PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Transcriptomic congruence analysis for evaluating model organisms

Wei Zong[a], Tanbin Rahman[b], Li Zhu[a], Xiangrui Zeng[c], Yingjin Zhang[a], Jian Zou[a], Song Liu[d], Zhao Ren[e], Jingyi Jessica Li[f], Etienne Sibille[g], Adrian V. Lee[h,i], Steffi Oesterreich[h,i], Tianzhou Ma[j,1], and George C. Tseng[a,k,l,1]

Edited by Anton Berns, Antoni van Leeuwenhoek Nederlands Kanker Instituut, Amsterdam, Netherlands; received February 15, 2022; accepted November 17, 2022

Model organisms are instrumental substitutes for human studies to expedite basic, translational, and clinical research. Despite their indispensable role in mechanistic investigation and drug development, molecular congruence of animal models to humans has long been questioned and debated. Little effort has been made for an objective quantification and mechanistic exploration of a model organism's resemblance to humans in terms of molecular response under disease or drug treatment. We hereby propose a framework, namely Congruence Analysis for Model Organisms (CAMO), for transcriptomic response analysis by developing threshold-free differential expression analysis, quantitative concordance/discordance scores incorporating data variabilities, pathway-centric downstream investigation, knowledge retrieval by text mining, and topological gene module detection for hypothesis generation. Instead of a genome-wide vague and dichotomous answer of "poorly" or "greatly" mimicking humans, CAMO assists researchers to numerically quantify congruence, to dissect true cross-species differences from unwanted biological or cohort variabilities, and to visually identify molecular mechanisms and pathway subnetworks that are best or least mimicked by model organisms, which altogether provides foundations for hypothesis generation and subsequent translational decisions.

model organism | molecular congruence analysis | transcriptome | translational research

Significance

As human clinical studies are often expensive, lengthy, and with many constraints, model organisms, such as mouse and rat, play an indispensable role in almost all disease domains. Although instrumental and popular, the application of model organisms has raised caution. Two previous PNAS reports presented controversial conclusions of mouse model's resemblance to humans in inflammatory transcriptomic responses, which triggered debates on its usefulness. To date, no objective and quantitative tools are available to describe the congruence of a mouse model to humans. The proposed methodology in this paper fills this gap to facilitate mechanistic understanding and hypothesis generation when evaluating an animal model.

As human studies often encounter numerous constraints, including larger biological heterogeneity, hidden confounding factors, greater cost and time, and potential ethical concerns, model organisms have played an indispensable role in preclinical research to understand the pathogenesis of human diseases at the behavioral, cellular, and molecular levels. Their clinical validity and translational values are, however, long debated with controversial opinions (1–4). A notable example is the opposite conclusions from two articles analyzing an identical transcriptomic response dataset in human and mouse inflammation (5, 6), with the former concluding that mouse models (MMs) "poorly" mimic humans while the latter reporting "greatly" mimicking. The contradictory results triggered further debates of merits and limitations of animal models (7–9). To date, efforts have been made to compare or predict model organism responses using association analysis (5, 6), machine learning (10, 11), pathway enrichment (12, 13), or meta-analysis (14) approaches. An objective and quantitative approach to identify biomarkers, pathways, and topological gene regulatory modules that are best or least mimicked by the model organism is, however, still lacking. Our research aims to fill this gap and to facilitate data-driven mechanistic understanding, hypothesis generation, and translational guidance of animal models.

Fig. 1 presents an overview of the Congruence Analysis for Model Organisms (CAMO) pipeline, consisting of state-of-the-art methods and approaches for a thorough congruence evaluation. In Fig. 1*A*, differential analyses contrasting case and control groups (e.g., disease vs. healthy or treated vs. un-treated) are first performed in mouse and human cohorts separately. Threshold-free Bayesian differential analysis is implemented to transform *P* values from conventional pipelines (e.g., Linear Model for Microarray Data (LIMMA) or Differential expression analysis based on the Negative Binomial (DESeq2)) to differential posterior probabilities, with which cross-species concordance/discordance scores (abbreviated as c-scores/d-scores hereafter) are calculated by the F-measure concept in machine learning with the associated inference of *P* values. Pathway-specific c-scores/d-scores are calculated similarly by constraining genes to a specific pathway. When multiple cohorts are jointly analyzed, c-scores and d-scores are calculated for all pairs of studies in each individual pathway provided by users or pathway databases. Fig. 1*B* heatmap illustrates an imaginary example of two human and two mouse studies, labeled as H_1, H_2, M_1, and M_2, where H_2, M_1, and M_2 are more congruent to each other while H_1 is dissimilar to the other three cohorts for the first pathway shown.

Author contributions: T.M. and G.C.T. designed research; W.Z., T.R., L.Z., X.Z., Y.Z., J.Z., S.L., T.M., and G.C.T. performed research; J.J.L. contributed new reagents/analytic tools; W.Z., T.R., Y.Z., J.Z., S.L., and T.M. analyzed data; L.Z. wrote software; Y.Z. wrote software package; and W.Z., Z.R., J.J.L., E.S., A.V.L., S.O., T.M., and G.C.T. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: tma0929@umd.edu or ctseng@pitt.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2202584120/-/DCSupplemental.

Published February 2, 2023.

After c-score/d-score calculation, the "Mechanistic investigation and hypothesis generation" component in Fig. 1*B* can perform "pathway knowledge retrieval" and "topological gene module detection". In "pathway knowledge retrieval", an unsupervised clustering method is applied to cluster enriched pathways with similar concordance patterns across studies to extract and simplify highly overlapped and redundant information of pathways annotated from different database sources, such as Gene Ontology (GO) (15), Kyoto Encyclopedia of Genes and Genomes (KEGG) (16), and Reactome (17). A text mining algorithm is then applied to retrieve statistically enriched keywords in each pathway cluster. Finally, in "topological gene module detection", a community detection algorithm is developed for any selected pair of models to identify concordant or discordant subnetworks in a selected pathway based on its topological regulatory information (e.g., see Fig.3). For a pathway topological plot of two selected models, it is possible to detect both a cross-species concordant subnetwork and a discordant subnetwork. The resulting concordant/discordant subnetworks together with retrieved pathway knowledge provide an objective and disciplined basis for mechanistic understanding of cross-species congruence and for further hypothesis generation.

To allow application by other researchers, we develop a user-friendly R-shiny app (https://github.com/CAMO-R/Rshiny) to interactively implement the streamlined workflow. The results can then be visualized and investigated in pathway clusters, individual pathway, or subnetwork modules inside pathways for interactive exploration. The Bayesian differential analysis tool allows flexible input of processed expression data or precalculated *P* values from conventional differential analysis tools. Ortholog genes can be automatically mapped for Mus musculus, Rattus norvegicus, *Drosophila melanogaster*, and *Caenorhabditis elegans*. Popular pathway databases, such as GO (15), KEGG (16), Reactome (17), and BioCarta (18), are included in the package, where KEGG (16) and Reactome (17) contain topological regulatory information for visualization.

Results

Below, we demonstrate the CAMO framework using two real examples. We first revisit the previously reported controversial example of human vs. mouse inflammatory disease models (5, 6). Our CAMO result reconciles the two dichotomized and subjective conclusions of "greatly mimicking" or "poorly mimicking" by numerically quantifying the concordance and discordance of human–MM comparisons at genome-wide, pathway, or gene subnetwork module level. In the second example, we apply CAMO to compare developmental stages from embryo to adult in two model organisms, worm (*C. elegans*) and fruit fly (*D. melanogaster*) in the modENCODE project (19, 20). Results of the two case studies show flexibility and extensibility of CAMO for simultaneously comparing multiple models (i.e., six MMs and six human models in case study 1) or two models in multiple developmental stages (i.e., five developmental stages in *C. elegans* and five in *D. melanogaster* in case study 2).

Case Study 1: Congruence of Inflammatory MMs and Human. Two papers published in PNAS reached contradicting conclusions on MM resemblance to humans in transcriptomic response of inflammatory diseases (5, 6). In an earlier paper, Seok et al. (5)

