Briefings in Bioinformatics, 22(5), 2021, 1–11

https://doi.org/10.1093/bib/bbaa406
Problem Solving Protocol

XGSEA: CROSS-species gene set enrichment analysis via domain adaptation

Menglan Cai, Canh Hao Nguyen, Hiroshi Mamitsuka and Limin Li

Corresponding author: Limin Li, Xi'an Jiaotong University, 28 Xianning W Rd, Jiaoda Commerce Block, Beilin, Xi'an 710049, China.
Tel.: +86 029 82660967; E-mail: liminli@mail.xjtu.edu.cn

Abstract

Motivation: Gene set enrichment analysis (GSEA) has been widely used to identify gene sets with statistically significant difference between cases and controls against a large gene set. GSEA needs both phenotype labels and expression of genes. However, gene expression are assessed more often for model organisms than minor species. Also, importantly gene expression are not measured well under specific conditions for human, due to high risk of direct experiments, such as non-approved treatment or gene knockout, and then often substituted by mouse. Thus, predicting enrichment significance (on a phenotype) of a given gene set of a species (target, say human), by using gene expression measured under the same phenotype of the other species (source, say mouse) is a vital and challenging problem, which we call CROSS-species gene set enrichment problem (XGSEP). **Results:** For XGSEP, we propose the CROSS-species gene set enrichment analysis (XGSEA), with three steps of: (1) running GSEA for a source species to obtain enrichment scores and *p*-values of source gene sets; (2) representing the relation between source and target gene sets by domain adaptation; and (3) using regression to predict *p*-values of target gene sets, based on the representation in (2). We extensively validated the XGSEA by using five regression and one classification measurements on four real data sets under various settings, proving that the XGSEA significantly outperformed three baseline methods in most cases. A case study of identifying important human pathways for T -cell dysfunction and reprogramming from mouse ATAC-Seq data further confirmed the reliability of the XGSEA. **Availability:** Source code of the XGSEA is available through https://github.com/LiminLi-xjtu/XGSEA.

Key words: cross-species study; domain adaptation; gene set enrichment analysis

Introduction

Due to recent advancement of modern experimental technologies, currently we have a massive amount of basic biological data. For example, next-generation sequencing technology has made sequencing faster and lower-cost, generating an incredible number of sequences. This situation makes bioinformatics tools more promising in retrieving biological knowledge from the data. For example, gene set enrichment analysis (GSEA) [1] has been well used in biology and related areas, which can rank gene set(s) most relevant (precisely, statistically significant) to binary-labeled gene expression measurement. However, GSEA needs gene expression data labeled binary, such as control and case, and is heavily affected by missing data.

Indeed gene expression are now measured by more speedy and precise techniques like RNA-Seq than cDNA microarray, while measuring gene expression is still costly both on money and time. Existing expression data often have strong bias in measured organisms or species. Model organisms, such as

Menglan Cai is currently working toward the PhD degree in Xi'an Jiaotong University, China. Her research interest includes bioinformatics.
Canh Hao Nguyen has been working in machine learning and bioinformatics. His current interest includes machine learning for graph data with applications in biological networks.
Hiroshi Mamitsuka is a professor of Kyoto University, Japan. His current research interests include mining from graphs and networks in biology and chemistry.
Limin Li is a professor of Xi'an Jiaotong University, China. Her research interests include machine learning and the applications in bioinformatics and biostatistics.
Submitted: 18 September 2020; **Received (in revised form):** 12 December 2020

Mus musculus, *Caenorhabditis elegans*, *Arabidopsis thaliana*, etc., are well measured, while data on minor species are relatively insufficient. Additionally, human gene expression data are unable to be measured under some specific conditions, due to high risk of direct experiments on human, such as non-approved treatment or gene knockout. On the other hand, mouse is usually used to study human disease [2, 3] because of lower cost, lower risk and relatively strong homology relationship with human [4]. However, there exists essential differences between mouse and human [5–8]. Effective treatments developed by mouse data often fail in human clinical trials [9, 10]. Thus it would be strongly expected to develop a method to bridge the gap between expression data of different species, such as human and mouse.

We consider a problem of predicting enrichment significance of given gene sets of one species (such as human) without gene expression, by using sufficient gene expression data of another species (such as mouse). The assumption behind this problem is that both expression data are measured under the same phenotype. We call this problem cross-species gene set enrichment problem (XGSEP). Assume that we have enough data behind XGSEP for human and mouse (more generally target and source), except target expression data. A gene set, either from mouse or human, could be represented as a binary annotation vector with dimension being the number of all genes in the expression data, representing whether the corresponding gene is in the gene set. The enrichment significance (such as p-value) of a source gene set S with an annotation vector $\boldsymbol{x}_s$ can be computed by traditional GSEA. The goal of XGSEP is to predict the enrichment significance for a target gene set T with an annotation vector $\boldsymbol{x}_t$, which might have a different dimension from $\boldsymbol{x}_s$ since the number the total genes for target (human) and source (mouse) are different. Note that the sequence homology between target genes and source genes is assumed to be represented by binary matrix $\boldsymbol{M}$, which should be important information for the prediction.

A naive idea for XGSEP would be to first find a source gene set $\boldsymbol{x}_s$, most homologous to genes in a particular target gene set $\boldsymbol{x}_t$, by using $\boldsymbol{M}$. Then, GSEA is run over source expression data and $\boldsymbol{x}_s$. The resultant p-value for $\boldsymbol{x}_s$ is considered as a prediction of the enrichment p-value for $\boldsymbol{x}_t$. The method is simple and fast, but the homology relationship between source and target is often complex, and thus homologous source gene set $\boldsymbol{x}_s$ cannot be clearly defined. Also using $\boldsymbol{M}$ directly would be not robust.

Our idea for XGSEP is, rather than focusing on only one gene set, to consider many gene sets at once and train a predictive machine learning model by these gene sets. Suppose that we have source gene sets $S_1, \ldots, S_m$ and target gene sets $T_1, \ldots, T_n$, with annotation matrices $\boldsymbol{X}_s = [\boldsymbol{x}_s^1, \ldots, \boldsymbol{x}_s^m]$ and $\boldsymbol{X}_t = [\boldsymbol{x}_t^1, \ldots, \boldsymbol{x}_t^n]$, respectively. Then the enrichment p-values for the source gene sets can be computed beforehand (by traditional GSEA). The goal of XGSEP is to predict enrichment p-values for target gene sets $\boldsymbol{x}_t^1, \ldots, \boldsymbol{x}_t^n$. Note that $\boldsymbol{X}_s$ (training data) and $\boldsymbol{X}_t$ (test data) are different in size of rows (number of genes), and thus it is difficult to compare the two matrices directly, meaning that a regular machine learning model such as a classifier generated by $\boldsymbol{X}_s$ cannot be run directly over test data $\boldsymbol{X}_t$. Thus, a further idea is to transform both the target and source species into a common space so that the target and source genes can be compared. However, this idea cannot be realized by regular machine learning models by the above problem of difference in size between training and test data. We solve this problem by domain adaptation, transfer learning between two domains: target and source. In general domain adaption, a machine learning model, trained by a larger amount of labeled samples from a source domain, is applied to a target domain with very few or no labeled samples [11]. This is exactly the same situation of XGSEP. A common way of domain adaptation methods is to train a model so that the model can reduce the probability gap between two domains. A possible measure for the probability gap, i.e. the difference of two data distributions, is maximum mean discrepancy (MMD) [12–15]. We will borrow the idea of domain adaptation and MMD to solve XGSEP.

We propose a method, XGSEA, standing for Cross-species Gene Set Enrichment Analysis (XGSEA). The XGSEA solves XGSEP by three steps: (1) we run GSEA over the source gene sets to obtain gene enrichment scores $\boldsymbol{E}_s$ and gene enrichment significance $\boldsymbol{v}_s$. (2) We first define pairwise similarities among gene sets based on $\boldsymbol{M}$, and then propose a MMD-based domain adaptation method to project $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$ into a latent common space with affine mappings $\boldsymbol{P}_s$ and $\boldsymbol{P}_t$ to obtain $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_t$, respectively, so that (i) the probability gap between $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_t$ in the latent space is minimized and (ii) $\boldsymbol{P}_s$ and $\boldsymbol{P}_t$ are smooth over the connection $\boldsymbol{M}$ between source and target gene sets.

