CAMO: A molecular congruence analysis framework for evaluating model organisms

Wei Zong[1], Tanbin Rahman[2], Li Zhu[1], Xiangrui Zeng[3], Yingjin Zhang[1], Jian Zou[1], Song Liu[4], Zhao Ren[5], Jingyi Jessica Li[6], Steffi Osterreich[7,8], Tianzhou Ma[9*] and George C. Tseng[1,10,11*]

[1]Department of Biostatistics, Graduate school of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA

[2]Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

[3]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[4]Department of Computer Science and Technology, Qilu University of Technology, Jinan, Shandong, China

[5]Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15261, USA

[6]Department of Statistics, University of California, Los Angeles, CA 90095, USA

[7]Departments of Pharmacology and Chemical Biology, University of Pittsburgh, Pittsburgh, PA 15261, USA

[8]Magee-Womens Research Institute, UPMC, Pittsburgh, PA 15123, USA

[9]Department of Epidemiology and Biostatistics, School of Public Health, University of Maryland, College Park, MD 20742, USA

[10]Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA 15261, USA.

[11]Department of Computational and System Biology, University of Pittsburgh, Pittsburgh, PA 15261, USA.

[*]To whom correspondence should be addressed: E-mail:ctseng@pitt.edu (GCT); tma0929@umd.edu (TM)

ABSTRACT

CAMO provides a rigorous and user-friendly solution for quantification and mechanistic exploration of omics congruence in model organisms and humans. It performs threshold-free differential analysis, quantitative concordance/discordance scoring, pathway-centric investigation, and topological subnetwork detection. Instead of dichotomous claims of "poorly" or "greatly" mimicking humans, CAMO facilitates discovery and visualization of specific molecular mechanisms that are best or least mimicked, providing foundations for hypothesis generation and subsequent translational investigations.

KEYWORDS

Model organism, Molecular congruence analysis, Transcriptome, Translational research

As human studies often encounter numerous recruitment and ethical constraints, model organisms have played an indispensable role in pre-clinical research to understand pathogenesis and treatment response at the behavioral, cellular, and molecular levels. Their clinical validity and translational values are, however, long been debated with controversial opinions[1-4]. A notable example is the contradictory conclusions from two articles analyzing an identical transcriptomic response dataset in human and mouse inflammation[5,6], with the former concluding "poorly mimicking" of the mouse model and the latter reporting "greatly mimicking", which triggered further debates[7-9]. Although 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, methods for exploring mechanistic insights are lacking. To meet the gap, we develop a Congruence Analysis of Model Organisms (CAMO) framework to evaluate omics congruence of animal models and aid mechanistic understanding, hypothesis generation, and translational guidance.

Fig. 1a-b overview CAMO's pipeline. "Bayesian differential analysis" contrasting case/control or treated/non-treated groups are performed in human and mouse cohorts separately and "concordance and discordance scores" (abbreviated as c-score and d-score) are calculated (Fig. 1a), reflecting degree of cross-species congruence and discrepancy. The threshold-free Bayesian differential model transforms p-values obtained from routine pipelines, such as LIMMA or DEseq2, into differential posterior probabilities, which in turn are input to cross-species c-score/d-score calculation based on a stochastic version of confusion matrix and F-measure in the machine learning setting with p-values assessed by permutation. When multiple cohorts are jointly analyzed, c-scores and d-scores are calculated for all pair-wise studies in each pathway. Next, the "Mechanistic investigation and hypothesis generation" component can perform "pathway knowledge retrieval" and "topological gene module detection" (Fig. 1b). In pathway knowledge retrieval, top (concordance or discordance) enriched pathways are clustered with similar congruence patterns across studies and a text

mining algorithm is implemented to retrieve representative keywords to interpret each pathway cluster. In each pathway, a community detection algorithm can further identify concordant or discordant subnetworks based on topological regulatory information.

Our first case study re-evaluates the contradicting papers in human-mouse transcriptomic response to inflammation. Supplementary Table 1 lists 12 studies in human and mouse (Burns, Infection, Trauma , Sepsis, LPS and ARDS), abbreviated as HB, HI, HT, HS, HL and HA, and MB, MI, MT, MS, ML and MA (H for human and M for mouse). Supplementary Table 2 summarizes analytical differences and arbitrary thresholds in the two papers that may have contributed to the contradicting conclusions. CAMO avoids such subjective analytical thresholds and decisions and extends the investigation into pathways and gene regulatory modules for insightful mechanistic understanding. Supplementary Table 3 contains genome-wide c-scores and d-scores of pair-wise studies and the c-score-based multidimensional scaling (MDS) plot in Supplementary Fig. 1 shows that four human studies HB, HI, HT and HS resemble each other well (c-scores=0.25~0.52). MI, MB and MT are relatively similar to the four human studies while MA, ML and MS have almost no genome-wide congruence to human, implying that cross-species congruence is condition specific. Unlike most of the human studies, the six mouse studies generally do not mimic each other, implying complexity and high variability of mouse models in inflammatory diseases. We next apply consensus tight clustering to 219 concordance enriched pathways and identify four pathway clusters (Supplementary Fig. 2-3). For example, the heatmap and text mining results show high congruence between human (HB, HS, HT and HI) and mouse models (MB, MI and MT) in both innate and adaptive (e.g., B and T cell related) immunity (Supplementary Fig. 3 and Supplementary Table 4). Despite the difference in neutrophil and lymphocyte abundance, it has been reported that the overall immune system is relatively similar in mouse and human[15]. Cluster IV shows mouse models do not mimic human studies

in ribosome and protein translation. Such findings agree with earlier studies that profound cross-species differences exist in translation machinery[16].