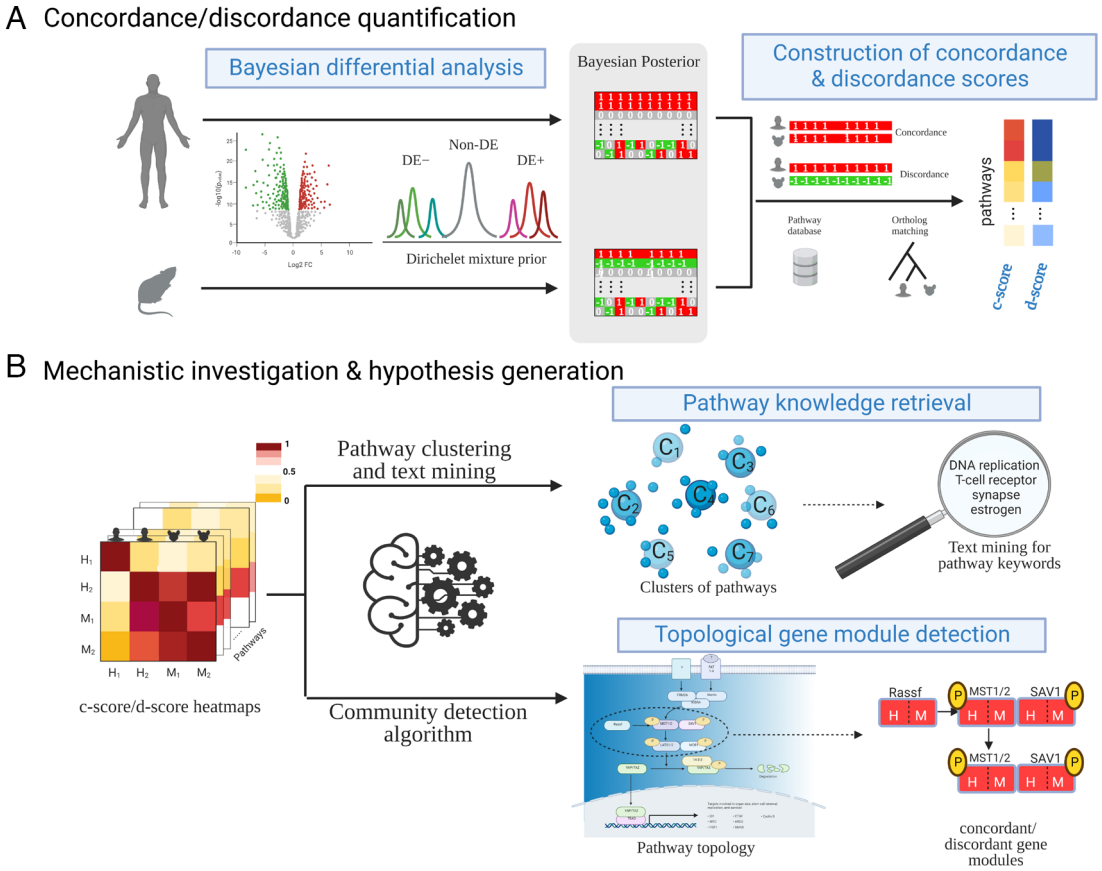

Fig. 1. Workflow of the "CAMO" framework. (*A*) Procedures to calculate genome-wide and pathway level c-scores and d-scores for a pair of human sepsis (HS) and MM. (*B*) Downstream machine learning and bioinformatics interactive visualization tools for pathway knowledge retrieval and topological gene nodule detection.

analyzed microarray studies to investigate gene expression changes in human inflammatory diseases and the corresponding MMs and found a "poor" correlation between the genome-wide expression changes in humans and in mice. Takao et al. (6) later reanalyzed the same dataset and concluded that the transcriptomic changes in MMs greatly mimicked those in humans. A series of comments and debates followed to discuss potential differences in cell-type composition, time frame, and genetic variation between mice and humans, but no converging message or practical guidance has been formed (7–9). In *SI Appendix*, Table S1, we compare the analytical procedures in the two papers and identify multiple major differences that may have contributed to the contradicting conclusions: a) The two papers select different sets of differential expression (DE) genes to calculate the correlation (i.e., using DE genes in humans only in ref. 5 but using intersected DE genes in both humans and mice in ref. 6), b) They use different statistical significance (q value or *P* value) and biological significance (effect size) thresholds for determining DE genes, c) They report different correlation measure ($R^2 \approx (\text{Pearson correlation})^2$ or Spearman correlation), d) Their sources of data and preprocessing steps are different and with no reproducible code. Observing the risk of ad hoc and potentially subjective analytical decisions in these cross-species congruence analyses, we are motivated to develop a threshold-free and rigorous statistical framework in CAMO. In contrast to genome-wide correlation assessment in these papers, we extend the investigation into pathways and gene regulatory modules for insightful mechanistic understanding.

SI Appendix, Table S2 lists six inflammatory response studies in humans (Burns, Infection, Trauma, Sepsis, lipopolysaccharide endotoxins, and acute respiratory distress syndrome; abbreviated as HB, HI, HT, HS, HL, and HA, respectively) and six corresponding conditions in mice (abbreviated as MB, MI, MT, MS, ML, and MA, respectively), where data are previously curated from Gene Expression Omnibus repository in the KERIS package (21). The 12 microarray datasets in Affymetrix and Illumina platforms are preprocessed and normalized as uniformly as possible (see *Methods* section). Since inflammatory diseases progress over time and differ in humans and mice, the selection of matched time points across species is critical. We evaluate the cross-species time series and select the best-matched time points by c-score in case-control transcriptomic response (*Methods*). After cross-species gene matching, 8,317 ortholog genes are remained for CAMO analysis. For any pair of the 12 studies, genome-wide and pathway-specific c-scores and d-scores are calculated. The resulting genome-wide c-scores and d-scores of each pair of studies are shown in *SI Appendix*, Tables S4 and S5 respectively. The congruence visualization by multidimensional scaling (MDS) plot, a visual representation of dissimilarity structure, using transformation of c-scores as dissimilarity measure is shown in *SI Appendix*, Fig. S1. We find that four human inflammatory studies HB, HI, HT, and HS resemble each other well with genome-wide c-scores ranging from 0.25 to 0.52 (shaded dark green in *SI Appendix*, Table S4), consistent with previous findings in the two PNAS papers, while no resemblance evidence is suggested for HL and HA (c-scores ≈ 0; shaded light green, *SI Appendix*, Table S4). MB and MI are overall more similar to the four human studies (HB, HI, HT, and HS) in a weaker congruence level (c-scores = 0.081 to 0.20; shaded pink, *SI Appendix*, Table S4) while MT, MS, MA, and ML have almost no genome-wide congruence, implying the concordance of cross-species transcriptomic response in inflammatory diseases is condition specific. Interestingly, mouse studies are not necessarily more similar to human studies of the same inflammatory condition. For example, c-scores of MI-HI and MB-HB are 0.2 and 0.11 (marked red, *SI Appendix*, Table S4), while c-scores of the other four pairs, MT-HT, MS-HS, ML-HL, and MA-HA, are almost 0 (marked blue, *SI Appendix*, Table S4). Unlike most of the human studies, the six mouse studies generally do not mimic each other, implying unknown complexity and high variability of MMs in inflammatory diseases.

We next apply consensus tight clustering to 219 selected pathways with enriched meta-analyzed DE genes (*Methods*) to reduce redundancy of highly overlapped pathways and to improve interpretation. Four pathway clusters are identified with 41 scattered (i.e., unclustered) pathways (*SI Appendix*, Table S6), where the number of clusters is selected by the consensus CDF and scree plot (*SI Appendix*, Fig. S2). Heatmap and MDS plot of pathway clusters are shown in *SI Appendix*, Fig. S3 *A* and *B*. The comembership heatmaps are used to summarize the proportion of significantly concordant pathways within each pathway cluster between each pair of studies (*SI Appendix*, Fig. S3*C*). Through text mining of pathway names and descriptions, top significantly enriched keywords are shown (*SI Appendix*, Fig. S3*C* and Table S7). These results suggest molecular mechanisms underlying the commonalities and differences in various inflammatory response models between the two species, providing insights beyond the subjective and dichotomous conclusions of the previous two contradicting PNAS papers. For example, pathway Cluster I and II (*SI Appendix*, Fig. S3*C*) show that several MMs (e.g., MB, MI, and MT in Cluster I) mimic most human studies (e.g., HB, HS, HT, and HI in Cluster I) well in both innate (e.g., natural killer cell related) and adaptive (e.g., T cell related; *SI Appendix*, Table S7) immunity. Despite the difference in neutrophil and lymphocyte abundance and other phenotypes, it has been reported that the overall immune system is relatively similar between humans and mice (22). Pathway Cluster IV shows that MMs do not mimic human studies in ribosome and protein translation while responses in HS, HB, HT, and HI are similar to each other. Such findings agree with earlier studies that profound cross-species differences exist in translation machinery (23).

CAMO next provides interactive exploration in the shiny app to select pathways and facilitate further topological visualization of regulatory networks. To demonstrate the idea, Fig. 2 contains a selected display of burn and sepsis studies in humans and mice (HB, HS, MB, and MS). Fig. 2*A* shows DE evidence with concordance (the upper right region) and discordance (the lower left region) information in each pair of model comparison, where each dot represents a pathway, X-axis and Y-axis represent the aggregated DE evidence (average posterior probability of DE) of a pathway, and color intensity of the dots refers to statistical significance (minus-log-transformed *P* values) of c-scores or d-scores in each pathway. In the shiny app, a user can click on any location of a plot to obtain information of the nearby pathways. After interactively exploring individual pathways of interest, a pathway with high DE evidence in both axes and strong congruence/discordance can be selected for scrutinizing its gene-specific congruence information and mechanistic investigation. In this demonstration, two KEGG pathways (hsa04760 and hsa04662) with strong DE evidence and high c-score or d-score statistical significance are identified and further investigated. Fig. 2*B* shows the signed DE posterior probability (red for upregulation and green for downregulation) of each study (on columns) for selected DE genes (on rows). The "Leukocyte transendothelial migration" pathway (hsa04670) has high DE evidence with average DE posterior probabilities = 0.42 and 0.35 in HS and MS in Fig. 2*A*. Intriguingly, the pathway exhibits discordant DE signals in 25 genes in the HS vs. MS comparison (marked blue in Fig. 2*B*; upregulated in HS and downregulated in MS or vice versa) and only one concordant gene (marked orange in Fig. 2*B*). Fig. 3*A*