By solving this optimization problem, we can obtain the optimal new representations $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_t$ for source and target gene sets, respectively. (3) We train a regression model by $(\boldsymbol{Z}_s, \boldsymbol{E}_s)$ and run the trained model over $\boldsymbol{Z}_t$ to predict enrichment scores $\boldsymbol{E}_t$ for target gene sets and finally p-values $\boldsymbol{v}_t$ with the principle of null hypothesis. Schematically, we may be able to explain our idea by using arrows: $\boldsymbol{X}_s \xrightarrow{\boldsymbol{P}_s} \boldsymbol{Z}_s$ and $\boldsymbol{X}_t \xrightarrow{\boldsymbol{P}_t} \boldsymbol{Z}_t$, so that the adaptive representations $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_t$ for source and target gene sets should have the smallest distribution divergence and preserve their pairwise homology similarities.

The contribution of this work can be summarized into three-fold: (1) we define a problem, XGSEP, which is helpful for understanding a particular phenotype (label) of a species with too limited data to run GSEA. (2) We propose a three-step method called XGSEA for XGSEP through domain adaptation that projects gene sets from two species into a common latent space. This projection is formulated as a nonlinear optimization problem, by which we can estimate the latent space and also estimate the enrichment scores and p-values of target gene sets through the latent space. Furthermore, the computational complexity of the optimization problem is low enough so that the computation of the XGSEA becomes feasible over regular gene annotation matrices. (3) We empirically validated the XGSEA by using four different real phenotypes with expression data. The experimental results show that the XGSEA significantly outperformed three baseline methods under various settings. The advantage of the XGSEA was further confirmed by a case study of finding significant unknown human pathways for T-cell dysfunction and reprogramming from a mouse ATAC-Seq data set.

Method: CROSS-species Gene Set Enrichment Analysis (XGSEA)

To the best of our knowledge, there is no existing work for XGSEP. A similar problem setting might be cross-species gene set analysis (XGSA) [16]. The goal of the XGSA is different with our XGSEP. XGSA aims to compare a gene set from one species with a gene set from another species. That is, XGSA directly examines if two gene sets (from two different species) are significantly different or not, only through the homology between genes in given two gene sets.

Problem definition

We have two species, source and target. Let $A = \{a_1, \cdots, a_p\}$ be a source (say mouse) gene set, and $B = \{b_1, \cdots, b_q\}$ be a target (say human) gene set. Let $\mathbf{M} \in \mathbb{R}^{p\times q}$ be a binary matrix of sequence homology, where the (i,j)-element $\mathbf{M}(i,j)$ is one if source gene a_i is homologous to target gene b_j; otherwise zero. Suppose that we have gene expression matrix $\mathbf{G}_s$ with phenotype vector $\mathbf{y}_s$ for source genes only, meaning that we can run GSEA over $\mathbf{G}_s$ and $\mathbf{y}_s$ to compute gene set enrichment significance for an arbitrary source gene set.

Suppose further that we have multiple gene sets for both source and target. Let $\mathcal{S} = \{S_1, \cdots, S_m\}$ be m source gene sets and $\mathcal{T} = \{T_1, \cdots, T_n\}$ be n target gene sets. We define an annotation matrix for source gene sets $\mathcal{S}$ (for columns) by A (for rows) as $\mathbf{X}_s = [\boldsymbol{x}_s^1, \cdots, \boldsymbol{x}_s^m] \in \{0,1\}^{p\times m}$ for source gene sets $S_1, \cdots, S_m$, where the i-th element of $\boldsymbol{x}_s^j$ is one if gene a_i is in gene set S_j and zero otherwise. Similarly, let $\mathbf{X}_t = [\boldsymbol{x}_t^1, \cdots, \boldsymbol{x}_t^n] \in \{0,1\}^{q\times n}$ be the annotation matrix for target gene sets $\mathcal{T}$. Then the problem, XGSEP standing for CROSS-species Geneset Enrichment Problem, is, given $\mathbf{G}_s, \mathbf{y}_s, \mathbf{X}_s, \mathbf{X}_t$ and $\mathbf{M}$, to estimate the enrichment p-value of each gene set in $\mathcal{T}$ with respect to the same phenotype of $\mathbf{y}_s$. We propose our method XGSEA, standing for CROSS-species Gene Set Enrichment Analysis, to solve XGSEP by using three steps. Figure 1 shows a schematic picture of the three-step procedure of the XGSEA. Below, we will explain each of these three steps in detail.

Step 1: gene set enrichment analysis for source

Since gene expression $\mathbf{G}_s$ and phenotype $\mathbf{y}_s$ are both available for the source side, we can directly use regular GSEA to obtain p-values, $v_{s,1}, \cdots, v_{s,m}$ for $S_1, \cdots, S_m$, respectively. In fact, p-value $v_{s,i}$ corresponds to null hypothesis $H_0^{s,i}$: gene set S_i has no association with phenotype $\mathbf{y}_s$ (against the entire set of genes) and can be computed by the following procedure [1].

1a. Compute enrichment score $E_{s,i}^0$ for gene set S_i by using gene expression $\mathbf{G}_s$ and phenotype $\mathbf{y}_s$.

1b. Permute the entries in $\mathbf{y}_s$ and recompute the enrichment score for gene set S_i. Repeat this step B times to generate an empirical null distribution of the enrichment score: E_{NULL} with $E_{s,i}^1, \cdots, E_{s,i}^B$.

1c. Compute empirical, nominal p-value $v_{s,i}$ for S_i from null distribution E_{NULL} by using the positive (or negative) region of the distribution corresponding to observed enrichment score $E_{s,i}^0$.

For source gene set S_i, we can compute $B+1$ enrichment scores $E_{s,i}^0, \cdots, E_{s,i}^B$ in **1a** and **1b** to compute p-value $v_{s,i}$ in **1c**. Similarly for target gene set T_j, we can first predict $B+1$ enrichment scores $E_{t,j}^0, \cdots, E_{t,j}^B$ for target gene set T_j and then p-value $v_{t,j}$ in **1c**.

Step 2: domain adaptation for source and target gene sets

We project the target and source genes into a common space, to maximally use the information from the source gene side for the target gene sets. The objective function is formulated as follows.

We project $\mathbf{X}_s$ and $\mathbf{X}_t$ to a common subspace in $\mathbb{R}^d$ by using affine mappings $\mathbf{P}_s \in \mathbb{R}^{p\times d}$ and $\mathbf{P}_t \in \mathbb{R}^{q\times d}$, respectively, such that $\mathbf{Z}_s = [\mathbf{z}_s^1, \cdots, \mathbf{z}_s^m] = \mathbf{P}_s^T\mathbf{X}_s$ and $\mathbf{Z}_t = [\mathbf{z}_t^1, \cdots, \mathbf{z}_t^n] = \mathbf{P}_t^T\mathbf{X}_t$.

In this process, we can set the following two reasonable objectives:

(1) Probability divergence between $\mathbf{Z}_s$ and $\mathbf{Z}_t$ should be small.

(2) Pairwise distances among the gene sets in $\mathbf{Z}_s$ and $\mathbf{Z}_t$ should be preserved.