CAMO next provides interactive exploration in the shiny app to select pathways and zoom in for regulatory topological visualization. Fig. 1c shows differential expression (DE) evidence and c-scores/d-scores of individual pathways in pair-wise comparisons of HB, HS, MB and MS. Fig. 1d displays gene-specific concordance or discordance information in two selected KEGG pathways. Fig. 2a highlights the KEGG gene-gene regulatory topological plot for "Leukocyte transendothelial migration" pathway (hsa04670) with side-by-side display of the differential regulation signals (red for up-regulation and green for down-regulation) in HS and MS. The community detection algorithm identifies a module of 14 DE genes (RHOA, PTK2B, RAC2, RAC1, CDC42, ITGA4, ITGB2, MSN, PXN, NCF2, CYBA, GNAI1, GNAI2 and GNAI3) with opposite effect sizes (green in HS and red in MS or vice versa; $p = 0.002$). The co-localized discordant module is directly related to cell motility and direct sensing, a critical function that allows leukocytes to attach to the vessel wall to initiate immune response during inflammation[17]. The striking mouse-human discordant result may reflect the discrepancy in proportions of different cell types of blood leukocytes between human and mouse as pointed out in a previous critique[9]. The topological plot for "B cell receptor signaling pathway" (hsa04662) (Fig. 2b) shows a gene module of 7 discordant genes (PTPN6, DAPP1, CD79A, RAC1, RAC2, GRB2, CD19; $p=0.009$). CD79A and CD19 are antigen receptors on B cell membrane to regulate signaling molecules, such as GRB2 and RAC family, with important roles in the regulation of cell growth and movement. On the other hand, Fig. 2c shows generally concordant DE signals between HB and MB (both red or both green) in the B cell membranes receptor and signaling, including CD72, CD79A, CD79B, IFITM1, CD19, CR2 and BLNK. Previous literature has pointed out the similarity but also significant differences between mouse and human immunology, specifically in B cell development[15].

The second case study evaluates transcriptomic congruence of *C. elegans* (ce) and *D. melanogaster* (dm) in developmental stages using the modENCODE data[18,19], where five developmental stages in each species are confirmed by hierarchical clustering (Supplementary Fig. 4): early embryo (ce.e0), mid embryo (ce.e1), late embryo (ce.e2), larvae (ce.lar), dauer (ce.dau) using *C. elegans* adult as the reference; early embryo (dm.e0), mid embryo (dm.e1), late embryo (dm.e2), larvae (dm.lar), pupae (dm.pup) using female Drosophila adult as reference. Supplementary Fig. 5a shows MDS plot of genome-wide c-scores for the five Drosophila and five *C. elegans* stages. The y-axis shows a clear separation between the two species. The x-axis presents a developmental transition in the embryonic stages e0→e1→e2, while the larvae and pupae/dauer stages are not exactly ordered. Adjacent developmental stages are found to be more similar within species. ce.e2 shows some resemblance with all developmental stages in Drosophila except for dm.e0. This unintuitive result is better visualized by an intriguing bipartite graph between Drosophila and *C. elegans* stages (Supplementary Fig.5b) by creating solid edges when pair-wise genome-wide c-scores are greater than 0.1 (Supplementary Table 5). 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) 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[18,19]. 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.

From 269 concordance enriched pathways, consensus tight clustering identifies six pathway clusters (Supplementary Fig. 6,7), including Cluster III related to cell cycle and DNA replication with cross-species late-stage congruence, Cluster IV related to estrogen and hormone in *C. elegans*, and Cluster II specific to Drosophila developmental stages.

Supplementary Fig. 8 shows DE evidence and c-scores/d-scores in ce.e2, ce.dau, dm.e2 and dm.pup. Pathways “Homologous recombination” (KEGG: cel03440) and “Mismatch repair” (KEGG: cel03430) exhibited high concordance between ce.e2 and dm.e2 (Supplementary Fig. 9), implying similar molecular events taking place in 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 (Supplementary Fig. 9). 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 present in the pupae stage of Drosophila but not in the dauer stage of *C. elegans*[20].

Notwithstanding the fact that human and animal studies are both fundamental in disease and drug investigation, objective and interactive congruence evaluation to distinguish and integrate cross-species omics information is currently lacking. We expect that CAMO and its future extension to multi-omics and single cell data will provide profound cross-species mechanistic understanding to improve animal models and to accelerate treatment development in human diseases.