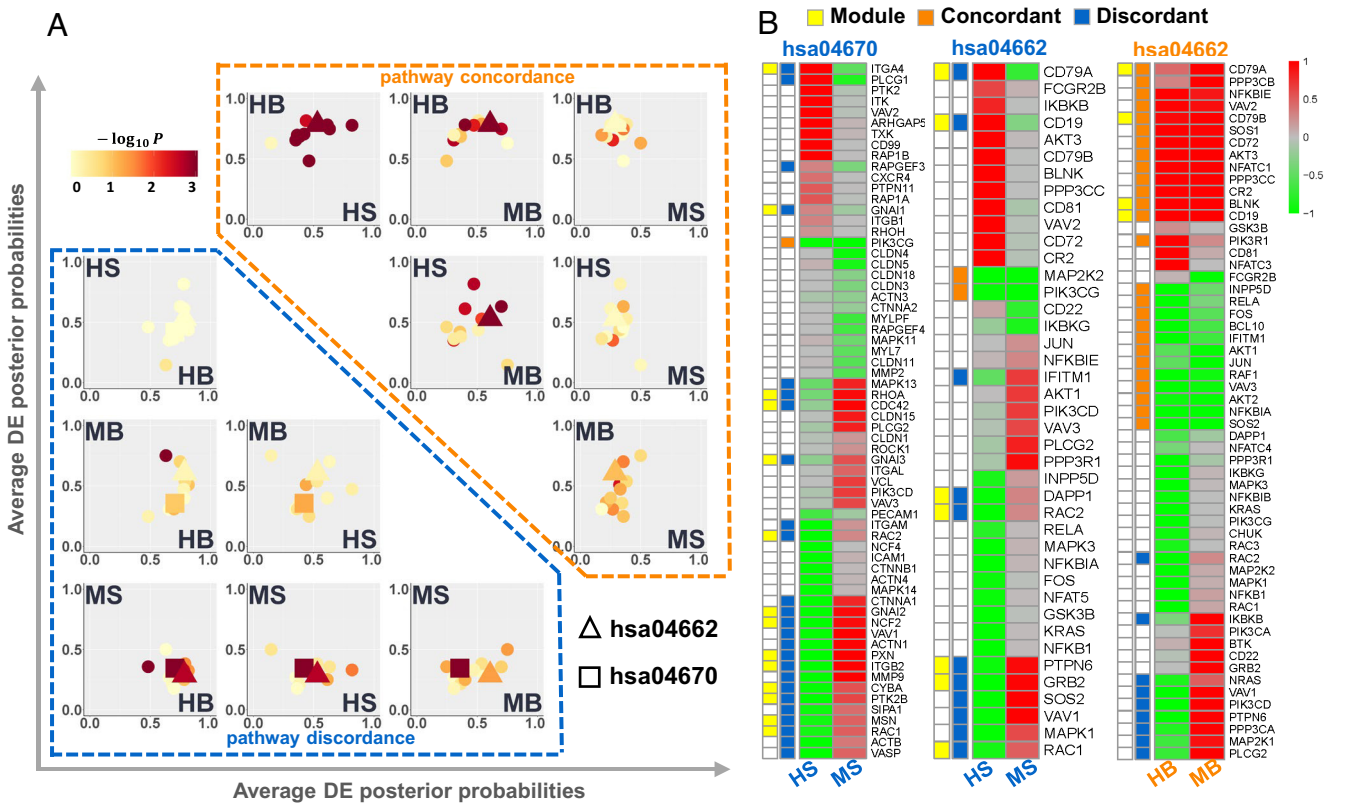

Fig. 2. Results from the case study 1. (*A*). Summary plot of DE evidence and pathway level concordance (orange in the *Upper Right* region) and discordance (blue in *Lower Left* region). Each ○/△/□ is a pathway. X- and Y-axes represent the average DE posterior probabilities, and color represents the magnitudes of $-\log_{10}$ transformed *P* value of c-scores or d-scores. Two example pathways are highlighted using different shapes ("△": hsa04662—KEGG: B cell receptor signaling pathway; "□": hsa04670—KEGG: leukocyte transendothelial migration). Ten additional randomly selected pathways are shown as comparison. Pathways with high average DE posterior probabilities in both models (x-axis and y-axis) and with high significance of c-score or d-score (darker color) are prioritized pathways for further mechanistic investigation. (*B*). Gene-wise heatmap of posterior mean of DE indicators of the HS-MS comparison in hsa04670, HS-MS in hsa04662, and HB-MB in hsa04662. Genes identified by community detection algorithm (yellow) and genes with concordant (orange) or discordant (blue) are shown in two columns beside the heatmaps.

shows the KEGG pathway topological plot with gene–gene regulatory network information and with side-by-side display of the differential regulation signals in HS and MS (red for upregulation and green for downregulation; HS on the left and MS on the right). The community detection algorithm (*Methods*) identifies a subnetwork module containing 14 DE genes (RHOA, PTK2B, RAC2, RAC1, CDC42, ITGA4, ITGB2, MSN, PXN, NCF2, CYBA, GNAI1, GNAI2, and GNAI3) with opposite expression response directions in human-mouse comparison of sepsis (i.e., green in HS and red in MS or vice versa; $P = 0.002$ for this detected subnetwork module). The colocalized discordant module is directly related to cell motility and direct sensing (Fig. 3*A* pop-out plot), a critical function that allows leukocytes to attach to the vessel wall to initiate immune response during inflammation (24). The striking mouse–human discordant result may reflect the discrepancy in proportions of different cell types of blood leukocytes between humans and mice as pointed out in a previous critique (9).

The second pathway "B cell receptor signaling pathway" (hsa04662) exhibits high discordance between HS and MS (11 discordant genes in blue and two concordant genes in orange in Fig. 2*B*) while exhibits more concordance between HB and MB (26 concordant genes in orange and nine discordant genes in blue). From the KEGG topological plot of HS/MS comparison in Fig. 3*B*, a subnetwork module of seven discordant genes (PTPN6, DAPP1, CD79A, RAC1, RAC2, GRB2, CD19; $P = 0.009$) is detected. CD79A and CD19 are antigen receptors on the B cell membrane to regulate signaling molecules, such as GRB2 and RAC family, with important roles in the regulation of cell growth and movement (25). On the other hand, Fig. 3*C* shows the general concordance of DE signals between HB and MB (both red or both green) in the B cell membrane receptor and signaling. The community detection algorithm identifies a subnetwork module of four concordant genes: CD79A (Ig-Alpha), CD79B (Ig-Beta), CD19, and BLNK (yellow genes in Fig. 3*B*). Additionally, three genes, CD72 (coinhibitor), IFITM1 (LEU13; costimulator) and CR2 (costimulator) work together to regulate the integral membrane protein complex and show concordance between HB and MB (Fig.3*C* pop-out plot). Results in Fig. 3 *B* and *C* are consistent with previous literature showing similarity but also a significant difference between mouse and human immunology, specifically in B cell development (22).

To elucidate true cross-species differences in c-scores/d-scores and dissect from within-cohort and cross-cohort variabilities in humans, we perform two additional analyses: Intracohort heterogeneity analysis and intercohort heterogeneity analysis for human subjects. An intracohort heterogeneity analysis is performed by randomly splitting samples into two equal-size subsets (e.g., HI randomly split to HI1 and HI2). The c-scores/d-scores are calculated to evaluate the experimental noise within the cohort and to assess the "ceiling" of (i.e., the highest possible) concordance given the cohort, sample size, and experimental design in humans (see details in *SI Appendix*, section 1C and Fig. S4*A*). *SI Appendix*, Fig. S4*B* compares the intracohort c-scores with cross-species