For the first objective, we use MMD [12, 14] to measure the divergence. An empirical estimate of MMD can be defined as follows:

$$\begin{aligned}\mathcal{D}(\mathbf{Z}_s, \mathbf{Z}_t) &= \|\frac{1}{m}\sum_{i=1}^{m}\phi(\mathbf{z}_s^i) - \frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{z}_t^i)\|_H^2, \\ &= \sum_{i,j=1}^{m}\frac{k(\mathbf{z}_s^i, \mathbf{z}_s^j)}{m^2} + \sum_{i,j=1}^{n}\frac{k(\mathbf{z}_t^i, \mathbf{z}_t^j)}{n^2} - 2\sum_{i,j=1}^{m,n}\frac{k(\mathbf{z}_s^i, \mathbf{z}_t^j)}{mn} \\ &= \text{trace}(\mathbf{KL}), \end{aligned} \tag{1}$$

where $\phi(\cdot)$ is a mapping to reproducible kernel Hilbert space H, $k(\cdot,\cdot) = (\phi(\cdot), \phi(\cdot))$ is the kernel associated to this mapping, and

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{ss} & \mathbf{K}_{st} \\ \mathbf{K}_{ts} & \mathbf{K}_{tt} \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)}, \tag{2}$$

where the (i,j)-element of $\mathbf{K}_{ab}$ is

$$\mathbf{K}_{ab}(i,j) = k(\mathbf{z}_a^i, \mathbf{z}_b^j), a, b \in \{s,t\}, i = 1, \cdots, m, j = 1, \cdots, n,$$

and the (i,j)-element of $\mathbf{L}$ is

$$\mathbf{L}(i,j) = \begin{cases} 1/m^2 & i,j \in \{1, \cdots, m\}; \\ 1/n^2 & i,j \in \{m+1, \cdots, m+n\} \\ -1/mn & \text{otherwise.} \end{cases} \tag{3}$$

For the second objective, we can first define the pairwise homologous similarity between source gene sets $S_1, \cdots, S_m$ and target gene sets $T_1, \cdots, T_n$ from given data directly as follows:

$$\begin{aligned} \mathbf{W}_{ss}(i,j) &= \min\{\tfrac{|S_i\cap S_j|}{|S_i|}, \tfrac{|S_i\cap S_j|}{|S_j|}\} \quad i,j \in \{1, \cdots, m\}; \\ \mathbf{W}_{tt}(i,j) &= \min\{\tfrac{|T_i\cap T_j|}{|T_i|}, \tfrac{|T_i\cap T_j|}{|T_j|}\} \quad i,j \in \{1, \cdots, n\}; \\ \mathbf{W}_{st}(i,j) &= \min\{\tfrac{|T_j\cap \widetilde{S}_i|}{|T_j|}, \tfrac{|S_i\cap \widetilde{T}_j|}{|S_i|}\} \quad i \in \{1, \cdots, m\}, \\ & \qquad\qquad\qquad\qquad\quad j \in \{1, \cdots, n\}, \end{aligned} \tag{4}$$

where $|A|$ is the number of genes in set A, $\widetilde{S}_i = \phi_M(S_i) \subset T$ is the set with the target genes homologous to the source genes in S_i, and $\widetilde{T}_j = \phi_M(T_i) \subset S$ is the set with the source genes homologous to the target genes in T_j. The projections $\mathbf{P}_s$ and $\mathbf{P}_t$ should be smooth over homologous similarity matrix $\mathbf{W} = \begin{bmatrix} \mathbf{W}_{ss} & \mathbf{W}_{st} \\ \mathbf{W}_{st}^T & \mathbf{W}_{tt} \end{bmatrix}$ so that Z_s and Z_t could preserve the gene sets' pairwise distances captured by $\mathbf{W}$.

Thus, the divergence $\mathcal{D}$ in Equation (1) should be minimized, being regularized by the smoothness of the projections $\mathbf{P}_s$ and $\mathbf{P}_t$ over similarity matrix $\mathbf{W}$. Overall the objective function can be

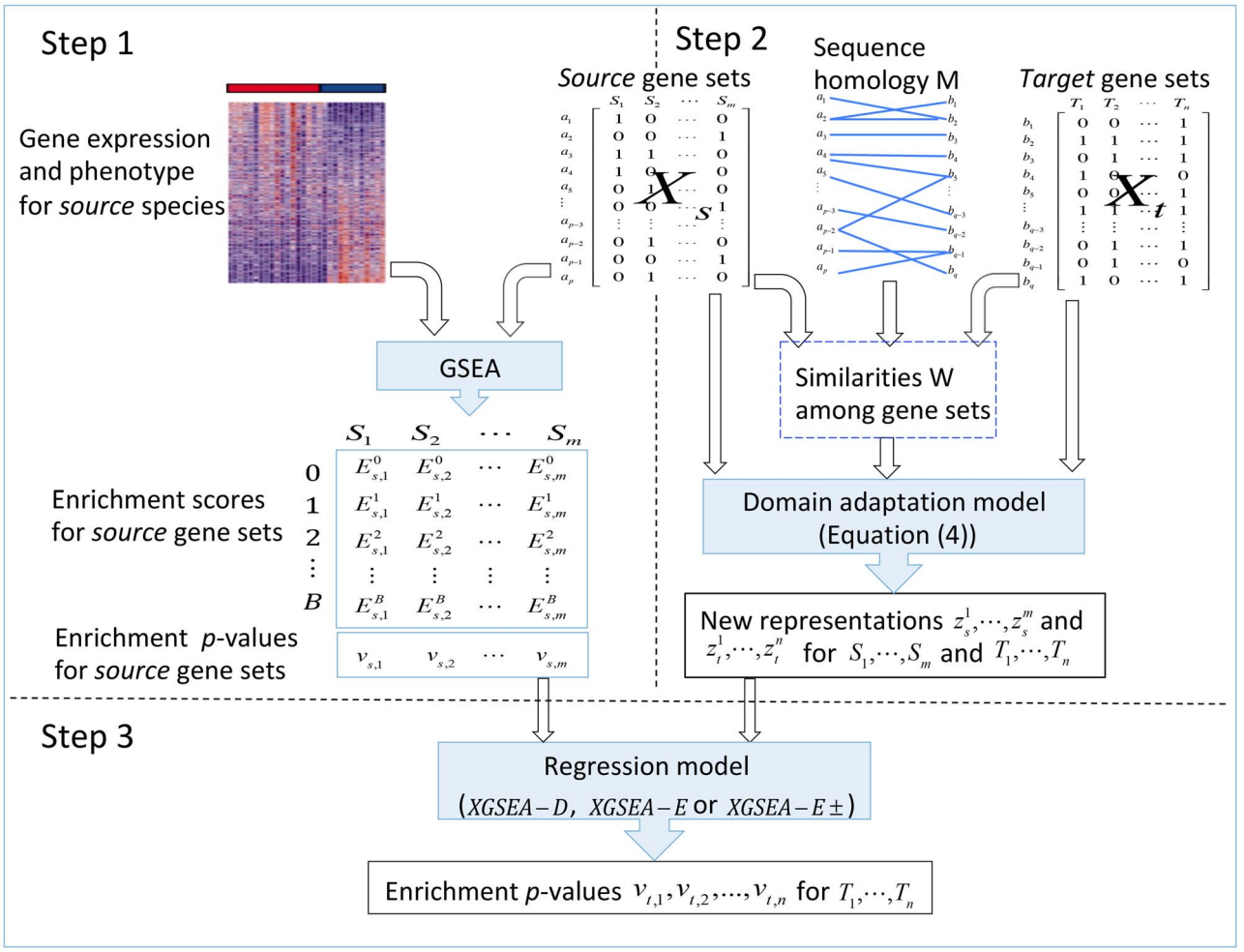

Figure 1. Flow chart of the XGSEA: we (A) compute B+1 enrichment scores and *p*-values for each source gene set by GSEA, where B is the number of permutation, (B) obtain new representations for all source and target gene sets by domain adaptation and (C) predict enrichment *p*-values for target gene sets by a regression model based on the new representations.

given as follows:

$$\min_{P_s^TP_s+P_t^TP_t=I} \mathcal{D}(\mathbf{P}_s^T\mathbf{X}_s,\mathbf{P}_t^T\mathbf{X}_t)+\lambda\Big(\frac{1}{2}\sum_{i,j=1}^{m}\mathbf{W}_{ss}(i,j)\|\mathbf{z}_s^i-\mathbf{z}_s^j\|_2^2$$
$$+\sum_{i,j=1}^{m,n}\mathbf{W}_{st}(i,j)\|\mathbf{z}_s^i-\mathbf{z}_t^j\|_2^2+\frac{1}{2}\sum_{i,j=1}^{n}\mathbf{W}_{tt}(i,j)\|\mathbf{z}_t^i-\mathbf{z}_t^j\|_2^2\Big), \quad (5)$$

where the constraint $\mathbf{P}_s^T\mathbf{P}_s+\mathbf{P}_t^T\mathbf{P}_t=I$ is used to avoid trivial solutions.

The model (5) is rewritten in a simpler form and then can be solved by the optimization algorithm on Grassmann manifold. The details of our optimization algorithm and our computational complexity reduction strategy are given in Section 1.1 of the supplementary material. Algorithm 1 in the supplementary material shows the pseudo code for the optimization algorithm.