ONLINE METHODS

Threshold-free Bayesian differential analysis

CAMO applies a Bayesian mixture (BayesP) model to derive differential expression posterior probabilities and to facilitate calculation of c-scores and d-scores in the next section, where the input of BayesP can be DE results from any conventional pipeline (e.g., "limma" for microarray and "DESeq2" for RNA-seq are used in this paper). Specifically, the one-sided p-values, p_g for gene g, from conventional DE analysis are first transformed to z-scores $Z_g = \Phi^{-1}(p_g)$ ($\Phi^{-1}(\cdot)$ is inverse CDF of standard normal distribution), which incorporates information of both statistical significance and directionality. BayesP adopts the following three-component Gaussian mixture model for up-regulated, down-regulated and no-change genes similar to Huo et al. 2019 [21]:

$$f(Z_g|\delta_g) = I(\delta_g = 0)N(Z_g; 0,1) + I(\delta_g = 1)N(Z_g; \mu_{g+}, 1) + I(\delta_g = -1)N(Z_g; \mu_{g-}, 1),$$

where δ_g is the DE indicator of gene g ($\delta_g = 1$ indicates up-regulation, $\delta_g = -1$ for down-regulation and $\delta_g = 0$ for no change), and μ_{g+} and μ_{g-} are the grand means of z-scores of the up-regulated and down-regulated groups. We assume a non-parametric Dirichlet process prior on the grand means: $\mu_{g+} \sim G_+, G_+ \sim DP(G_{0+}, 1)$; $\mu_{g-} \sim G_-, G_- \sim DP(G_{0-}, 1)$, where G_{0+} (G_{0-}) denotes a left (right) truncated $N(0, 10^2)$ and 1 is the concentration parameter of the Dirichlet process. A Chinese Restaurant Process is used to update δ_g, where we define an auxiliary component variable $C_g \in \{\dots, -2, -1, 0, 1, 2, \dots\}$ such that $C_g = 0$ indicates $\delta_g = 0$, $C_g > 0$ indicates $\delta_g = 1$ and $C_g < 0$ indicates $\delta_g = -1$. The prior for δ_g is specified as: $P(\delta_g \neq 0) = \pi_g, P(\delta_g = 1|\delta_g \neq 0) = \rho_g$; $\pi_g \sim Beta\left(\frac{\gamma}{G-\gamma}, 1\right), \rho_g \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right)$, where $1 - \pi_g$, $\pi_g\rho_g$ and $\pi_g(1 - \rho_g)$ are the prior probabilities of being no-change, up-regulated and down-regulated genes respectively.

Markov chain Monte Carlo (MCMC) using Gibbs sampling is used to update all parameters $(\pi_g, \rho_g, \delta_g)$ sequentially as follows:

1. Update π_g: $\pi_g|\cdot \sim Beta\left(\frac{\gamma}{G-\gamma} + \sum I\{\delta_g = 1\} + \sum I\{\delta_g = -1\}, \sum I\{\delta_g = 0\} + 1\right)$

2. Update ρ_g: $\rho_g|\cdot \sim Beta\left(\frac{1}{2}+\sum I\{\delta_g=1\}, \frac{1}{2}+\sum I\{\delta_g=-1\}\right)$
3. Update $C_g: P(C_g=k|\cdot) \propto h_k(Z_g|C_{-g})(1-\pi_g)^{I(k=0)}(\pi_g\rho_g)^{I(k>0)}(\pi_g(1-\rho_g))^{I(k<0)}$, where $h_k(Z_g|C_{-g})$ is derived according to Neal 2000 [22].
4. Update δ_g: $\delta_g = sgn(C_g)$ where $sgn(\cdot)$ is the sign function.

The empirical distribution of the posterior probabilities of δ_g will be used to derive the cross-species c-scores and d-scores later.

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 (Supplementary Table 7) when human and mouse DE status of up-regulation (Ω^{H+} and Ω^{M+}), down-regulation (Ω^{H-} and Ω^{M-}) and no change (Ω^{H0} and Ω^{M0}) are deterministically known, where $\Omega^{H+}=\{g:\delta_g^H=1\}$, $\Omega^{H-}=\{g:\delta_g^H=-1\}$ and $\Omega^{H0}=\{g:\delta_g^H=0\}$ in human, and similarly for mouse. Denote by a, e and i the number of cross-species DE concordant genes: $a=\#(\Omega^{H+}\cap\Omega^{M+})$ (number of concordant up-regulated genes), $e=\#(\Omega^{H0}\cap\Omega^{M0})$ (number of concordant no-change genes), and $i=\#(\Omega^{H-}\cap\Omega^{M-})$ (number of concordant down-regulated genes). The numbers of DE discordant genes can be similarly defined for b,c,d,f,g,h in the contingency table. From the viewpoint of machine learning prediction benchmark assuming we use mouse DE status to predict human DE status, one can define concordance sensitivity$_C$ (a.k.a. recall$_C$)=$\frac{a+i}{D+F}$, and 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 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, 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 DE concordance score between human and mouse as the F-measure: c' =2(precision$_C$ x recall$_C$)/(precision$_C$+recall$_C$). Similarly, we can focus on DE discordant genes (i.e., genes up-regulated in human but down-regulated in mouse or vice versa) and define