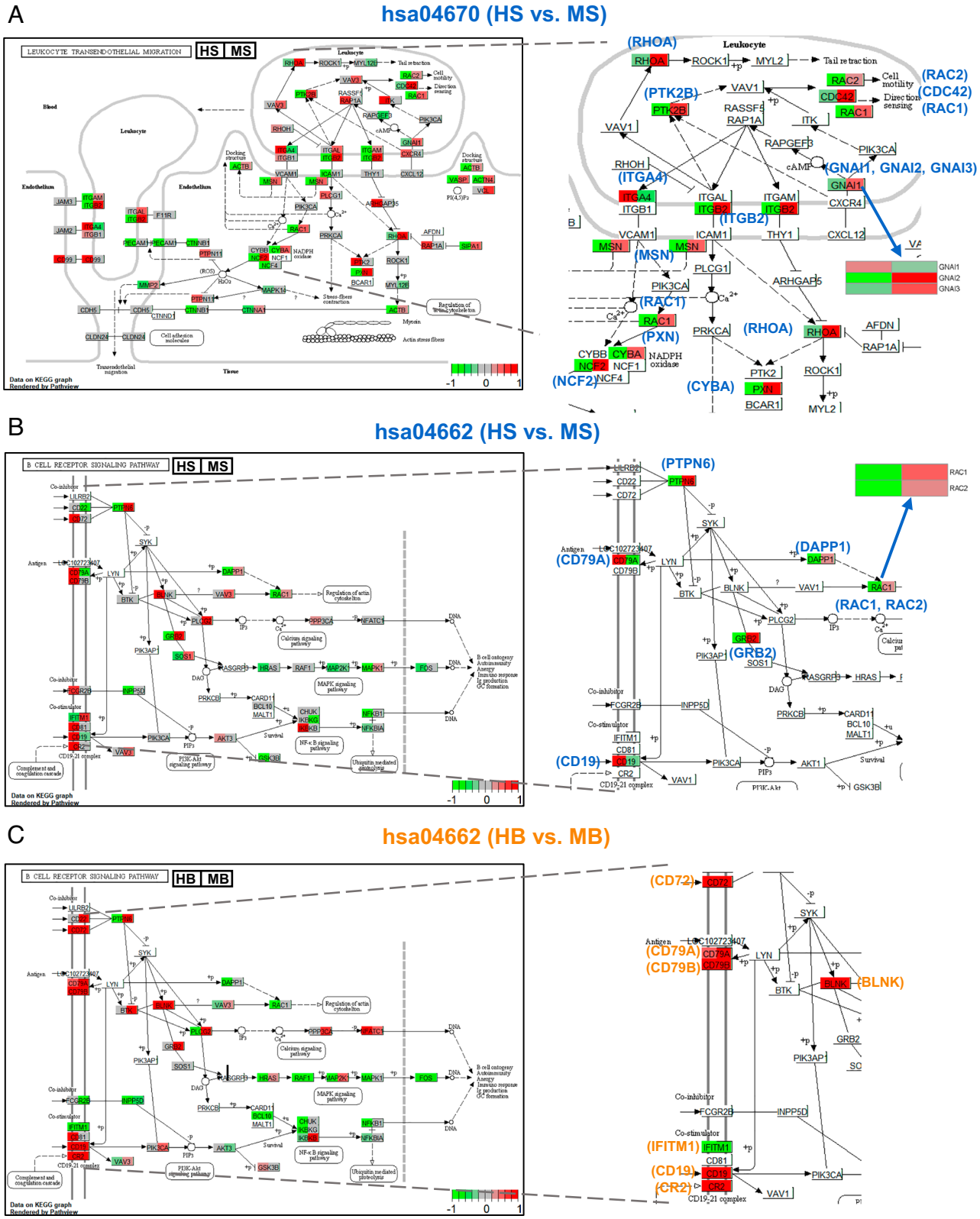

Fig. 3. Pathway topological plots of the example pathways in case study 1. (*A*) hsa04670 (HS-MS), (*B*) hsa04662 (HS-MS), and (*C*) hsa04662 (HB-MB). Pop-out plots represent the colocalized concordant/discordant modules identified from the pathway topology by the community detection algorithm. Colors in the nodes refer to the posterior mean of DE indicators in each corresponding study pair (red for upregulation and green for downregulation).

c-scores in both genome-wide and selected pathways. Higher genome-wide intracohort c-scores in HB, HI, HT, and HS (between 0.4 and 0.7) suggest their potential of meaningful cross-species evaluation while ~0 intracohort c-scores in HA and HL indicate incompetence of using these human data for cross-species evaluation. Intriguingly, although MI only shows moderate genome-wide cross-species c-score compared to intracohort c-score (median c-score = 0.15 for HI1 vs. MI vs. 0.43 for HI1 vs. HI2), its cross-species c-scores in all four selected pathways are close to the human "ceilings" in *SI Appendix*, Fig. S4*B*. For

example, the median c-score of HI1 vs. MI is 0.26 compared to 0.31 in HI1 vs. HI2 in hsa04662, showing impressively high cross-species congruence of MI in the B cell receptor signaling. We note that, in this human intracohort analysis, mouse studies (e.g., MI) has to be compared to the halved human subsampled data (i.e., HI1) to avoid sample size bias. Since the reduced sample size could significantly reduce statistical power in the downstream biological findings, we always apply CAMO to the full human dataset as the main analysis and this analysis is an auxiliary diagnostic tool to dissect cross-species differences from inherent human within-cohort variability.

Next, a human intercohort heterogeneity analysis is performed to validate the intriguing discordance findings between HS and MS using independent human datasets to factor in cross-cohort variability (see detailed analysis procedure in *SI Appendix*, section 1D and Fig. S5*A*). Four independent human sepsis studies (a. GSE26378, b. GSE26440, c. GSE4607, and d. GSE8121) with similar study designs are downloaded from the GEO repository and preprocessed similarly. In *SI Appendix*, Fig. S5*B*, HS has high genome-wide c-scores with all four independent studies and even higher in the three selected KEGG pathways, showing satisfying replicability across human cohorts. In contrast, MS has almost zero genome-wide and pathway-specific c-scores with all five human studies. The result strongly validates the lack of congruence of the MS model with human. In *SI Appendix*, Fig. S5*C*, MS shows d-scores = 0.2 to 0.4 for the three selected KEGG pathways when compared to HS and the four human validation studies. When HS is compared to the four human validation studies, the d-scores dropped to ~−0.3, which is likely an impact of high congruence (large c-scores). In summary, the result provides a strong confirmation that the discordance of the MS model with human in the three KEGG pathways are true biological signals.

Case Study 2: Congruence of Developmental Stages in *C. elegans* and *D. melanogaster*. *C. elegans* (ce) and *D. melanogaster* (dm) are effective model organisms for studying molecular, cellular, and developmental processes. Using the modENCODE RNA-seq data (19, 20), Li et al. performed comprehensive comparison in developmental time courses between the two species and provided new insights into similarities in their development. We reanalyze and preprocess 35 worm samples measured at four developmental stages (embryo, larvae, dauer, and adult) and 30 fruit fly samples measured at four developmental stages (embryo, larvae, pupae, and adult). Heatmaps of hierarchical clustering (*SI Appendix*, Fig. S6) show clear separation across the developmental stages while the embryonic stage of both species can further split into three subphases: Early embryonic phase, middle embryonic phase, and late embryonic phase. In addition, the female and male adults in fruit fly are very different from each other on the heatmap. In light of these observations and also considering the similarity between female adult fruit fly and adult worm as reported in Li et al. (20), five transcriptomic studies in each species are identified and the worm adults and fruit fly female adults are treated as the reference group in the DE analysis, which generates 10 models for molecular congruence analysis: 1) *C. elegans*: early embryo (ce.e0), mid embryo (ce.e1), late embryo (ce.e2), larvae (ce.lar), and dauer (ce.dau); 2) *D. melanogaster*: early embryo (dm.e0), mid embryo (dm.e1), late embryo (dm.e2), larvae (dm.lar), and pupae (dm.pup). After gene matching and standard preprocessing, 6,869 common ortholog genes are remained for CAMO analysis.

The MDS plot (*SI Appendix*, Fig. S7) of genome-wide c-scores for the five Drosophila and five *C. elegans* developmental stages shows a clear separation between the two species on the y-axis. The x-axis presents a developmental transition in the embryonic stages e0→e1→e2 while the lar and pup/dau stages are not exactly ordered. Adjacent developmental stages are found to be more similar to each other within species and the late embryonic stage in *C. elegans* (ce.e2) appears to somewhat resemble all stages in Drosophila except for the early embryonic stage (dm.e0). This unintuitive result is better visualized by an intriguing bipartite graph between Drosophila and *C. elegans* stages (Fig. 4*A*) by creating solid edges when the genome-wide c-score between any pair of stage is greater than 0.1. We first observe reasonable within-stage cross-species resemblance (i.e., solid yellow edges: ce.e0—dm.e0, ce.e1—dm.e1, ce.e2—dm.e2, and ce.e2—dm.e1; dashed yellow edge: ce.lar—dm.lar with a slightly lower c-score = 0.087) and then identify surprising cross-stage resemblance between species (i.e., purple edges: ce.dau—dm.e2, ce.e2—dm.lar, and ce.e2—dm.pup). Resemblance of ce.dau—dm.e2 has been suggested by the original modENCODE paper (20). Resemblance of ce.e2—dm.lar and ce.e2—dm.pup confirms the second large wave of cell proliferation and differentiation in Drosophila's life cycle. The complete c-score matrix and d-score matrix are shown in *SI Appendix*, Tables S8 and S9, respectively.

From the 269 selected pathways with enriched meta-analyzed DE genes, consensus tight clustering identifies six pathway clusters (see consensus CDF plot and scree plot in *SI Appendix*, Fig. S8; heatmap and MDS plot in *SI Appendix*, Fig. S9 *A* and *B*). Pathway Cluster III is found related to cell cycle and DNA replication with cross-species congruence in late stages (e2, lar, and pup/dau). Cluster IV is related to hormones, such as estrogen and other steroid hormones, with congruence mostly in *C. elegans* stages. Pathway Cluster II is specific to Drosophila developmental stages and contains several pathways related to RUNX family of transcription factors shown to be orchestrators of development (26) (see results of six pathway clusters in comembership heatmaps in *SI Appendix*, Fig. S9*C*; pathways memberships in *SI Appendix*, Table S10; text mining results in *SI Appendix*, Table S11). For demonstration purposes, DE evidence and c-scores/d-scores in ce.e2, ce.dau, dm.e2, and dm.pup are shown in Fig. 4*B*. Three pathways of interest with strong DE evidence and large c-scores or d-scores between the two species are further explored by genewise heatmap of posterior mean of DE (Fig. 4*C*) and pathway topology (*SI Appendix*, Fig. S10). Pathways "Homologous recombination" (KEGG: cel03440) and "Mismatch repair" (KEGG: cel03430) exhibit high concordance between ce.e2 and dm.e2 (*SI Appendix*, Fig. S10 *A* and *B*), implying similar molecular events taking place in the late embryo stage for both species. The pathway "Nucleotide-binding domain, leucine-rich repeat containing receptor (NLR) signaling pathways" (Reactome: R-CEL-168643) exhibits discordance between ce.dau and dm.pup (*SI Appendix*, Fig. S10*C*). The NOD1/2 and inflammasomes components of the pathway are both related to the innate immune system, the first line of defense against invading microorganisms that are more expressed in the pupae stage of Drosophila but not in the dauer stage of *C. elegans* (27).