Step 3: enrichment scores and *p*-values for target

In order to estimate *p*-values for target gene sets from adaptive representations $\mathbf{Z}_s$ and $\mathbf{Z}_t$ (obtained in Step 2), we present the following three methods:

XGSEA-D: we train parameter $\boldsymbol{\alpha}$ in logistic regression by source:

$$\text{logit}(v_{s,i}):=\log\Big(\frac{v_{s,i}}{1-v_{s,i}}\Big)=\sum_{l=1}^{d}\boldsymbol{\alpha}(l)\mathbf{z}_s^i(l)+\epsilon_i,\ \text{fori}=1,\cdots,m,$$

and predict *p*-values for target using $\mathbf{z}_t^1,\cdots,\mathbf{z}_t^n$, directly.

XGSEA-E (enrichment scores and *p*-values): We already compute enrichment score $E_{s,i}^b$ of source gene set S_i at the *b*-th permutation ($b=0$ means no permutation) in Step 1. For any $b\in\{0,\cdots,B\}$, we train a linear regression model for enrichment scores from source gene sets:

$$E_{s,i}^b=\sum_{l=1}^{d}\boldsymbol{\beta}_b(l)\mathbf{z}_s^i(l)+\epsilon_i,\ \text{fori}=1,\cdots,m,$$

and predict enrichment scores $E_{t,j}^b$ for target gene sets using $\mathbf{z}_t^j$. We further estimate enrichment *p*-values for target gene sets by using **1c** of GSEA in Section 'Step 1: gene set enrichment analysis for source'.

XGSEA-E± (two enrichment scores and *p*-values): we train two linear regression models for positive and negative enrichment

scores, separately as follows:

$$E^b_{s,i} = \sum_{l=1}^{d} \gamma^+_b(l)\mathbf{z}^i_s(l) + \epsilon_i, \text{if} E^b_{s,i} \geq 0$$
$$E^b_{s,i} = \sum_{l=1}^{d} \gamma^-_b(l)\mathbf{z}^i_s(l) + \epsilon_i, \text{if} E^b_{s,i} < 0,$$

for $i = 1, \cdots, m$. We then compute distances $\|\mathbf{z}^i_t - \mathbf{z}^-_t\|_2$ and $\|\mathbf{z}^i_t - \mathbf{z}^+_t\|_2$, where $\mathbf{z}^+_t$ and $\mathbf{z}^-_t$ are the centers for $\mathbf{Z}_s$ with positive and negative enrichment scores, respectively, and then assign the positive model to target gene set $\mathbf{z}^i_t$ if the positive distance is smaller; otherwise the negative model. We predict the enrichment scores of target gene sets for each b and estimate p-values by using **1c** of GSEA in Section 'Step 1: Gene set enrichment analysis for source'.

More details of Step 3 can be found in Section 1.2 of the supplementary material.

Experiments

Comparison methods

We compared the XGSEA with three baseline methods, HM_1, HM_A and HM_O, which all directly map each target gene to source genes based on sequence homology, and estimate the enrichment p-value of target gene set T from enrichment p-value of particular source gene set S. These three baseline methods correspond to different strategies to generate S:

HM_1: S has a randomly chosen gene homologous to each gene in T (i.e. $|S| = |T|$).

HM_A: S has all genes homologous to each gene in T (i.e. $|S| \geq |T|$).

HM_O: S has, out of gene sets predefined by biological pathways and GO terms, the set with genes most overlapped with those in T.

Note that it is reasonable to compare our methods with the three baseline methods due to two reasons. First, there are no other advanced methods proposed for XGSEP. Second, the three baseline methods show fairly good performance in our experiments later. Since we propose three methods, thus we compared totally six methods : XGSEA-D, XGSEA-E, XGSEA-E±, HM_1, HM_A and HM_O.

Data sets

To evaluate the performance of the XGSEA, we need target expression data, so that we can compute ground truth enrichment p-values. We collected four gene expression data sets as below, where each data set consists of human (target) and another species (source: mouse or zebrafish) which share the same phenotype. Table 1 shows the statistics of the four data sets.

1. Embryonic development (human and mouse): the two data sets were collected from www.ncbi.nlm.nih.gov/geo with accessing number GSE44183. Both gene expression data sets were obtained from single cell RNA sequencing. In the human data set, there are 29 samples with 14 766 genes and seven embryonic development stages, including oocytes, pronucleus, zygote, 2-cell, 4-cell, 8-cell and morula. For the mouse, there are 17 samples with gene expression levels of 13 879 genes at six embryonic development stages, including oocytes, pronucleus,2-cell, 4-cell, 8-cell and morula. These data sets were used in a cross-species study [17] already, while this study is not on GSEA.
2. Brain cancer (human and mouse): the data sets of the two species were downloaded from GEO with accession number GSE45874 and GSE38591, respectively. Both data sets were measured by microarray. The human data set has 44 030 genes with six disease and six control samples, while the mouse data set has 9653 genes with four disease and four control samples. These data sets were also used in another cross-species study [18], while this study is also not on GSEA at all.
3. Ovarian cancer (human and mouse): the two microarray gene expression data sets were downloaded from GEO with accession number GSE6008 and GSE5987, respectively. The human data set has 21 188 genes with 13 mucinous ovarian tumors and four control samples, while the mouse data set has 45 101 genes with seven disease and four control samples. These data sets were also used in the cross-species study [18].
4. Melanomas (human and zebrafish): the microarray gene expression data sets of the two species were downloaded from GEO with accession number GSE83343 and GSE83399, respectively. The human data set has 42 346 genes with eight disease and four control samples, while the zebrafish data set has 13 620 genes with five disease and three control samples. These data sets were collected from two different studies [19, 20].

We then accessed Ensembl BioMart through http://www.ensembl.org/ [21] to retrieve homology relationships between 19 404 human and 19 614 mouse genes, and also 16 070 human and 18 324 zebrafish genes. The homology data from Ensembl is produced at the protein level rather than the DNA level by whole-genome alignments of vertebrate species [22, 23].

We showed two homology matrices between human and mouse (left) and between human and zebrafish (right), respectively, in Figure S1 (see the supplementary material). We can see that genes cannot be assigned in a simple manner of one-to-one correspondence.

We collected 674 human gene sets (pathways) from Reactome in Molecular Signatures Database (MSigDB), 2250 mouse gene sets from http://baderlab.org/GeneSets and 1550 zebrafish gene sets from http://bioinformatics.org/go2msig/. Note that for the melanomas data set, the genes in some gene sets do not have gene expression data, and thus GSEA could not be performed to obtain the ground truth p-values. After removing these gene sets, only 664 human gene sets are left in the melanomas data set.

Experimental setting

In our experiments, we took human species as the target species, and took mouse or zebrafish as the source species. We apply our XGSEA approach to predict the enrichment p-values for the 674 human pathway gene sets $\mathcal{T} = \{T_1, \cdots, T_n\}(n = 674)$, for embryonic development, brain, ovarian and melanomas, respectively. For the target gene sets $\mathcal{T} = \{T_1, \cdots, T_n\}$, we take the training source gene sets $\mathcal{S} = \{S_1, \cdots, S_n\}$ in the XGSEA, where S_i corresponds to T_i, meaning that each gene in S_i is homologous to one or more genes in T_i.

To sufficiently evaluate our XGESA methods, we predict enrichment p-values for target gene sets with three experimental settings. Note that the homology between two genes can be classified into four types: one-to-one, many-to-one, one-to-many and many-to-many, where one-to-one means only one gene in one side is homologous to only one gene in the other side. First level is for simple target gene sets $\mathcal{T}^{(1)} = \{T^{(1)}_1, \cdots, T^{(1)}_n\}$,

Table 1. Statistics of expression data and human gene sets ($\mathcal{T}^{(3)}$, where the cutoff p-values were 0.01 and 0.05 for embryonic development and the others, respectively)

Data set	Species	Gene expression Number of genes	Number of samples	Test sets in human Number of labels	Number of sets	Number of positive sets
Embryonic	Human	14 766	29	7	674	24
	Mouse	13 879	17	6		
Brain	Human	44 030	12	2	674	24
	Mouse	9 653	8	2		
Ovarian	Human	21 188	17	2	674	13
	Mouse	45 101	11	2		
Melanomas	Human	42 346	12	2	664	15
	Zebrafish	13 620	8	2		

where each $T_i^{(1)} \subset T_i$ only includes the target genes in T_i with label 'one-to-one'. For this case, each target gene g in set $T_i^{(1)}$ only has one homologous source gene, which does not have any other homologous target gene except g. The second case is for more complex target gene sets $\mathcal{T}^{(2)} == \{T_1^{(2)}, \cdots, T_n^{(2)}\}$, where each $T_i^{(2)} \subset T_i$ only includes the target genes in T_i with label 'one-to-one' and 'one-to-many'. For this case, each target gene g in set $T_i^{(2)}$ only has one homologous source gene, which may or may not have other homologous target genes besides g. The third case is the most complicated case with pathway target gene sets $\mathcal{T}^3 = \mathcal{T} = \{T_1, \cdots, T_n\}$, where the target genes may have any of four labels.

