented (p<0.001)	797	10.5%
	Unique to Three	23	2.9%			
	Unique to Four	3	0.4%			
	Two to Three	176	22.1%	overrepresented (p<0.001)		
	Two to Four	33	4.1%			
	Three to Four	362	45.4%	overrepresented (p<0.001)		

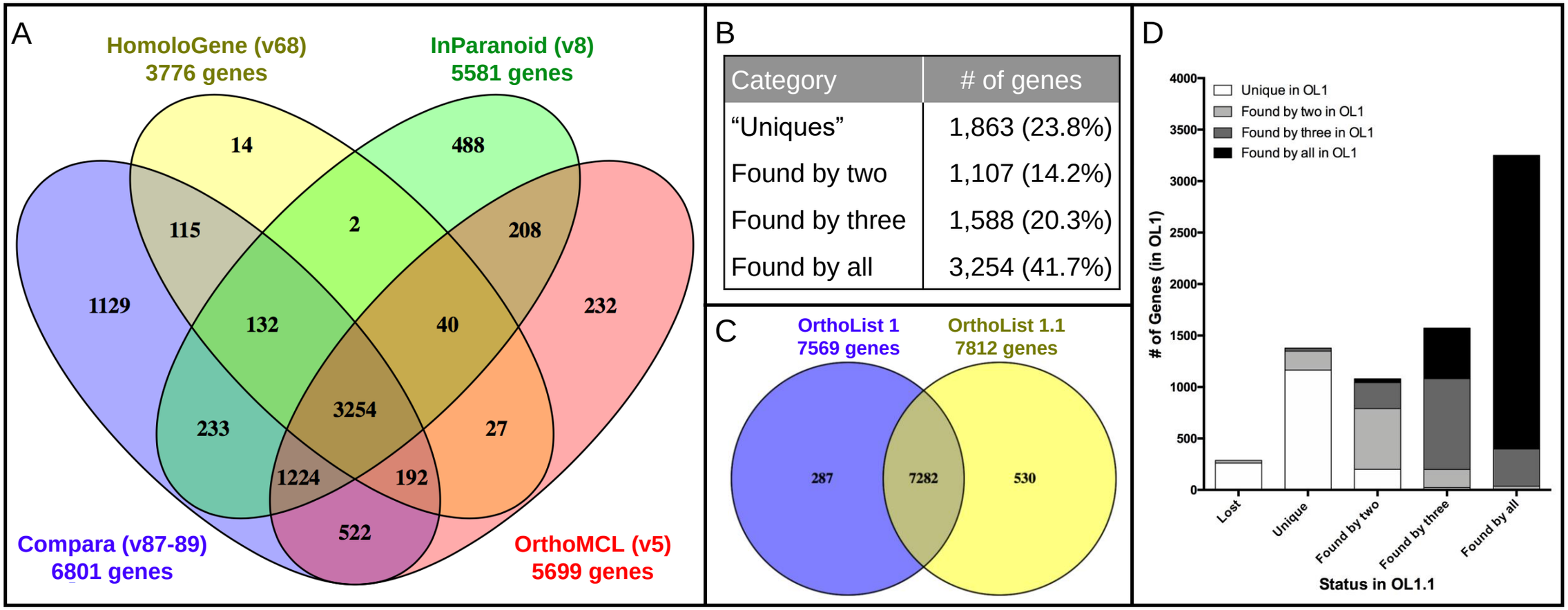
Table 3: Changes in support after updates to orthology prediction methods. All statistics in this table are calculated by a two-tailed chi-square with Yates correction. The majority of genes (72.5%) from OL1 retained the same level of support, but significantly more lost support than gained it after updates (p>0.001). To ask if there was a trend towards stability based on degree of support, we looked at whether genes supported by more programs in OL1 were overrepresented in the class that retained, or gained, support, or whether they were underrepresented in the class of genes that lost support. Conversely, we looked for whether genes supported by fewer programs were overrepresented in the class of genes that lost support. We did not find strong evidence for such a trend. The proportion of “uniques” within the class that retained the same level of support, or lost it, was not significantly different from the proportion of “uniques” in OL1. Moreover, “uniques” were overrepresented in the class that gained support. Therefore, being a unique is not a predictor for remaining unique or losing support. We also noticed that genes supported by two programs were as likely to lose support as they were to gain it (overrepresented in both classes), while genes supported by three or four programs are less likely to lose support upon updates.