Discussion

Model organisms have played critical roles in biomedical research. However, molecular congruence analysis between these models and human has been largely lacking or understudied, resulting in confusion and loss of opportunities for best use of these animal models. In this paper, we develop the CAMO pipeline for a rigorous quantification, visualization, and exploratory system to study molecular congruence of an animal model to human. We also propose intracohort and intercohort heterogeneity analyses in human to isolate experimental variabilities from true cross-species differences when the sample size in the human sepsis is sufficiently large and/or independent human studies are available for

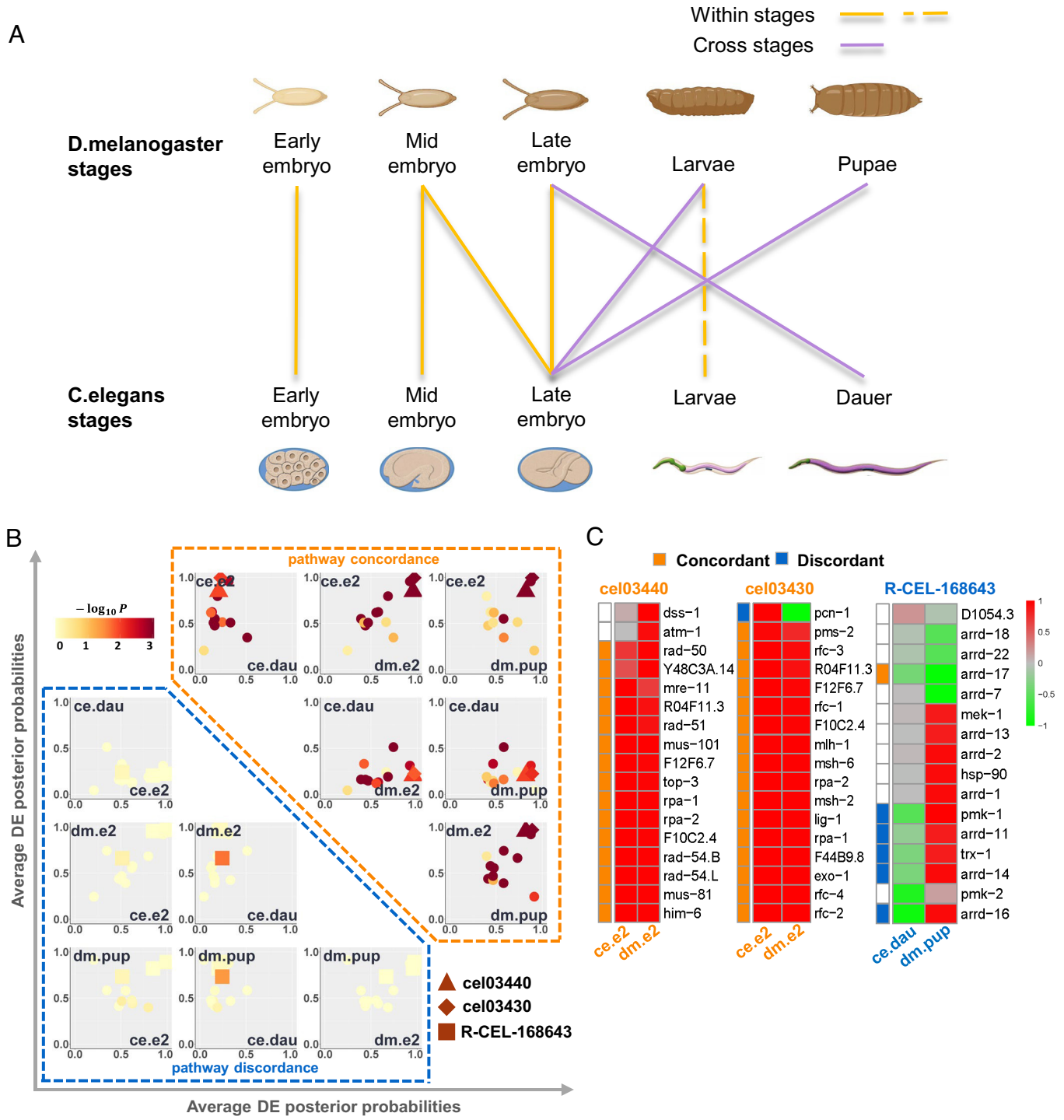

Fig. 4. Results from the case study 2. (*A*) Bipartite graph between Drosophila and *C. elegans* where solid edges are drawn when the genome-wide c/d-scores between any pair of cross-species stages are greater than 0.1. The yellow dashed line indicates a slightly weaker within-stage concordance with c-score = 0.087. (*B*) Summary plot of DE evidence and pathway level concordance (the *Upper Right* region) and discordance (the *Lower Left* region). Each ○/△/◇/□ is a pathway. X- and Y-axes represent the average DE posterior probabilities, and color represents the magnitudes of $-\log_{10}$ transformed *P* value of c-scores or d-scores. ce.e2: Late embryo stage of CE; ce.dau: Dauer stage of CE; dm.e2: Late embryo stage of DM; and dm.pup: Pupae stage of DM. Three example pathways are highlighted using different shapes ("△": cel03440—KEGG: Homologous recombination; "◇": cel03430—KEGG: DNA mismatch repair; "□": R-CEL-168643—Reactome: NLR signaling pathways). Ten additional randomly selected pathways are shown as comparisons. Pathways with high average DE posterior probabilities in both models (x-axis and y-axis) and with high significance of c-score or d-score (darker color) are prioritized pathways for further mechanistic investigation. (*C*) Gene-wise heatmap of posterior mean of DE indicators of the ce.e2-dm.e2 in cel03440, ce.e2-dm.e2 in cel03430, and ce.dau-dm.pup in R-CEL-168643. Genes with concordant (orange) or discordant (blue) are shown in column beside the heatmaps.

validation. The workflow is flexible and extensible to most disease contents and general experimental design.

In Fig. 5, we present a guideline for practitioners to apply and interpret toward decision making. In Step 1, we perform DE analysis and intracohort congruence analysis in the reference (human) and comparison (mouse) groups, respectively, to confirm sufficient and stable DE information for the subsequent cross-species evaluation (see *SI Appendix*, Table S12 for case

study 1). Conceptually, intracohort c-score provides stability evaluation to troubleshoot DE analysis in a single cohort, where small intracohort c-scores point to underpowered and inconsistent DE analysis result. If the reference group has low number of DE genes and low intracohort concordance, the congruence analysis is not expected to succeed (Scenario I). For example, HL and HA in case study 1 have smaller sample sizes compared to the other four comparisons (N = 26 and 34 vs. N = 57 to 60 in *SI Appendix*, Table S2). This resulted in smaller numbers of DE genes and intracohort c-scores and thus almost no genome-wide congruence in HL-ML and HA-MA comparisons. When there is sufficient and stable DE information in the reference group but not in the comparison group (Scenario II), one can continue with the congruence analysis with caution. In this case, the genome-wide congruence is likely low but some pathway-specific congruence results may be valuable. In addition, increasing sample size in the comparison group is expected to provide a more conclusive finding. This is exactly the case of HS-MS and HT-MT comparisons. Finally, when both reference and comparison groups show sufficient differential response signals, the cross-species congruence framework (Step 2) can be performed with confidence, and a congruence decision can be made based on genome-wide congruence and pathway-specific evaluation incorporating prior biological knowledge, which is the case of HB-MB and HI-MI comparisons.

CAMO has two critical factors to consider when applying to different model organisms. First, when a model organism has fewer orthologous genes mapped, the quantification of pathway-specific congruence may be impacted. *SI Appendix*, section 1E and Fig. S11 present simulation analysis and show robust c-score quantification at genome-wide level when the number of mapped orthologs decreases. For pathway-specific analysis, however, variabilities of c-scores are much increased and results are less reliable. Second, pathway annotations are often biased toward human biology and understudied in model organisms. Since many pathway annotations in model organisms are computationally inferred and not accurate (28), users can apply annotation quality scores [e.g., GO Annotation Quality, GAQ (29)] to prefilter and select high-quality pathway annotations for congruence analysis. For this purpose, CAMO allows users to modify or customize the pathway definitions. Note that restricting pathways to only high-quality annotations may improve hypothesis generation during investigation, but potentially at a cost of reduced statistical power. The decision and balance should be contextual, for example, whether the research aim is exploratory or explanatory.

The current CAMO framework focuses on bulk-RNA transcriptome data with biological replicates. Multiple future directions are under development. When multilevel omics data, such as single-nucleotide polymorphism, methylation, miRNA expression, and protein expression, are available, an extended CAMO with