In summary, we consider three levels for $\mathcal{T}$, i.e. $\mathcal{T}^{(1)}$, $\mathcal{T}^{(2)}$ and $\mathcal{T}^{(3)}$, where $T_i^{(1)} \subset T_i^{(2)} \subset T_i^{(3)}$ ($i = 1, \ldots, n$):

$\mathcal{T}^{(1)}$ (simple): each set in $\mathcal{T}^{(1)}$ has one-to-one genes only. That is, target gene $g \in T_i^{(1)}$ has only one homologous source gene s, which has no other homologous target genes except g.

$\mathcal{T}^{(2)}$ (medium): each set in $\mathcal{T}^{(2)}$ has one-to-one or many-to-one genes. That is, target gene $g \in T_i^{(2)}$ has always only one homologous source gene s, which has one or more homologous target genes including g.

$\mathcal{T}^{(3)}$ (complex): each set in $\mathcal{T}^{(3)}$ target gene g may have one or more homologous source genes, and one of them s also may have one or more homologous target genes, including g.

To evaluate the methods sufficiently, we conducted bootstrapping on 674 human gene sets in $\mathcal{T}^{(3)}$ (or $\mathcal{T}^{(1)}$ and $\mathcal{T}^{(2)}$) as follows. We first sampled 674 gene sets with replacement from $\mathcal{T}^{(3)}$ for 20 times. For each sampling, the baseline methods and our methods were then used to predict p-values for the sampled gene sets that were used to calculate the evaluating measurements. Finally, we reported average results over the 20 trials. The parameters d and λ were chosen from $\{5, 10, 20, 30, 40, 50\}$ and $\{0.01, 0.1, 1, 10, 100\}$, respectively, to give the best performance under each experimental setting.

Evaluating the XGSEA methods by regression and classification measurements

The goal of the XGSEA methods is to predict p-values for target gene sets, and thus to further discover significant cross-species gene sets. We first used five regression measurements to evaluate the performance of our methods for p-value prediction, and then used one classification measurement (AUC area under ROC curve) to evaluate the performance of the XGSEA methods for discovering significant cross-species gene sets.

Evaluating regression performance

In this section, we first predicted p-values for the target gene sets in $\mathcal{T}^{(3)}$ of the four data sets, respectively, by the XGSEA methods and the baseline methods, and then compared the regression performance by using five regression measurements, including mean square errors (MSEs), mean absolute errors (MAEs), concordance index (CI), Pearson correlation and Cosine similarity. The definition of the five regression measurements can be found in Section 1.3 of the supplementary material. We first transformed p-values to negative log p-values ($-\log p$) and then computed the measurements.

We reported the bootstrapped MSEs for four real data sets under $\mathcal{T}^{(3)}$ by the six methods in Table 2. We also reported the results for the MAEs and the Cosine similarities in Table S1 and Table S2 in the supplementary material. Based on the results for MSEs and MAEs, the three XGSEA methods performed better than the three baseline methods for most cases, and the XGSEA-D method tends to perform the best for all the four data sets. As for the Cosine similarity, although the baseline methods are even better than the XGSEA-D and XGSEA-E for most cases, the XGSEA-E± performed the best among all the methods.

For the measurements of concordance index and Pearson correlation, our experiments showed that none of the six methods could obtain reasonable values if we consider all the tested gene sets. Thus, we took an alternative strategy to evaluate the performance by the measurements based on part of the gene sets. Based on the ground truth p-values for the target gene sets, we selected the k most significant and the k most insignificant gene sets. We then calculated the concordance indices and Pearson correlations between the negative logarithm of the ground truth p-values and the predicted p-values for the selected $2k$ gene sets. We reported the bootstrapped CIs and Pearson correlations by changing k from the set of $\{20 : 20 : 300\}$ for all competing methods on the four data sets with $\mathcal{T}^{(3)}$ in Figure 2 and Figure S2, respectively. The results show that the CIs or the Pearson correlations basically decrease along with increasing k for most of the methods and data sets, and the XGSEA methods could obtain higher CIs and Pearson correlations than the baseline methods for most cases.

Evaluating classification performance

The XGSEA methods predict p-values for cross-species gene sets. One important application of these p-values is to discover the significant gene sets, by setting up a significant level. Thus, we further evaluated the classification performance of the XGSEA methods by comparing AUCs (area under ROC curves) with the

Table 2. Bootstrapped mean square errors of six competing methods on four data sets under gene set $\mathcal{T}^{(3)}$

Data set	HM_1	HM_A	HM_O	XGSEA-D	XGSEA-E	XGSEA-E±
Embryonic	0.33	0.32	0.41	**0.30**	0.34	0.32
Brain	0.18	0.18	0.16	**0.15**	0.18	**0.15**
Ovarian	0.20	0.19	0.16	**0.12**	0.13	**0.12**
Melanomas	0.23	0.22	0.14	**0.13**	**0.13**	**0.13**

Note: The best in each row are in bold.