sensitivity$_D$=$\frac{c+g}{D+F}$ and precision$_D$=$\frac{c+g}{A+C}$. The raw DE discordance score between human and mouse becomes: d' =2(precision$_D$ x recall$_D$)/(precision$_D$+recall$_D$). In addition to F-measure, we can also use Youden index (=sensitivity+specificity-1) or the geometric mean of sensitivity and specificity, where specificity$_C$=$\frac{e}{E}$ and specificity$_D$=$\frac{e}{B}$. When there is no reference study specified or under the general multi-cohort 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 [23], although both c'-score and d'-score range between 0 and 1, their expected value under null hypothesis (i.e., no resemblance between mouse and human) 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-score = $\frac{c\prime-E(c\prime|H_0)}{1-E(c\prime|H_0)}$ and d-score = $\frac{d\prime-E(d\prime|H_0)}{1-E(d\prime|H_0)}$, where H_0 is the null hypothesis when mouse and human have no resemblance, $E(c'|H_0) = \frac{2(AD+CF)}{G(A+C+D+F)}$ and $E(d'|H_0) = \frac{2(AF+CD)}{G(A+C+D+F)}$ by computing the expected counts from the table margins for each cell (e.g. $E(a|H_0) = \frac{AD}{G}$).

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, denote by $\hat{\delta}_{gb}^{H}$ the simulated estimation of δ_g^H in the b-th MCMC iteration for gene g in the human study and similarly $\hat{\delta}_{gb}^{M}$ for the mouse study. The unbiased estimators are obtained as: $\hat{A} =$

$\sum_g \sum_{b=1}^{B} \chi\left(\hat{\delta}_{gb}^{M} = 1\right)/B$, $\hat{C} = \sum_g \sum_{b=1}^{B} \chi\left(\hat{\delta}_{gb}^{M} = -1\right)/B$, $\hat{D} = \sum_g \sum_{b=1}^{B} \chi\left(\hat{\delta}_{gb}^{H} = 1\right)/B$, $\hat{F} = \sum_g \sum_{b=1}^{B} \chi\left(\hat{\delta}_{gb}^{H} = -1\right)/B$, $\hat{a} = \sum_g \sum_{b=1}^{B} \chi\left(\hat{\delta}_{gb}^{H} = 1 \,\&\, \hat{\delta}_{gb}^{M} = 1\right)/B$, $\hat{\imath} = \sum_g \sum_{b=1}^{B} \chi\left(\hat{\delta}_{gb}^{H} = -1 \,\&\, \hat{\delta}_{gb}^{M} = -1\right)/B$, $\hat{g} = \sum_g \sum_{b=1}^{B} \chi\left(\hat{\delta}_{gb}^{H} = 1 \,\&\, \hat{\delta}_{gb}^{M} = -1\right)/B$, $\hat{c} = \sum_g \sum_{b=1}^{B} \chi\left(\hat{\delta}_{gb}^{H} = -1 \,\&\, \hat{\delta}_{gb}^{M} = 1\right)/B$, 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 $\hat{E}^{(j)}(c'|H_0)$ and $\hat{E}^{(j)}(d'|H_0)$ in the j-th sampling. We then estimate $\hat{E}(c'|H_0) = \frac{1}{J}\sum_{j=1}^{J} \hat{E}^{(j)}(c'|H_0)$ and $\hat{E}(d'|H_0) = \frac{1}{J}\sum_{j=1}^{J} \hat{E}^{(j)}(d'|H_0)$ 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 $p(\hat{c}) = \left(\sum_{t=1}^{T} \chi\left(\hat{c}^{(t)} \geq \hat{c}\right) + 1\right)/(T+1)$ and $p(\hat{d}) = \left(\sum_{t=1}^{T} \chi\left(\hat{d}^{(t)} \geq \hat{d}\right) + 1\right)/(T+1)$, where $\hat{c}$ and $\hat{d}$ are the calculated c-score and d-score, and $\hat{c}^{(t)}$ and $\hat{d}^{(t)}$ are the derived c-score and d-score in the t-th permutation. Note that we count $\hat{c}$ and $\hat{d}$ as one of the permutation observations to avoid obtaining zero p-values [24]. Benjamini-Hochberg (BH) procedure [25] is applied to adjust for

multiple comparisons of testing many pathways. 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, congruence analysis is evaluated in a pair of studies. When we assess M studies, CAMO will create $Q = C_2^M$ congruence analysis results. Depending on selection of pathway databases, hundreds or up to thousands of pathways are assessed for c-scores and d-scores, and the result can contain high redundancy since different pathway databases may describe a related biological function using similar gene sets. Denote by $C_{K\times Q} = \{c_{kq}\}$ and $\Theta_{K\times Q} = \{\theta_{kq} = -log_{10}p(c_{kq})\}$ 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 $d(\vec{\theta}_k, \vec{\theta}_{k'})$) between $\vec{\theta}_k = (\theta_{k1}, \dots, \theta_{kQ})$ and $\vec{\theta}_{k'} = (\theta_{k'1}, \dots, \theta_{k'Q})$ of pathways k and k' and cluster the statistically significant pathways (i.e., meta-analyzed q-values by Fisher's method across Q comparisons ≤ 0.05) using a consensus tight clustering algorithm. The algorithm uses the resampling-based consensus clustering [26] for identifying stable patterns in data followed by removing the scattered pathways with low silhouette width [27] 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 (Supplementary Fig.3a,7a). A multidimensional scaling (MDS) algorithm is applied to the dissimilarity matrix generated from $\Theta_{K\times Q}$ for visualization (Supplementary Fig.3b,7b). Finally, the co-membership heatmaps are

used to summarize the proportion of significantly concordant pathways within each pathway cluster between each pair of studies (Supplementary Fig.3c,7c).