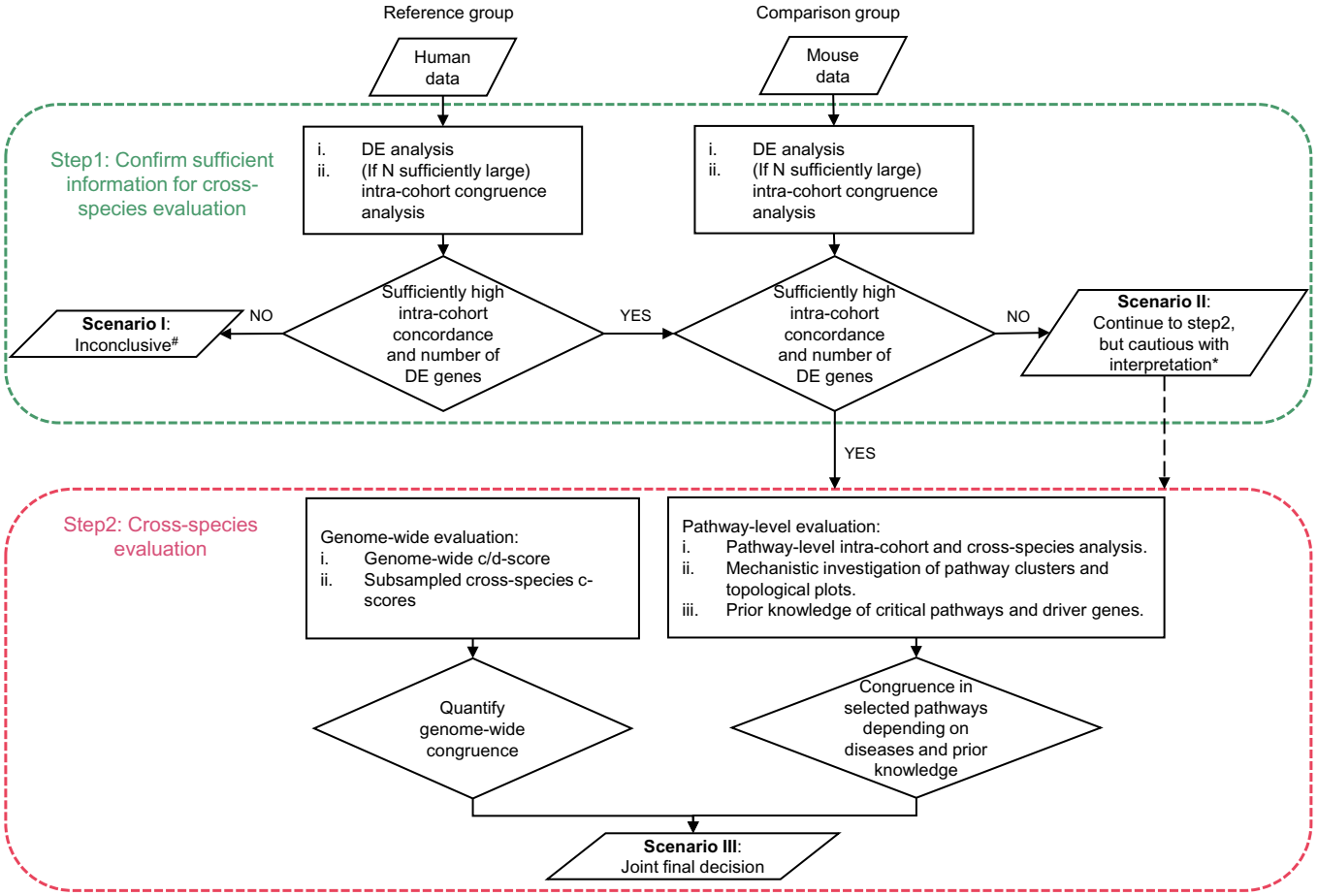

Fig. 5. Flowchart for practitioners to apply and interpret toward decision making. In Scenario I#, the DE signal of the reference group (humans) is insufficient, and the congruence analysis is not expected to succeed. In Scenario II*, since the comparison group (mice) does not have sufficient and stable DE information, genome-wide congruence is expected to be low, but one can continue with the congruence analysis with caution as some pathway-specific congruence results may be valuable. In Scenario III, when both reference and comparison groups show sufficient differential response signals, the cross-species congruence framework can be performed with confidence, and a congruence decision can be made based on genome-wide congruence and pathway-specific evaluation incorporating prior biological knowledge.

cross-species comparisons of genomes, epigenomes, transcriptomes, and gene regulation will provide holistic understanding of the congruence. Second, single-cell experimental data (e.g., scRNA-seq) will provide further congruence information specific to cell types. Finally, some types of experimental data typically do not contain replicates to capture biological heterogeneity and thus, the current framework based on conventional DE analysis does not fit. Notable examples include cell lines and 3 dimensional organoids in cancer research, where omics experiments are usually performed once without replicates. Extended modeling for investigating molecular congruence between cell lines or organoids and human cancer patients is an ongoing work.

Materials and Methods

Data Preprocessing and Pathway Database. The datasets used in case study 1 are downloaded from the publicly available GEO repository with accession numbers listed in *SI Appendix*, Table S2. All datasets are preprocessed separately by applying the suggested pipeline in R packages "affy" and "lumi" for Affymetrix and Illumina platforms, respectively. For datasets comparing human and mouse samples collected at multiple time points, we compare each time point to controls within species, respectively, and select the best-matched time points with the highest cross-species c-score while also consider d-score when it has prominent discordance. (*SI Appendix*, Table S3 for c-scores, d-scores, sample sizes, and matched time points are provided in *SI Appendix*). The processed RNA-seq datasets (in FPKM) of both species in case study 2 are downloaded from the author's website (http://jsb.ucla.edu/software-and-data). Genes with average FPKM smaller than one are filtered out followed by log2 transformation. Cross-species genes (worm vs. fly) are matched by Drosophila RNAi Screening Center Integrative Ortholog Prediction Tool (http://www.flyrnai.org/diopt) (30).

In case study 1, 219 KEGG (16) and Reactome (17) human pathways satisfying 1) gene set size between 5 and 200, 2) Fisher combined enrichment q value smaller than 0.05, 3) the minimum number of overlapping genes across studies greater than five, and 4) the median number of overlapping DE genes across studies greater than three are selected. In case study 2, 269 KEGG and Reactome worm pathways are selected similarly to have size between 3 and 500, the minimum number of overlapping genes across studies greater than three, the minimum number of overlapping DE genes across studies greater than two, and belonging to the top 50 enriched pathways in at least one of the studies.

Threshold-Free Bayesian Differential Analysis. CAMO applies a Bayesian mixture (BayesP) model (31) to derive DE posterior probabilities and to facilitate the calculation of c-scores and d-scores in the next section, where the input of BayesP can be single-study DE results from any conventional pipeline (e.g., "LIMMA" for microarray or log2-transformed and normalized RNA-seq data and "DESeq2" for RNA-seq counts). "LIMMA" was used for both case studies 1 and 2 since only normalized expression values are available for these public data. By assuming a nonparametric Dirichlet process prior on the grand means, Markov chain Monte Carlo (MCMC) using Gibbs sampling is used to sample the posterior probabilities of DE indicator δ_g, which will be used to derive the cross-species concordance and discordance scores later. The detailed modeling and MCMC procedure are outlines in *SI Appendix*, section 1A.

Deterministic Version of Cross-Species c-Scores and d-Scores. The foundation of cross-species c-scores and d-scores comes from a natural definition of confusion matrix and F-measure in machine learning (Table 1) when human and mouse DE status of upregulation (Ω^{H+} and Ω^{M+}), downregulation (Ω^{H-} and Ω^{M-}), and no change (Ω^{H0} and Ω^{M0}) are deterministically known, where $\Omega^{\mathrm{H+}} = \left\{g: \delta_g^H = 1\right\}$, $\Omega^{\mathrm{H-}} = \left\{g: \delta_g^H = -1\right\}$, and $\Omega^{\mathrm{H0}} = \left\{g: \delta_g^H = 0\right\}$ in human and δ_g^H is the DE indicator of gene g in human and similarly for mouse. a, e, and i denote the number of cross-species concordant genes (i.e., DE with the same directionality or no change in both): $a = |\Omega^{H+} \cap \Omega^{M+}|$ (number of concordant upregulated genes), $e = |\Omega^{H0} \cap \Omega^{M0}|$ (number of concordant no-change genes), and $i = |\Omega^{H-} \cap \Omega^{M-}|$ (number of concordant downregulated genes). The numbers of discordant genes (i.e., DE in one but DE with opposite directionality or no change in the other) can be similarly defined for b, c, d, f, g, h in the contingency table.

Table 1. Confusion matrix of the DE gene status comparing between a human study (H) and a mouse study (M).

	$\Omega^{\mathrm{H+}}$	Ω^{H0}	$\Omega^{\mathrm{H-}}$	sum
$\Omega^{\mathrm{M+}}$	a	b	c	A
Ω^{M0}	d	e	f	B
$\Omega^{\mathrm{M-}}$	g	h	i	C
sum	D	E	F	G

Ω^{H+} Ω^{H-}, and Ω^{H0} are collections of upregulated, downregulated, and nondifferential genes in humans, where $\Omega^{H+} = \{g: \delta_g^H = 1\}$, $\Omega^{H-} = \{g: \delta_g^H = -1\}$, and $\Omega^{H0} = \{g: \delta_g^H = 0\}$, and similarly for mouse. a, e, and i denote the no. of cross-species DE concordant genes: $a = |\Omega^{H+} \cap \Omega^{M+}|$ (no. of concordant upregulated genes), $e = |\Omega^{H0} \cap \Omega^{M0}|$ (number of concordant no-change genes), and $i = |\Omega^{H-} \cap \Omega^{M-}|$ (no. of concordant downregulated genes). b, c, d, f, g, h are defined similarly.