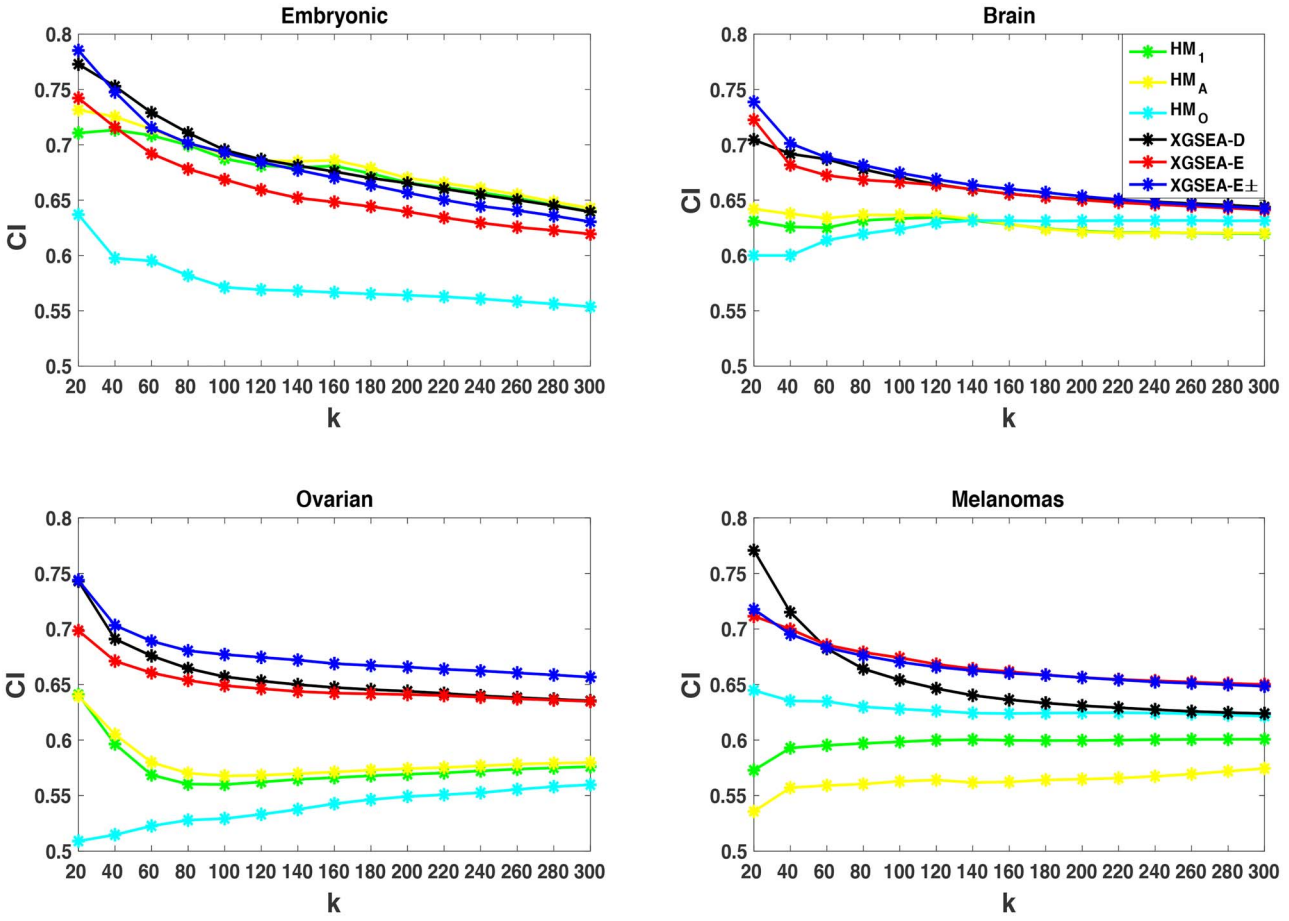

Figure 2. Bootstrapped CIs of six methods on four data sets ($\mathcal{T}^{(3)}$)for the selected 2k human gene sets (target), including the k most significant and the k most insignificant gene sets based on the ground truth p-values of human gene sets.

baseline methods. We first labeled the target gene sets to positive and negative ones by setting a cutoff p-value (significance level) for the ground truth p-values so that a gene set is a positive instance if the true p-value of this instance is lower than the cutoff; otherwise a negative. This means that we can obtain positive and negative gene sets by changing the cutoff p-value. We then could compute the AUC values by comparing the predicted p-values and the ground truth labels for the target gene sets. Note that the AUC values rely on the chosen cutoff p-values.

We first examined the performance of the competing methods by fixing the cutoff value for labeling. Table 3 shows bootstrapped AUCs under three different gene sets ($\mathcal{T}^{(1)}$, $\mathcal{T}^{(3)}$ and $\mathcal{T}^{(3)}$) by all six methods, fixing the cutoff at 0.01 for embryonic development and 0.05 for the other data sets. This table shows that the XGSEA significantly outperformed the baseline methods for most cases. For example, the XGSEA-E± achieved the best in 9 out of all 12 cases, followed by the XGSEA-E of three cases. Any naive method could neither be the best nor the second best in all cases, the difference from the best being statistically significant in t-test over 20 trials. Also the AUC of $\mathcal{T}^{(1)}$ was not necessarily higher than $\mathcal{T}^{(2)}$ (also $\mathcal{T}^{(3)}$), since each one-to-one homologous gene pair between two species is not necessarily the same gene, which would be prediction-wise harder than the case that the target and source gene sets share the same gene.

We then checked how the AUCs change along with cutoff p-values in the set {5e-1, 1e-1, 5e-2, 2.5e-2, 1e-2, 5e-3, 2.5e-3, 1e-3}. Figure 3 shows the bootstrapped AUCs of all methods on all four real data sets, under $\mathcal{T}^{(3)}$, by changing cutoff p-values. The bootstrapped AUCs basically increase as the cutoff p-values decrease, due to the decreasing number of the true positives. For most cutoff p-values, the XGSEA methods could obtain higher bootstrapped AUCs than the three baseline methods.

Table 3. Bootstrapped AUCs of six competing methods on four data sets and three target gene sets

Data set		HM_1	HM_A	HM_O	XGSEA-D	XGSEA-E	XGSEA-E±
Embryonic	$\mathcal{T}^{(1)}$	0.81 (6.59e-06)	0.81 (6.59e-06)	0.75 (1.48e-08)	0.86	0.80	**0.89**
	$\mathcal{T}^{(2)}$	0.80 (2.12e-06)	0.80 (2.12e-06)	0.74 (6.84e-09)	0.86	0.83	**0.89**
	$\mathcal{T}^{(3)}$	0.79 (1.58e-09)	0.80 (4.43e-09)	0.75 (3.73e-11)	0.87	0.83	**0.90**
Brain	$\mathcal{T}^{(1)}$	0.66 (3.14e-01)	0.66 (3.14e-01)	0.58 (1.14e-05)	0.60	**0.68**	0.67
	$\mathcal{T}^{(2)}$	0.59 (1.00e-04)	0.59 (1.00e-04)	0.57 (5.59e-06)	0.60	0.66	**0.67**
	$\mathcal{T}^{(3)}$	0.58 (1.75e-07)	0.60 (1.36e-05)	0.55 (2.64e-07)	0.61	0.63	**0.68**
Ovarian	$\mathcal{T}^{(1)}$	0.45 (2.53e-12)	0.45 (2.53e-12)	0.57 (1.65e-04)	0.67	0.64	**0.70**
	$\mathcal{T}^{(2)}$	0.56 (6.72e-09)	0.56 (6.72e-09)	0.50 (2.07e-08)	0.67	0.69	**0.75**
	$\mathcal{T}^{(3)}$	0.57 (5.60e-12)	0.61 (1.50e-07)	0.46 (6.60e-14)	0.65	0.70	**0.77**
Melanomas	$\mathcal{T}^{(1)}$	0.72 (3.65e-12)	0.72 (3.65e-12)	0.47 (2.10e-16)	0.84	**0.92**	0.87
	$\mathcal{T}^{(2)}$	0.63 (6.14e-05)	0.63 (6.14e-05)	0.48 (8.01e-14)	0.74	0.80	**0.81**
	$\mathcal{T}^{(3)}$	0.44 (1.74e-16)	0.44 (2.90e-15)	0.59 (4.68e-06)	0.64	**0.72**	0.71

Notes: The best and second best in each row are in bold and underlined, respectively. The p-value by t-test between the best and each corresponding baseline method is shown in brackets.