We next apply an automated text mining pipeline to extract summary annotations and retrieve knowledge from each pathway cluster [28]. 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*. The remaining 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 co-localized concordant/discordant gene module detection

Pathway databases such as KEGG [29] and Reactome [30] 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" [31] to render the topology graph and integrate the concordance/discordance information. For pathways from Reactome, we developed 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 assignment 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" [32] and "xml2", and the shortest path matrix is calculated by R package "igraph" [33]. 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 (SA) algorithm is used for fast search. We define the initial temperature T_0, the temperature multiplier μ, the number of iterations for reaching equilibrium N, and the final temperature T_f. Intuitively, T_0 controls the acceptance of a trial assignment, N controls the maximum number of annealing iterations, μ controls speed of cooling down process for T_0 to drop below T_f and stop the process. The annealing process is harder when T_0 is larger, μ is smaller and N is larger. Real data evaluation shows $T_0 = 10$, $\mu = 0.95$, $T_f = 1e-5$ and $N = 1000$ work well in general. The energy function is defined as the average shortest path of the current module denoted as $avgSP(G_m)$ where m is a given module size. A trial module G'_m is proposed by randomly substituting one node in G_m to another one in the searching space. If $avgSP(G'_m) < avgSP(G_m)$, G_m will be accepted immediately, otherwise it will initiate the annealing process. Another parameter R is introduced to control the total iterations in case it keeps hitting $avgSP(G'_m) = \infty$ throughout the algorithm in a very sparse graph. Detailed algorithm can be found in the Supplementary notes 1. When it is applied to a spectrum of module sizes, to further improve the performance, at each module size m, the algorithm runs x times and the top y results are stored and passed to the next $m + 1$ scenario as initials. Borrowing initial values from the previous step allows this procedure to converge faster and $y > 1$ helps to robustize the procedure when multiple close-to-optimal solutions exist. The overall SA algorithm including the initialization procedure is summarized in the Supplementary notes 2.

Permutation test is performed to assess the p-value of identified concordant or discordant gene modules. Specifically, for an observed smallest $avgSP_m$ at module size m, gene modules of the same size are sampled without replacement from the searching space B times resulted

in $avgSP_m^{(b)}, b = 1, \dots, B$. The p-value is derived as $p(avgSP_m) = \left(\sum_{b=1}^{B} \chi\left(avgSP_m^{(b)} \leq avgSP_m\right) + 1\right)/(B+1)$ and the standard deviation of the $p(avgSP_m)$ is estimated as $\sqrt{\frac{1}{B}p(avgSP_m)(1-p(avgSP_m))}$ by regarding the permuted p-value as the mean estimate of B approximately independent Bernoulli trials.

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 $(avgSP_m)$ over m is generated from $m = 4$ to the cardinality of searching space i.e., the total number of discordant genes (Supplementary Fig. 10). 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 2 standard deviations of the minimum p-value is reported (12 nodes containing 14 genes in hsa04670 and 6 nodes containing 7 genes in hsa04662). Corresponding KEGG topology plots with highlighted gene modules are shown in Fig. 2. We recommend users to consider the p-value elbow plot, KEGG topology plots together with their biological insights in determining an appropriate module size to report.

Data availability

The datasets used in Case Study 1[34] are publicly available on the NCBI GEO database with accession numbers in Supplement Table 1. The datasets used in Case Study 2 are available at http://jsb.ucla.edu/software-and-data.

Code availability

An open-source R package can be downloaded from https://github.com/weiiizong/CAMO.

Acknowledgements

The authors acknowledge Diane Litman for helpful discussions. WZ, TR, LZ, YZ, JZ, ZR, TM and GCT are supported by R21LM012752 and R01CA190766. JJL is supported by R35GM140888 and R01GM120507.

Author contributions

TM and GCT conceived the framework. WZ, TR, LZ, XZ, YZ, JZ, SL, ZR, TM and GCT developed the statistical and computational methods, and performed the analyses. WZ, JJL, SO, TM and GCT discussed the results and contributed to the final manuscript presentation. All authors proofread and agreed with the final manuscript.

Competing Interests

The authors declare no competing interests.