Kim et al. - Fig. 1

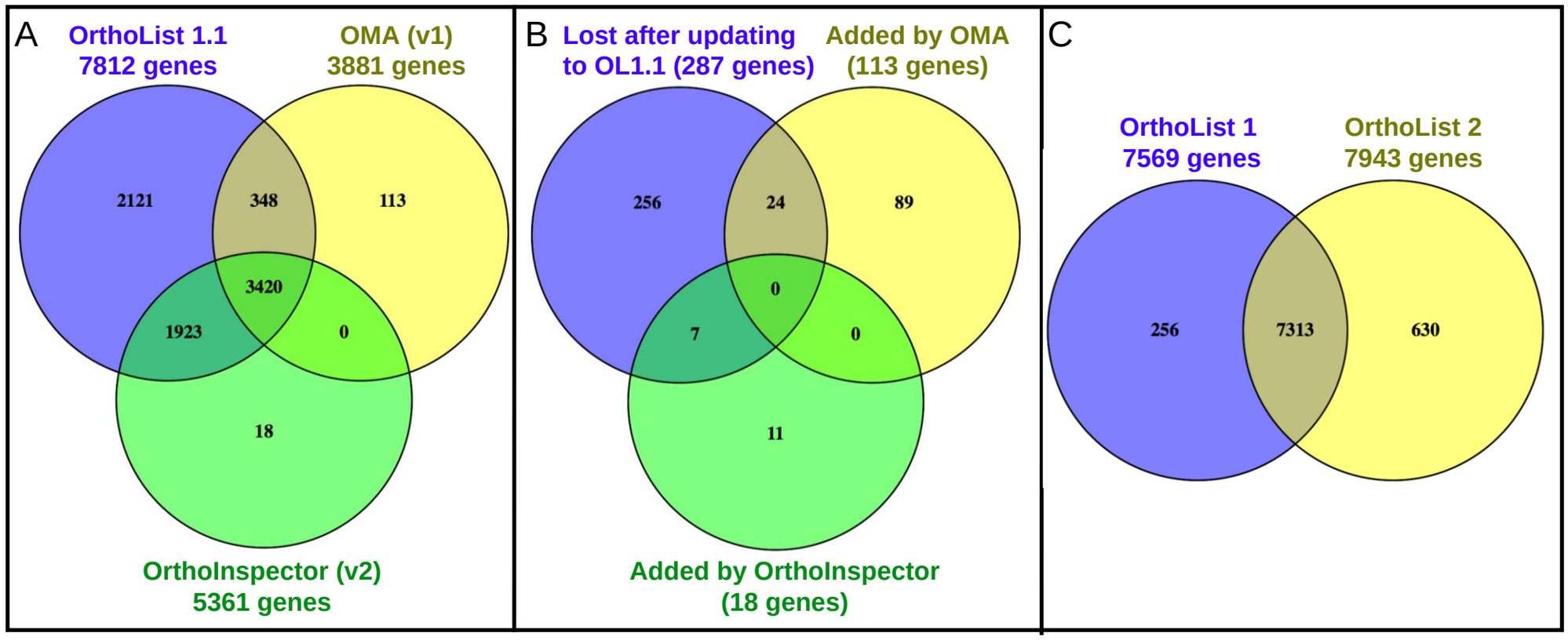


Kim et al. - Fig. 2



Category	# of genes
“Uniques”	1,863 (23.8%)
Found by two	1,107 (14.2%)
Found by three	1,588 (20.3%)
Found by all	3,254 (41.7%)

Kim et al. - Fig. 3



Kim et al. - Fig. 4

A

Ortholist 2

let-60

Fields searched:
- WormBaseID
- Common Name
- Locus ID
- Ensembl ID
- HGNC Symbol
- SMART IDs
- InterPro Domains
- OMIM Phenotypes

No. of Programs: No minimum (default)

Partial match allowed: No / Yes

Submit

B

Search in results

WormBase ID	Common Name	Locus ID	Ahringer RNAi Clone Location	No. of Programs	Ensembl ID	HGNC Symbol	SMART IDs (toggle)	InterPro Domains (toggle)	OMIM Phenotypes (toggle)
WBGene00002335	*let-60*	ZK792.6	IV-6A16	6 (?)	ENSG00000133703	KRAS	View	View	View
WBGene00002335	*let-60*	ZK792.6	IV-6A16	5 (?)	ENSG00000213281	NRAS	View	View	View
WBGene00002335	*let-60*	ZK792.6	IV-6A16	4 (?)	ENSG00000174775	HRAS	View	View	View
WBGene00002335	*let-60*	ZK792.6	IV-6A16	3 (?)	ENSG00000276536	HRAS	View	View	
WBGene00002335	*let-60*	ZK792.6	IV-6A16	1 (?)	ENSG00000133818	RRAS2	View	View	View
WBGene00002335	*let-60*	ZK792.6	IV-6A16	1 (?)	ENSG00000187682	ERAS	View	View	

Export to CSV

You searched for the following term(s): *let-60*

Kim et al. - Fig. 5