From the viewpoint of machine-learning prediction benchmark assuming we use mouse DE status to predict human DE status, one can define concordance $\mathrm{sensitivity_C}$ (a. k. a. $\mathrm{recall_c}$) $= \frac{a+i}{D+F}$ and $\mathrm{precision_c} = \frac{a+i}{A+C}$ when we focus on cross-species concordant DE genes, where $A = |\Omega^{M+}|, C = |\Omega^{M-}|, D = |\Omega^{H+}|$ and $F = |\Omega^{H-}|$. In $\mathrm{sensitivity_C}$, we calculate the number of concordant DE genes (i.e., $a + i$) among the true human DE genes (i.e., $D + F$). Similarly, $\mathrm{precision_C}$ is defined as the number of concordant DE genes (i.e., $a + i$) among the claimed mouse DE genes (i.e., $A + C$). We define the raw concordance score between humans and mice as the F-measure: $c' = 2\left(\mathrm{precision_c} \times \mathrm{recall_c}\right)/\left(\mathrm{precision_c} + \mathrm{recall_c}\right)$. Similarly, we can focus on discordant DE genes (i.e., genes upregulated in humans but downregulated in mice or vice versa) and define $\mathrm{sensitivity_D} = \frac{c+g}{D+F}$ and $\mathrm{precision_D} = \frac{c+g}{A+C}$. The raw discordance score between humans and mice becomes $d' = 2\left(\mathrm{precision_d} \times \mathrm{recall_d}\right)/\left(\mathrm{precision_d} + \mathrm{recall_d}\right)$. In addition to F-measure, we can also use the Youden index (=sensitivity + specificity − 1) or the geometric mean of sensitivity and specificity, where $\mathrm{specificity_C} = \frac{e}{E}$ and $\mathrm{specificity_D} = \frac{e}{B}$. When there is no reference study specified or under the general multicohort scenario, the F-measure is a better choice among the three because it is symmetric no matter which species is taken as the reference. With simple algebraic calculation, one can show that $c' = \frac{2(a+i)}{A+C+D+F}$ and $d' = \frac{2(c+g)}{A+C+D+F}$. Similar to Rand index used to evaluate clustering similarity and the adjusted Rand index subsequently developed (32), although both c'-score and d'-score range between 0 and 1, their expected value under null hypothesis (i.e., no resemblance between mice and humans) is not 0, making the interpretation difficult. To account for this pitfall, we adjust the scores to have maximum value at 1 for perfect resemblance and expected value at 0 when no resemblance exists using a linear transformation: $c-\mathrm{score} = \frac{c' - E(c' \mid H_0)}{1 - E(c' \mid H_0)}$ and $d-\mathrm{score} = \frac{d' - \mathrm{E}(d' \mid H_0)}{1 - E(d' \mid H_0)}$, where H_0 is the null hypothesis when mice and humans have no resemblance, $E\left(c' \mid H_0\right) = \frac{2(AD+CF)}{G(A+C+D+F)}$ and $E\left(d' \mid H_0\right) = \frac{2(AF+CD)}{G(A+C+D+F)}$ by computing the expected counts from the table margins for each cell (e.g., $E\left(a|H_0\right) = \frac{AD}{G}$).

Empirical (Data-Driven Estimation) Version of c-Scores and d-Scores. In practice, the underlying true DE statuses ($\Omega^{H+}, \Omega^{H0}, \Omega^{H-}$) and ($\Omega^{M+}, \Omega^{M0}, \Omega^{M-}$) are not known and are inferred from data. As previously mentioned, cross-species congruence analysis by applying arbitrary P value/FDR and fold change cutoffs can lead to subjective bias and inconsistent conclusions (5, 6). In CAMO, we infer Bayesian posterior probabilities and plug into the deterministic definition of c-scores and d-scores. Specifically, $\hat{\delta}_{gb}^H$ denotes the simulated estimation of δ_g^H in the b-th MCMC iteration for gene g in the HS and similarly $\hat{\delta}_{gb}^M$ for the mouse study. The unbiased estimators are obtained as $\hat{A} = \Sigma_g \Sigma_{b=1}^B \chi\left(\hat{\delta}_{gb}^M = 1\right)/B$, $\hat{C} = \Sigma_g \Sigma_{b=1}^B \chi\left(\hat{\delta}_{gb}^M = -1\right)/B$, $\hat{D} = \Sigma_g \Sigma_{b=1}^B \chi\left(\hat{\delta}_{gb}^H = 1\right)/B$, $\hat{F} = \Sigma_g \Sigma_{b=1}^B \chi\left(\hat{\delta}_{gb}^H = -1\right)/B$, $\hat{a} = \Sigma_g \Sigma_{b=1}^B \chi\left(\hat{\delta}_{gb}^H = 1 \;\&\; \hat{\delta}_{gb}^M = 1\right)/B$, $\hat{i} = \Sigma_g \Sigma_{b=1}^B \left(\hat{\delta}_{gb}^H = -1 \,\&\, \hat{\delta}_{gb}^M = -1\right)/B$, $\hat{g} = \Sigma_g \Sigma_{b=1}^B \chi\left(\hat{\delta}_{gb}^H = 1 \;\&\; \hat{\delta}_{gb}^M = -1\right)/B$, $\hat{c} = \Sigma_g \Sigma_{b=1}^B \chi\left(\hat{\delta}_{gb}^H = -1 \,\&\, \hat{\delta}_{gb}^M = 1\right)/B$, where B is the number of (post burn-in) MCMC simulations and $\chi(\cdot)$ is the indicator function taking value 1

if the statement is true and 0 otherwise. c-score and d-score are estimated by plugging these estimators into their deterministic definitions.

Pathway-Specific c-Scores and d-Scores. The aforementioned c-score and d-score estimations are calculated in the genome-wide scale. Since the cross-species congruence can vary by biological pathways, we analogously define pathway-specific c-scores and d-scores by constraining the calculation to each pathway. One major modification is when calculating the expected raw score under null hypothesis, a subsampled (sample without replacement) gene set with equivalent size of the target pathway is used to calculate $\widehat{E}^{(j)}\left(c' \mid H_0\right)$ and $\widehat{E}^{(j)}\left(d' \mid H_0\right)$ in the j-th sampling. We then estimate $\widehat{E}\left(c' \mid H_0\right) = \frac{1}{J}\Sigma_{j=1}^{J}\widehat{E}^{(j)}\left(c' \mid H_0\right)$ and $\widehat{E}\left(d' \mid H_0\right) = \frac{1}{J}\Sigma_{j=1}^{J}\widehat{E}^{(j)}\left(d' \mid H_0\right)$ to better represent the genome-wide status.

Statistical Significance (*P* Value) Assessment of c-score and d-score. We assess P values of genome-wide and pathway-specific c-scores and d-scores by permutation analysis. Specifically, we randomly permute cross-species ortholog gene annotation, so no cross-species congruence exists under the null hypothesis and the procedure is repeated for T times. The P values are calculated as $p\left(\widehat{c}\right) = \left(\Sigma_{t=1}^{T}\chi\left(\widehat{c}^{(t)} \geq \widehat{c}\right) + 1\right)/(T+1)$ and $p\left(\widehat{d}\right) = \left(\Sigma_{t=1}^{T}\chi\left(\widehat{d}^{(t)} \geq \widehat{d}\right) + 1\right)/(T+1)$, where $\widehat{c}$ and $\widehat{d}$ are the calculated c-score and d-score, and $\widehat{c}^{(t)}$ and $\widehat{d}^{(t)}$ are the derived c-score and d-score in the t-th permutation. Note that we count $\widehat{c}$ and $\widehat{d}$ as one of the permutation observations to avoid obtaining zero P values (33). Both pathway specific c-scores and d-scores and their associated P values are essential in CAMO to identify pathways most or least mimicked by the animal model and to investigate the underlying mechanism.

Pathway Clustering and Text Mining. In CAMO, the congruence analysis is evaluated in a pair of studies. When we assess M studies, CAMO will create $Q = C_2^M$ congruence analysis results. In practice, hundreds or up to thousands of pathways are assessed for c-scores and d-scores depending on selection criteria, and the result can contain high redundancy since different pathway databases may describe a related biological function using similar gene sets. $C_{K\times Q} = \left\{c_{kq}\right\}$ and $\Theta_{K\times Q} = \left\{\theta_{kq} = -log_{10}p\left(c_{kq}\right)\right\}$ denote the matrices of c-scores and associated minus-log-transformed P values of the Q congruence comparisons in K pathways. Note that large value of θ_{kq} represents high concordance in the q-th congruence evaluation of pathway k. To further decipher and interpret pathway-specific congruence result, we consider dissimilarity (Euclidean distance $d\left(\overrightarrow{\theta_k}, \overrightarrow{\theta_k}\right)$ between $\overrightarrow{\theta_k} = \left(\theta_{k1}, \cdots, \theta_{kQ}\right)$ and $\overrightarrow{\theta_{k'}} = \left(\theta_{k'1}, \cdots, \theta_{k'Q}\right)$ of pathways k and k' and using a consensus tight clustering algorithm to cluster the statistically significant pathways that pass the selection criteria in pathway size, minimum number of DE genes, q values from Fisher's combination method, etc. The algorithm uses the resampling-based consensus clustering (34) for identifying stable patterns in data followed by removing the scattered pathways with low silhouette width (35) iteratively until all pathways' silhouette widths are above a certain cutoff (e.g., 0.1) to improve the tightness of clusters. Pathways with similar concordance patterns across the Q pairwise comparisons of the M studies are clustered together to reduce redundancy and facilitate further investigation. A heatmap of the matrix $\Theta_{K\times Q}$ sorted by pathway clusters is shown to visualize the concordance patterns in different clusters (e.g., *SI Appendix*, Figs. S3*A* and S9*A*). A MDS algorithm is applied to the dissimilarity matrix generated from $\Theta_{K\times Q}$ for visualization (e.g., *SI Appendix*, Figs. S3A and S9*B*). Finally, the comembership heatmaps are used to summarize the proportion of significantly concordant pathways within each pathway cluster between each pair of studies (e.g., *SI Appendix*, Figs. S3*C* and S9*C*).

We next apply a text mining pipeline to extract summary annotations and retrieve knowledge from each pathway cluster (36). The method first collects names and summary descriptions of all pathways and extract noun phrases after filtering of biologically redundant phrases and merging synonyms using R packages *spacyr*, *tm*, *textstem*, and *wordnet*. Each noun phrases are tested for whether significantly enriched in selected pathway clusters by performing a permutation test on a cluster score weighted by length of pathway description. The output of text mining includes a list of key phrases most enriched and the corresponding permutation P values for each pathway cluster.