REFERENCES

1 Brubaker, D. K. & Lauffenburger, D. A. Translating preclinical models to humans. *Science* **367**, 742-743 (2020).
2 Mak, I. W., Evaniew, N. & Ghert, M. Lost in translation: animal models and clinical trials in cancer treatment. *American journal of translational research* **6**, 114 (2014).
3 Rhrissorrakrai, K. *et al.* Understanding the limits of animal models as predictors of human biology: lessons learned from the sbv IMPROVER Species Translation Challenge. *Bioinformatics* **31**, 471-483 (2015).
4 Kleiman, R. J. & Ehlers, M. D. Data gaps limit the translational potential of preclinical research. *Science translational medicine* **8**, 320ps321-320ps321 (2016).
5 Seok, J. *et al.* Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences* **110**, 3507-3512 (2013).
6 Takao, K. & Miyakawa, T. Genomic responses in mouse models greatly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences* **112**, 1167-1172 (2015).
7 Warren, H. S. *et al.* Mice are not men. *Proceedings of the National Academy of Sciences* **112**, E345-E345 (2015).
8 Takao, K., Hagihara, H. & Miyakawa, T. Reply to Warren et al. and Shay et al.: Commonalities across species do exist and are potentially important. *Proceedings of the National Academy of Sciences* **112**, E347-E348 (2015).
9 Shay, T., Lederer, J. A. & Benoist, C. Genomic responses to inflammation in mouse models mimic humans: we concur, apples to oranges comparisons won't do. *Proceedings of the National Academy of Sciences* **112**, E346-E346 (2015).
10 Brubaker, D. K., Proctor, E. A., Haigis, K. M. & Lauffenburger, D. A. Computational translation of genomic responses from experimental model systems to humans. *PLoS computational biology* **15**, e1006286 (2019).
11 Normand, R. *et al.* Found In Translation: a machine learning model for mouse-to-human inference. *Nature methods* **15**, 1067-1073 (2018).
12 Weidner, C., Steinfath, M., Opitz, E., Oelgeschläger, M. & Schönfelder, G. Defining the optimal animal model for translational research using gene set enrichment analysis. *EMBO molecular medicine* **8**, 831-838 (2016).
13 Cai, M., Hao Nguyen, C., Mamitsuka, H. & Li, L. XGSEA: CROSS-species gene set enrichment analysis via domain adaptation. *Brief Bioinform*, doi:10.1093/bib/bbaa406 (2021).
14 Sweeney, T. E., Lofgren, S., Khatri, P. & Rogers, A. J. Gene expression analysis to assess the relevance of rodent models to human lung injury. *American Journal of Respiratory Cell and Molecular Biology* **57**, 184-192 (2017).
15 Mestas, J. & Hughes, C. C. Of mice and not men: differences between mouse and human immunology. *The Journal of Immunology* **172**, 2731-2738 (2004).
16 Genuth, N. R. & Barna, M. Heterogeneity and specialized functions of translation machinery: from genes to organisms. *Nat Rev Genet* **19**, 431-452, doi:10.1038/s41576-018-0008-z (2018).
17 Muller, W. A. Mechanisms of leukocyte transendothelial migration. *Annual Review of Pathology: Mechanisms of Disease* **6**, 323-344 (2011).
18 Gerstein, M.B., Rozowsky, J., Yan, K.K., Wang, D., Cheng, C., Brown, J.B., Davis, C.A., Hillier, L., Sisu, C., Li, J.J. and Pei, B., 2014. Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445-448 (2014).
19 Li JJ, Huang H, Bickel PJ, Brenner SE. Comparison of D. melanogaster and C. elegans developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome research* **24**, 1086-101 (2014).
20 Govind, S. Innate immunity in Drosophila: Pathogens and pathways. *Insect science* **15**, 29-43 (2008).
21 Huo, Z., Song, C., Tseng, G. Bayesian latent hierarchical model for transcriptomic meta-analysis to detect biomarkers with clustered meta-patterns of differential expression signals. *The annals of applied statistics* **13**, 340 (2019).
22 Neal, R. M. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics* **9**, 249-265 (2000).
23 Hubert L, Arabie P. Comparing partitions. *Journal of classification* **2**, 193-218 (1985).
24 Phipson B, Smyth GK. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology* **9**, (2010).
25 Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289-300 (1995).
26 Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* **52**, 91-118 (2003).
27 Rousseeuw, PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53-65 (1987).
28 Zeng, X. *et al.* Comparative Pathway Integrator: a framework of meta-analytic integration of multiple transcriptomic studies for consensual and differential pathway analysis. *Genes* **11**, 696 (2020).
29 Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research.***28**, 27-30 (2000).
30 Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B. and Milacic, M. The reactome pathway knowledgebase. *Nucleic acids research* **46**, D649-D655 (2018).
31 Luo, W. & Brouwer, C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29**, 1830-1831 (2013).
32 Zhang JD, Wiemann S. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* **25**, 1470-1 (2009).
33 Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal, complex systems*. **1695**,1-9 (2006)..
34 Li, P., Tompkins, R. G., Xiao, W. & Program, I. a. H. R. t. I. L.-S. C. R. KERIS: kaleidoscope of gene responses to inflammation between species. *Nucleic acids research*, gkw974 (2016).