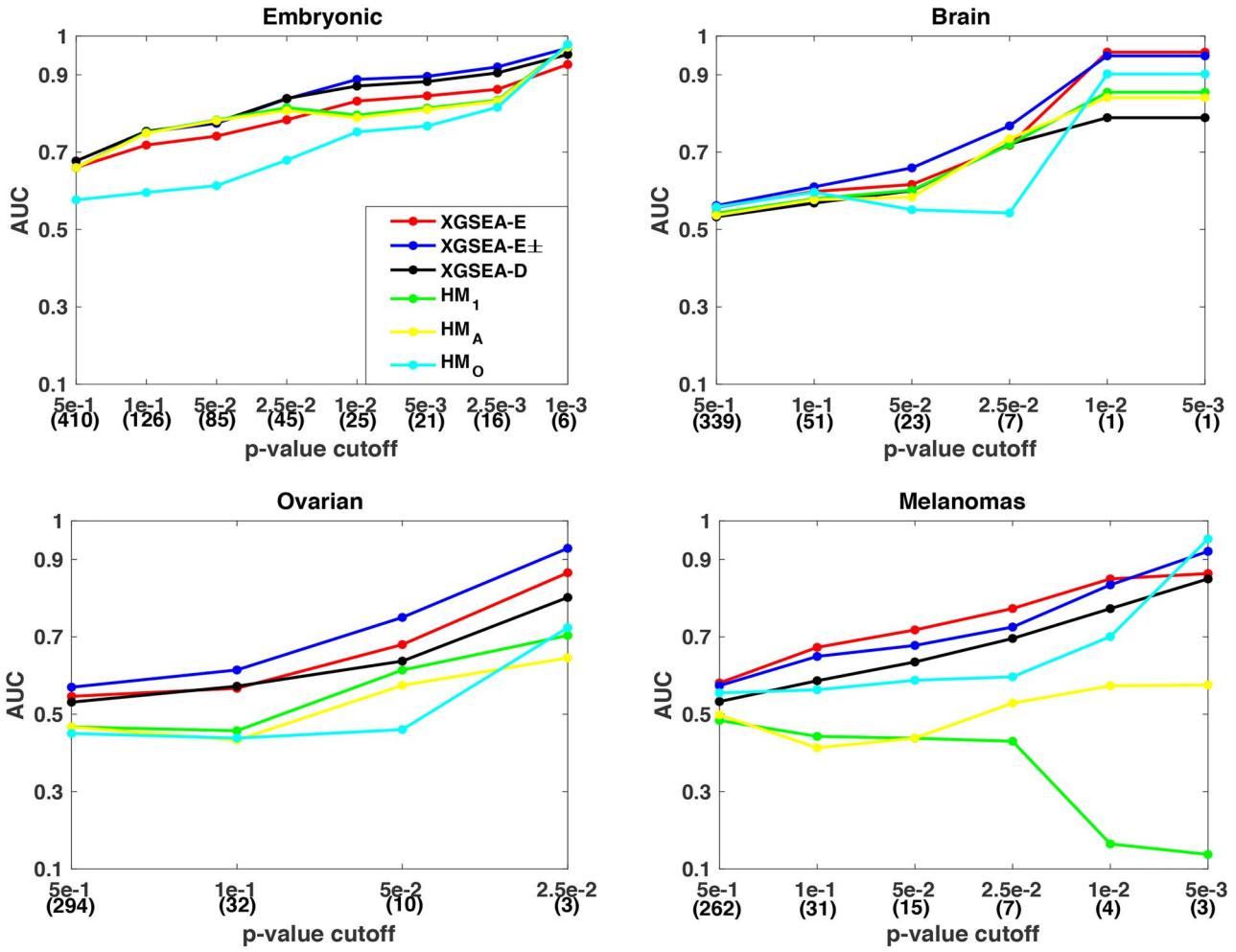

Figure 3. Bootstrapped AUCs on four data sets ($\mathcal{T}^{(3)}$) with changed cutoff p-values (and the corresponding numbers of true positives in the brackets).

Discussion

Prediction results without bootstrapping

In the above section, we evaluated our XGSEA methods by bootstrapping gene sets in $\mathcal{T}^{(1)}$, $\mathcal{T}^{(2)}$ or $\mathcal{T}^{(3)}$, and here we compared our methods with the baseline methods on the original 674 gene sets in $\mathcal{T}^{(3)}$ without bootstrapping.

We changed cutoff p-values from the same set in Section 'Evaluating classification performance'. Figure S3 (see the supplementary material) shows the AUCs of all six methods on all four data sets under the original $\mathcal{T}^{(3)}$ with varied cutoff p-values. We can see that Figure S3 is similar with the bootstrapped Figure 3, while the curves in Figure 3 seem more stabilized and differentiable due to the bootstrapped

Table 4. Eleven human pathways (with p-values) identified by the XGSEA-E for T-cell dysfunction and reprogramming

Pathway	p-value
Gene expression (Transcription)	0.03
A third proteolytic cleavage releases NICD	0.03
Signaling by NOTCH	0.03
Immune system	0.04
Signaling by NOTCH3	0.04
Signaling by NOTCH4	0.04
NOTCH2 activation and transmission of signal to the nucleus	0.04
Activated NOTCH1 transmits signal to the nucleus	0.04
Signaling by NOTCH2	0.04
Constitutive signaling by NOTCH1 HD+PEST domain mutants	0.04
Signaling by NOTCH1	0.04

procedure. Both figures show the performance advantage of the XGSEA methods over the three baseline methods for most cases.

Robustness against parameters

We first reported the most often best parameters in the bootstrapped results (Tables 2, 3, S1 and S2) in Table S3 (see the supplementary material), for the four data sets and four measurements including MSE, MAE, cosine similarity and AUC, respectively. We can see that different data sets tend to share similar best parameters, while different measurements may prefer different parameters. Under the measurements MSE and AUC, four data sets share the similar parameters $d = 5$ and $\lambda = 0.01$. MAE prefers the same λ and a larger $d = 50$, while cosine similarity prefers a larger $\lambda = 100$ and a larger $d = 50$.

We further examined the robustness of the XGSEA, regarding the parameter λ variation. Figure S4 (see the supplementary material) plotted the AUCs obtained by the XGSEA-E method versus λ in the set $\{1e\text{-}4, 1e\text{-}3, 1e\text{-}2, 1e\text{-}1\}$, under three original gene sets ($\mathcal{T}^{(1)}$, $\mathcal{T}^{(2)}$ and $\mathcal{T}^{(3)}$) of embryonic development and melanomas. The AUCs for the best baseline methods were also plotted. This figure shows that AUC of the XGSEA-E was rather stable within the given range, implying that the advantage over the baseline methods will be kept constantly.

Effect of similarity and homology on predictive performance

We examined the contribution of three types of gene set similarity, i.e. W_{ss}, W_{st} and W_{tt}, used in the XGSEA, by modifying the objective function in the formulation of the XGSEA. The objective function of the XGSEA is given by Equation (5), which has four terms, where the first term is the divergence and the last three terms are for W_{ss}, W_{st} and W_{tt}. We then generated four different variants of Equation (5), as follows:

MMD: only divergence, i.e. no terms on gene set similarity.
MMD+W: divergence and two terms on W_{ss} and W_{tt}.
MMD+B: divergence and the term on W_{st}.
MMD+WB: original objective function, i.e. Equation (5).

We applied these four variants to embryonic development data with target gene set $\mathcal{T}^{(3)}$. Table S4 (see the supplementary material) shows AUCs obtained with the cutoff (for p-values) of 0.01. From Table S4 (see the supplementary material), MMD+WB (i.e. original Equation (5)) achieved the best result for the XGSEA-E and the XGSEA-E±, and MMD was worst for them. This result implies that all gene set similarity contribute to the performance improvement.

We then evaluated the effect of sequence homology on predictive performance, by removing a certain amount of part in sequence homology matrix $\boldsymbol{M}$: being motivated by that less homology connectivity between two species would cause poorer performance.

In more detail, we first randomly chose a certain number of genes from the source and target gene sets, respectively, and kept only the part corresponding to these genes in $\boldsymbol{M}$. Practically, we used 50 500 and 5000 for this number of selected genes, resulting in three matrices: $\boldsymbol{M}_{50}$, $\boldsymbol{M}_{500}$ and $\boldsymbol{M}_{5000}$, respectively. Using each of the four sequence homology matrices (including original $\boldsymbol{M}$), we ran the XGSEA over embryonic development data under gene set $\mathcal{T}^{(3)}$ to predict enrichment p-values.

Table S4 (see the supplementary material) shows the performance results (AUC) of this experiment. The results show that the AUC was reduced by decreasing the number of randomly selected genes, while if the selected number is 5000, the performance was almost consistent with that of using the original $\boldsymbol{M}$, implying that interestingly 5000 genes might be good enough.

Case study: identifying human pathways for T-cell dysfunction and reprogramming from mouse ATAC-Seq

It is important for cancer immunotherapy to study the epigenetic regulation of T-cell dysfunction and therapeutic reprogrammability: a plastic dysfunctional state from which T-cells can be rescued, and a fixed dysfunctional state in which cells are resistant to reprogramming [24]. Identifying two (plastic or fixed) dysfunctional chromatin states, through which T-cells in tumours differentiate, would be very important to predict, for example, if a patient will respond to a therapy. Using GSE89308 of GEO on ATAC-Seq data of mouse, with 22 samples and the two chromatin states [24], we ran the XGSEA-E (B = 100 000, λ=0.01 and d=5) to identify human pathways out of 1960 Reactome pathways (downloaded from https://reactome.org/download-data).

Table 4 shows 11 human pathways identified by the XGSEA-E at the cutoff of 0.05, where the top, 'gene expression (transcription)', and the fourth 'immune system' are large pathways with 1367 and 2296 genes, respectively. Obviously due to important chromatin roles in transcription, 'gene expression (transcription)' is tightly related to the chromatin states. Also 'immune system' definitely plays important roles in T-cell dysfunction and reprogramming through a number of membrane proteins, such as CD38, CD101, CD30L, CD5, TCF1, IRF4, BCL2, CD44, PD1, LAG3 and CD62L [24].