Individual Pathway Topology and Colocalized Concordant/Discordant Gene Module Detection. Pathway databases such as KEGG (16) and Reactome (17) provide pathway topological graphs to visualize involved genes, gene–gene interactions, and regulatory information in the pathway. In the R-shiny interface of CAMO, we map and incorporate the gene-based concordance/discordance inference results in mouse–human comparison to the pathway graph to allow users for visual mechanistic investigation of the local concordance/discordance pattern. For pathways from KEGG, we use R package "Pathview" (37) to render the topology graph and integrate the concordance/discordance information. For pathways from Reactome, we develop our own tool to first retrieve and parse the pathway topology (.sbgn file) from Reactome database using the Python *minidom* parser (https://docs.python.org/3/library/xml.dom.minidom.html). Then, each node is colored by its posterior mean of DE indicators in the two studies side by side using the Python Imaging Library (https://pillow.readthedocs.io/en/stable/).

To avoid visual bias and to further investigate the local concordance/discordance pattern inside the pathway, we develop a community detection algorithm to identify closely connected concordant or discordant gene modules based on shortest path distance in the graph, where the unweighted graph is constructed using R packages "KEGGgraph" (38) and "xml2", and the shortest path matrix is calculated by R package "igraph" (39). Exhaustive search algorithm is implemented to identify the concordant/discordant gene set with the smallest average shortest path at a given module size. However, for a pathway with a large number of concordant/discordant genes (e.g., size > 30), exhaustive search is not feasible, and a simulated annealing algorithm is used for fast search. Finally, a permutation test is performed to assess the P value of identified concordant or discordant gene modules (see *SI Appendix*, section 1B for details of simulated annealing and permutation).

In case study 1, we apply this local community detection algorithm to KEGG pathways hsa04670 and hsa04662 to identify discordant modules using exhaustive search. An elbow plot of $\left(avgSP_m\right)$ over m is generated from $m = 4$ to the cardinality of searching space, i.e., the total number of discordant genes (*SI Appendix*, Fig. S12). The SA algorithm with $x = 1$ and $y = 1$ generates similar results as the exhaustive search. The maximum module size whose P value is within two SD of the minimum P value is reported (i.e., 12 nodes containing 14 genes in hsa04670 (HS-MS), 6 nodes containing 7 genes in hsa04662 (HS-MS), and 4 nodes containing 4 genes in hsa04662 (HB-MB)). Corresponding KEGG topological plots with highlighted gene modules are shown in Fig. 3. We recommend users to consider the P value elbow plot and KEGG topological plot, together with their biological insights into determining an appropriate module size for further investigation.

Data, Materials, and Software Availability. Datasets used in the two case studies are publicly available from the following two papers, Li et al. (21) and Li et al. (20). CAMO R package and R-shiny app are available in github (https://github.com/CAMO-R).

ACKNOWLEDGMENTS. We acknowledge Diane Litman for helpful discussions. W.Z., T.R., L.Z., Y.Z., J.Z., Z.R., T.M., and G.C.T. are supported by R21LM012752 and R01CA190766. J.J.L. is supported by R35GM140888 and R01GM120507. A.V.L., S.O., and G.C.T. are supported by R01CA252378. T.M. is supported by the University of Maryland MPower Brain Health and Human Performance seed grant.

Author affiliations: [a]Department of Biostatistics, School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261; [b]Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX 77030; [c]Martinos Center for Biomedical Imaging, Harvard Medical School, Boston, MA 02129; [d]Department of Computer Science and Technology, Qilu University of Technology, Jinan, Shandong 250353, China; [e]Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15261; [f]Department of Statistics, University of California, Los Angeles, CA 90095; [g]Campbell Family Mental Health Research Institute at the Centre for Addiction and Mental Health, Toronto, ON M5S 2S1, Canada; [h]Department of Pharmacology and Chemical Biology, University of Pittsburgh Medical Center Hillman Cancer Center University of Pittsburgh, Pittsburgh, PA 15261; [i]Magee-Womens Research Institute, University of Pittsburgh Medical Center, Pittsburgh, PA 15123; [j]Department of Epidemiology and Biostatistics, School of Public Health, University of Maryland, College Park, MD 20742; [k]Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA 15261; and [l]Department of Computational and System, Biology, University of Pittsburgh, Pittsburgh, PA 15261

1. D. K. Brubaker, D. A. Lauffenburger, Translating preclinical models to humans. *Science* **367**, 742–743 (2020).
2. I. W. Mak, N. Evaniew, M. Ghert, Lost in translation: Animal models and clinical trials in cancer treatment. *Am. J. Transl. Res.* **6**, 114–118 (2014).
3. K. Rhrissorrakrai *et al.*, Understanding the limits of animal models as predictors of human biology: Lessons learned from the sbv IMPROVER species translation challenge. *Bioinformatics* **31**, 471–83 (2015).
4. R. J. Kleiman, M. D. Ehlers, Data gaps limit the translational potential of preclinical research. *Sci. Transl. Med.* **8**, 320ps1 (2016).
5. J. Seok *et al.*, Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 3507–12 (2013).
6. K. Takao, T. Miyakawa, Genomic responses in mouse models greatly mimic human inflammatory diseases. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 1167–72 (2015).
7. H. S. Warren *et al.*, Mice are not men. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E345 (2015).
8. K. Takao, Commonalities across species do exist and are potentially important. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E347–E348 (2015).
9. T. Shay, J. A. Lederer, C. Benoist, Genomic responses to inflammation in mouse models mimic humans: we concur, apples to oranges comparisons won't do. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E346 (2015).
10. D. K. Brubaker *et al.*, Computational translation of genomic responses from experimental model systems to humans. *PLoS Comput. Biol.* **15**, e1006286 (2019).
11. R. Normand *et al.*, Found in translation: A machine learning model for mouse-to-human inference. *Nat. Methods* **15**, 1067–1073 (2018).
12. C. Weidner *et al.*, Defining the optimal animal model for translational research using gene set enrichment analysis. *EMBO Mol. Med.* **8**, 831–838 (2016).
13. M. Ca *et al.*, XGSEA: CROSS-species gene set enrichment analysis via domain adaptation. *Brief Bioinform.* **22**, bbaa406 (2021).
14. T. E. Sweeney *et al.*, Gene expression analysis to assess the relevance of rodent models to human lung injury. *Am. J. Respir. Cell Mol. Biol.* **57**, 184–192 (2017).
15. M. A. Harris *et al.*, The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
16. M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
17. A. Fabregat *et al.*, The reactome pathway knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
18. D. Nishimura, BioCarta. *Biotech Software Internet Rep. Comput. Software J. Sci.* **2**, 117–120 (2001).
19. M. B. Gerstein *et al.*, Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445–448 (2014).
20. J. J. Li *et al.*, Comparison of D. melanogaster and C. elegans developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome Res.* **24**, 1086–101 (2014).
21. P. Li *et al.*, KERIS: Kaleidoscope of gene responses to inflammation between species. *Nucleic Acids Res.* **45**, D908–D914 (2017).
22. J. Mestas, C. C. Hughes, Of mice and not men: Differences between mouse and human immunology. *J. Immunol.* **172**, 2731–2738 (2004).
23. N. R. Genuth, M. Barna, Heterogeneity and specialized functions of translation machinery: from genes to organisms. *Nat. Rev. Genet.* **19**, 431–452 (2018).
24. W. A. Muller, Mechanisms of leukocyte transendothelial migration. *Annu Rev Pathol.* **6**, 323–44 (2011).
25. T. Kurosaki, Regulation of B-cell signal transduction by adaptor proteins. *Nat. Rev. Immunol.* **2**, 354–63 (2002).
26. R. Mevel *et al.*, RUNX transcription factors: Orchestrators of development. *Development* **146**, dev148296 (2019).
27. S. Govind, Innate immunity in Drosophila: Pathogens and pathways. *Insect Sci.* **15**, 29–43 (2008).
28. N. T. Doncheva *et al.*, Human pathways in animal models: Possibilities and limitations. *Nucleic Acids Res.* **49**, 1859–1871 (2021).
29. T. J. Buza *et al.*, Gene ontology annotation quality analysis in model eukaryotes. *Nucleic Acids Res.* **36**, e12 (2008).
30. Y. Hu *et al.*, An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics* **12**, 357 (2011).
31. Z. Huo, C. Song, G. Tseng, Bayesian latent hierarchical model for transcriptomic meta-analysis to detect biomarkers with clustered meta-patterns of differential expression signals. *Ann. Appl. Stat.* **13**, 340–366 (2019).
32. L. Hubert, P. Arabie, Comparing partitions. *J. Classification.* **2**, 193–218 (1985).
33. B. Phipson, G. K. Smyth, Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.* **9**, Article39 (2010).
34. S. Monti *et al.*, Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52**, 91–118 (2003).
35. P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
36. X. Zeng *et al.*, Comparative pathway integrator: A framework of meta-analytic integration of multiple transcriptomic studies for consensual and differential pathway analysis. *Genes (Basel)* **11**, 696 (2020).
37. W. Luo, C. Brouwer, Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29**, 1830–1831 (2013).
38. J. D. Zhang, S. Wiemann, KEGGgraph: A graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* **25**, 1470–1471 (2009).
39. G. Csardi, T. Nepusz, The igraph software package for complex network research. *InterJ. Complex Syst.* **1695**, 1–9 (2006).