Figure Legends

Figure 1. Workflow of the "CAMO" framework with application results from Case Study 1. (a) Procedures to calculate genome-wide and pathway level c-scores and d-scores for a pair of human study (HS) and mouse model (MM). (b) Downstream machine learning and bioinformatics interactive visualization tools for pathway knowledge retrieval and topological gene nodule detection. (c) Summary of DE evidence with pathway level c-scores (orange in the upper right region) and d-scores (blue in lower left region). X- and Y-axes represent the average DE posterior probabilities, and size of dots represents the magnitudes of c-scores (orange) or d-scores (blue). Two example pathways are highlighted using different shapes ("◇": hsa04662 - KEGG: B cell receptor signaling pathway; "□": hsa04670 - KEGG: Leukocyte transendothelial migration). (d) 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.

Figure 2. Pathway topology plots of the selective pathways Case Study 1. (a). hsa04670 (HS-MS), (b). hsa04662 (HS-MS) and (c). hsa04662 (HB-MB). Pop-out plots represent the co-localized 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 up-regulation and green for down-regulation).

Figure 1

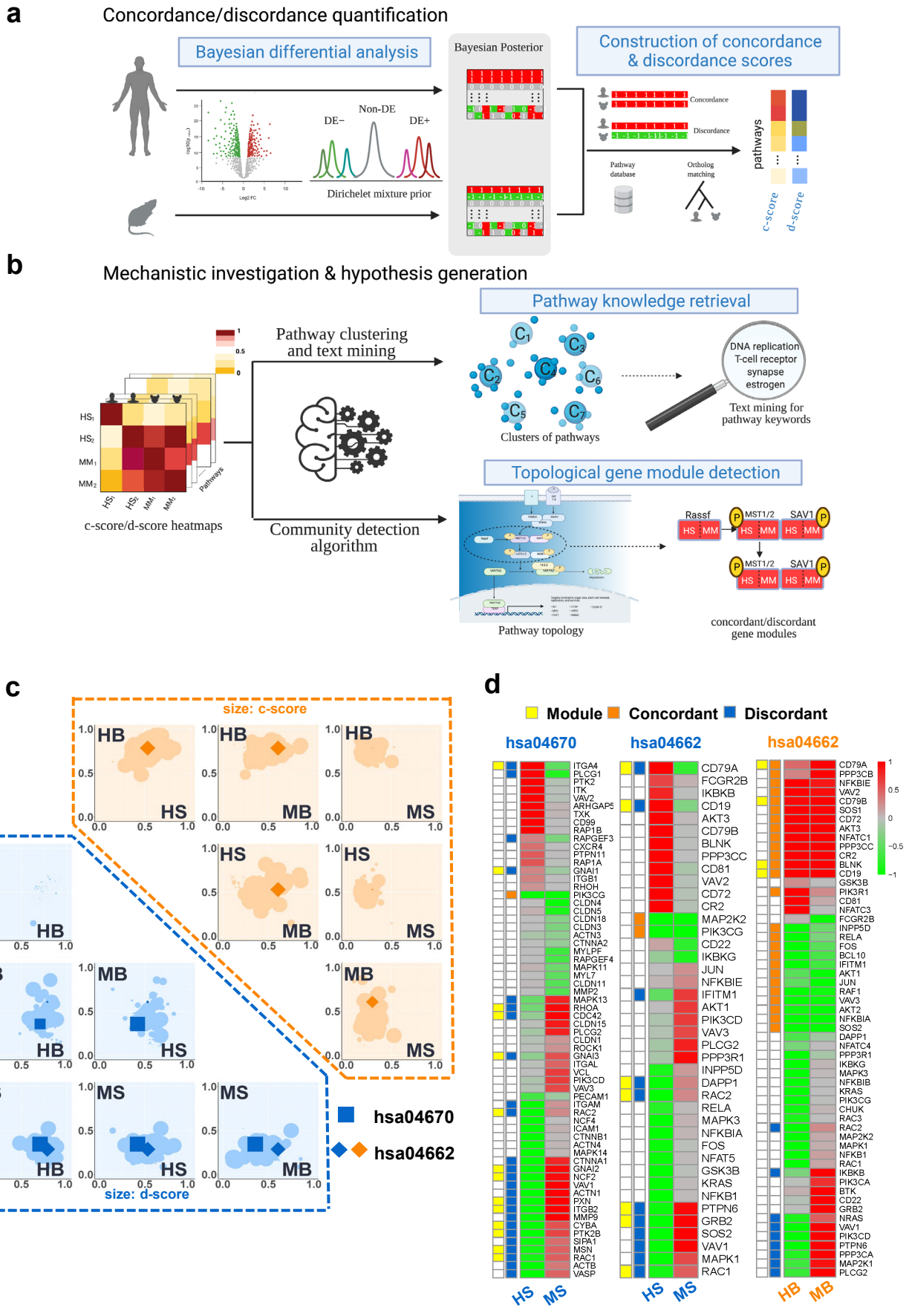
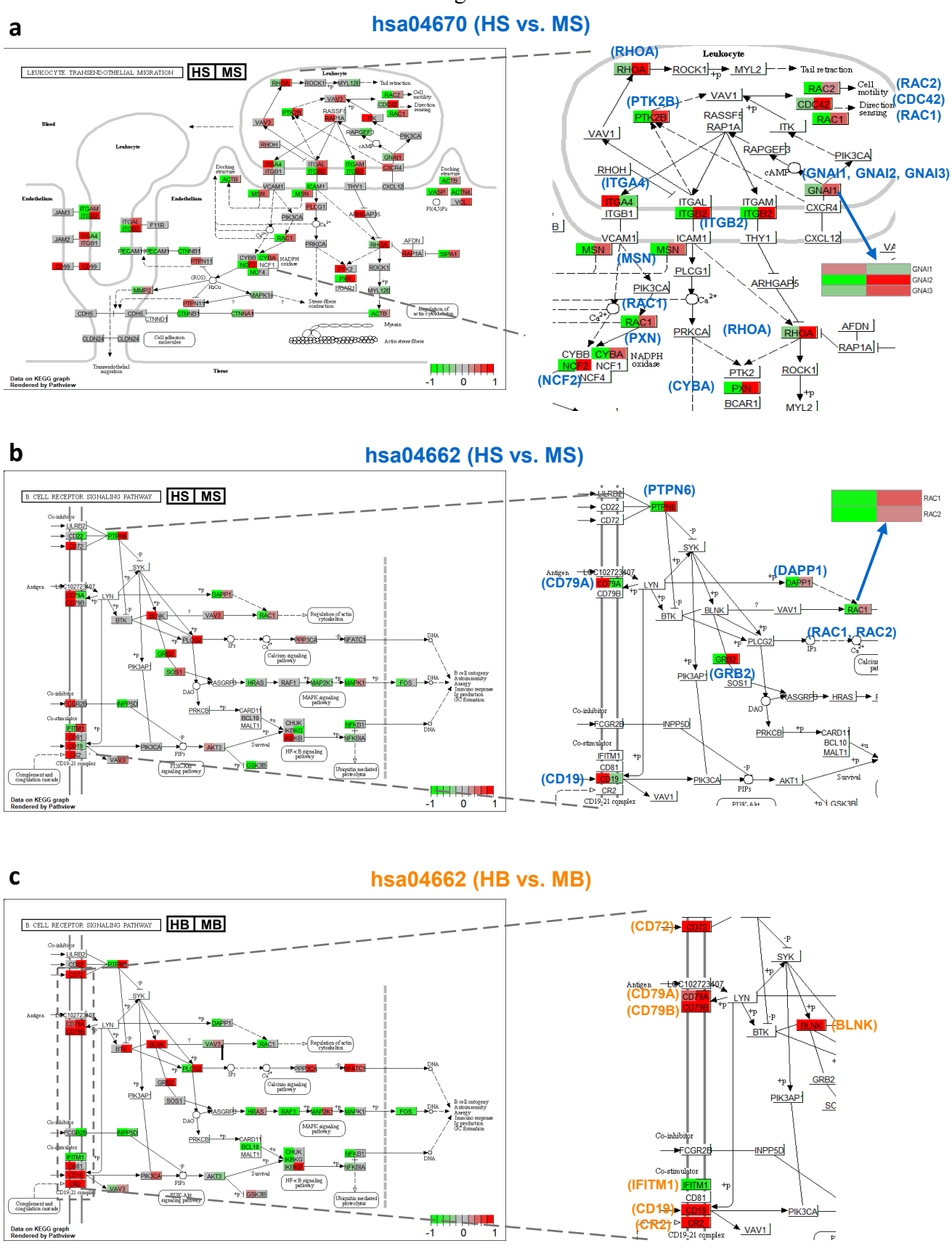