The remaining nine pathways are all on Notch signaling pathways, which affect T cells in various ways. Notch signaling pathways play multiple essential roles in thymic T cell development and peripheral T cell differentiation [25]. For example, Delta-like ligand 4 (DLL4) interacts with Notch 1 to specify thymic T cell commitment during lymphocyte development. This Notch pathway regulates CD8+ T cells by directly upregulating mRNA expression of granzyme B and perforin to maintain memory T cells [26].

Furthermore, the Notch pathway plays an important role in antitumor immunity. CD8+ T cell-specific Notch2 deletion impairs antitumor immunity, whereas the stimulation of the

Notch pathway can increase tumor suppression. Ezh2, a suppressor of the Notch pathway, regulates effector T-cell polyfunctionality and survival by targeting the Notch signaling pathway [27]. Down regulation of Ezh2 could elicit poor antitumor immunity.

Besides, Delta-like 1-mediated Notch signaling enhances the conversion of human memory CD4 T cells into FOXP3-expressing regulatory T cells [28]. These facts support the reliability of the pathways identified by the XGSEA.

On the other hand, we ran a naive approach, HM_A, over the same data, under the cutoff of 0.05, resulting in 20 pathways showed in Table S5 (see the supplementary material). Although the number of pathways is larger than Table 4, these 20 pathways were diverse and less connected to the chromatin states, such as only two being related to Notch signaling pathways.

Conclusion

We have defined XGSEP for promoting GSEA on species with scarce expression data, and proposed the XGSEA with three steps, which can be simply: (1) GSEA, (2) domain adaptation and (3) regression. Our empirical supervised validation, including regression and classification, over four real data sets revealed that the XGSEA outperformed three baseline approaches under the measurements MSE, MAE, CI, Pearson correlation, cosine similarity and AUC for most cases. Particularly, the advantage was also proved statistically by bootstrapping and *t*-test. In the case study, mouse ATAC-Seq expression data are used to identify significant human pathways for T-cell dysfunction and reprogramming. The XGSEA found rather general two pathways related with gene expression (transcription) and immune system, as well as nine Notch signal-related pathways, all being convincing, especially compared with pathways found by a baseline approach. Our XGSEA methods also have potential applications on minor species once the homology information between the minor species and one major species is available in the future.

There are two characteristics for the proposed XGSEA methods. On one hand, the XGSEA methods apply domain adaptation to reduce the gap of samples from different species. It's common that different species samples are drawn from distinguish distributions. To shorten the distance of two sample distributions between different species for better downstream analysis, the XGSEA methods take domain adaptation as an effective strategy to search a latent subspace, in which two distributions are as close as possible. When the subspace found and new representations of samples computed for different species, traditional machine learning methods, such as linear regression, can be used to train in one species and test in the other one. Besides, compared with three naive methods, which project linearly the target gene set to the corresponding source one based on homology network directly, better new representations of source and target gene sets are computed by the XGSEA methods with capturing the complex relationship among gene sets from different species. On the other hand, in the source species, the XGSEA methods learn a regression model with enrichment scores, instead of training a model with *p*-values directly. The power of this strategy is shown by the performance of the XGSEA methods on the results in Table 3, which further indicating that training on enrichment scores can capture more information than training on *p*-values for cross-species gene sets enrichment analysis.

Improvement of the XGSEA would be definitely interesting future work. It would be worth working on exploring a better variation on each of the three steps of the XGSEA: Step 1 can be generalized or focused on another statistical problem. Exploring more efficient, robust domain adaptation would be interesting future work for Step 2. Reasonably in Step 3, we can consider more sophisticated regression models. The most key point of the XGSEA is Step 2, i.e. domain adaptation, which would be useful for other problems between two species, such as genome wide association studies between a well- and the other less-sequenced species. This direction of applying domain adaptation to various problems would be also promising future work. On the statistical side, we could also further consider the problem of multiple testing and controlling the false discovery rate or family wise error rate, which have been well studied in regular GSEA.

Data Availability

All data used in our experiments is available and the details of data source can be found in Data sets Section.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

This work was supported by the National Natural Science Foundation of China (No.11631012 and No.11471256).

Conflict of Interest

The authors declare no conflict interest.

References

1. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* 2005;**102**(43):15545–50.
2. Stuart JM. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003;**302**(5643):249–55.
3. Zheng-Bradley X, Rung J, Parkinson H, *et al*. Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol* 2010;**11**(12):R124.
4. Debry RW, Seldin MF. Human/mouse homology relationships. *Genomics* 1996;**33**(3):0–351.
5. Liao BY, Zhang J. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* 2008;**105**(19):6987–92.
6. Mestas J, Hughes CCW. Of mice and not men: differences between mouse and human immunology. *J Immunol* 2004;**172**(5):2731–8.
7. Geifman N, Rubin E. The mouse age phenome knowledgebase and disease-specific inter-species age mapping. *Plos One* 2013;**8**.
8. Beura LK, Hamilton SE, Bi K, *et al*. Normalizing the environment recapitulates adult human immune traits in laboratory mice. *Nature* 2016;**532**:512–6.
9. Bugelski PJ, Martin PL. Concordance of preclinical and clinical pharmacology and toxicology of therapeutic monoclonal antibodies and fusion proteins: cell surface targets. *British Journal of Pharmacol* 2012;**166**.

10. Hünig and Thomas. The storm has cleared: lessons from the cd28 superagonist tgn1412 trial. *Nat Rev Immunol* 2012;**12**:317–8.
11. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Transaction on Knowledge and Data Engineering* 2010;**22**(10):1345–59.
12. Huang J, Gretton A, Borgwardt KM, *et al*. Correcting sample selection bias by unlabeled data. In: Schölkopf B, Platt J, Hoffman T (eds). *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press, 2006, 601–8.
13. Pan SJ, Kwok JT, Yang Q. Transfer learning via dimensionality reduction. In: *AAAI 2008, pages 677–682*, 2008.
14. Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, *et al*. A kernel two-sample test. *Journal of Machine Learning Research*, **13**:723–73, 2012.
15. Baktashmotlagh M, Harandi M, Salzmann M. Distribution-matching embedding for visual domain adaptation. *Journal of Machine Learning Research* 2016;**17**(108):1–30.
16. Djordjevic D, Kusumi K, Ho JWK. Xgsa: a statistical method for cross-species gene set analysis. *Bioinformatics* 2016;**32**(17):i620–8.
17. Sun J, Jiang Z, Tian X, *et al*. A cross-species bi-clustering approach to identifying conserved co-regulated genes. *Bioinformatics* 2016;**32**(12):i137–46.
18. Normand R, Du W. Found in translation: a machine learning model for mouse-to-human inference. *Nat Methods* 2018;**15**:1067–73.
19. FV Filipp CL, Boiko AD. Cd271 is a molecular switch with divergent roles in melanoma and melanocyte development. *Sci Rep* 2019;**9**(1):7696.
20. Venkatesan AM, Vyas R, Gramann AK, *et al*. Ligand-activated bmp signaling inhibits cell differentiation and death to promote melanoma. *J Clin Invest* 2018;**128**(1):294–308.
21. Durinck S, Spellman PT, Birney E, *et al*. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nat Protoc* 2009;**4**(8):1184–91.
22. Clamp M, Andrews D, Barker D, *et al*. Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res* 2002;**31**:38–54.
23. Flicek P. Ensembl 2014. *Nucleic Acids Res* 2014;**42**(Issue D1):D749–55.
24. Philip M, Fairchild L, Sun L, *et al*. Chromatin states define tumour-specific t cell dysfunction and reprogramming. *Nature* 2017;**545**(7655):452–6.
25. Freddy R, Robson MH, Fabienne TC. Regulation of innate and adaptive immunity by notch. *Nat Rev Immunol* 2013;**13**(6):427–37.
26. Tsukumo S-i, Yasutomo K. Regulation of cd8+ t cells and antitumor immunity by notch signaling. *Front Immunol* 2018;**9**:101.
27. Ende Z, Tomasz M, Ilona K. Et. al. cancer mediates effector t cell dysfunction by targeting microrna s and ezh2 via glycolysis restriction. *Nat Immunol* 2016;**17**(1):95–103.
28. Mota C, Nunes-Silva V, Pires AR, *et al*. Delta-like 1-mediated notch signaling enhances the in vitro conversion of human memory cd4 t cells into foxp3-expressing regulatory t cells. *J Immunol* 2014;**193**(12):5854–62.