a
Concordance/discordance quantification
Bayesian differential analysis
Bayesian Posterior
Construction of concordance & discordance scores
Non-DE
DE−
DE+
Dirichelet mixture prior
Concordance
Discordance
Pathway database
Ortholog matching
pathways
c-score
d-score
b
Mechanistic investigation & hypothesis generation
Pathway knowledge retrieval
Pathway clustering and text mining
c-score/d-score heatmaps
Clusters of pathways
DNA replication
T-cell receptor
synapse
estrogen
Text mining for pathway keywords
Topological gene module detection
Community detection algorithm
Pathway topology
Rassf
MST1/2
SAV1
concordant/discordant gene modules
c
size: c-score
size: d-score
hsa04670
hsa04662
d
Module
Concordant
Discordant
hsa04670
hsa04662
hsa04662
HS
MS
HB
MB

Figure 2


a
hsa04670 (HS vs. MS)
LEUKOCYTE TRANSENDOTHELIAL MIGRATION
HS MS
(RHOA)
(RAC2)
(CDC42)
(RAC1)
(PTK2B)
(ITGA4)
(ITGB2)
(GNAI1, GNAI2, GNAI3)
(MSN)
(RAC1)
(PXN)
(NCF2)
(CYBA)
b
hsa04662 (HS vs. MS)
B CELL RECEPTOR SIGNALING PATHWAY
HS MS
(PTPN6)
(DAPP1)
(CD79A)
(RAC1, RAC2)
(GRB2)
(CD19)
c
hsa04662 (HB vs. MB)
B CELL RECEPTOR SIGNALING PATHWAY
HB MB
(CD72)
(CD79A)
(CD79B)
(BLNK)
(IFITM1)
(CD19)
(CR2)
Data on KEGG graph
Rendered by Pathview