# XGSEA: CROSS-species Gene Set Enrichment Analysis via domain adaptation

Menglan Cai[1], Canh Hao Nguyen[2], Hiroshi Mamitsuka[2,3], Limin Li[1*]

**1** School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, 710049, China,
**2** Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, 6110011, Japan,
**3** Department of Computer Science, Aalto Unviersity, Espoo, Finland

* liminli@mail.xjtu.edu.cn

## Abstract

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. More importantly, gene expression could not be measured under specific conditions for human, due to high healthy 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). For XGSEP, we propose XGSEA (Cross-species Gene Set Enrichment Analysis), 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 XGSEA by using four real data sets under various settings, proving that XGSEA significantly outperformed three baseline methods. A case study of identifying important human pathways for T cell dysfunction and reprogramming from mouse ATAC-Seq data further confirmed the reliability of XGSEA. Source code is available through https://github.com/LiminLi-xjtu/XGSEA

## Author summary

Gene set enrichment analysis (GSEA) is a powerful tool in the gene sets differential analysis given a ranked gene list. GSEA requires complete data, gene expression with phenotype labels. However,gene expression could not be measured 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 no availability of gene expression leads to more challenging problem, CROSS-species Gene Set Enrichment Problem (XGSEP), in which enrichment significance (on a phenotype) of a given gene set of a species (target, say human) is predicted by using gene expression measured under the same phenotype of the other species (source, say mouse). In this work, we propose XGSEA (Cross-species Gene Set Enrichment Analysis) for XGSEP, with three steps of: 1) GSEA; 2) domain adaptation; and 3) regression. The results of four real data sets and a case study indicate that XGSEA significantly outperformed three baseline methods and confirmed the reliability of XGSEA.

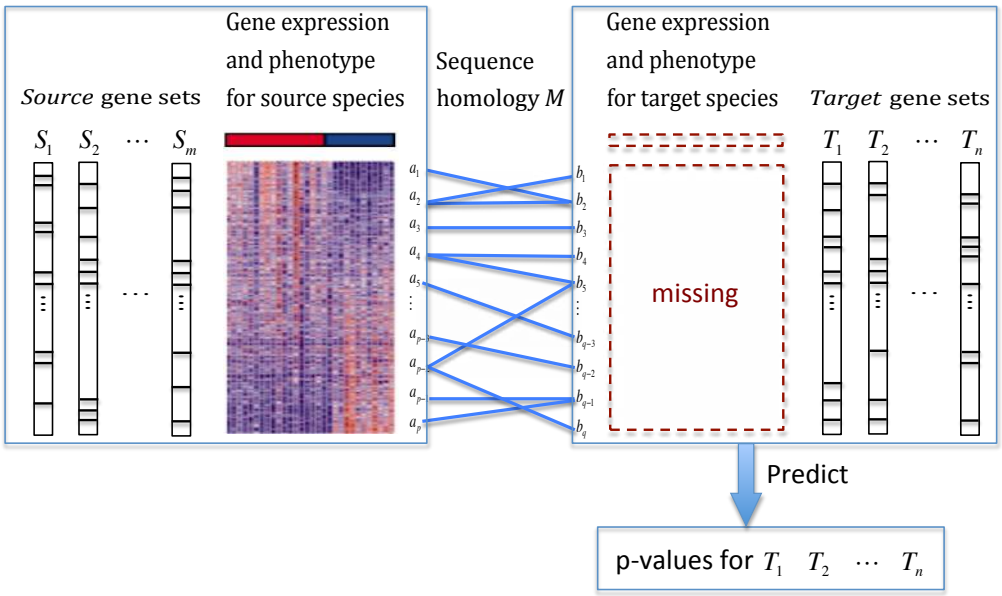


**Fig 1.** XGSEP: Cross-species gene set enrichment problem, to predict enrichment $p$-values of target gene sets by using source gene sets, gene expression data and sequence homology between target and source genes.

# 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 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 has strong bias in measured organisms or species. Model organisms, such as *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). Fig 1 shows a schematic picture of the problem setting of 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 $p$-values) 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 $p$-value 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 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: $\boldsymbol{X}_s \overset{\boldsymbol{P}_s}{\to} \boldsymbol{Z}_s$ and $\boldsymbol{X}_t \overset{\boldsymbol{P}_t}{\to} \boldsymbol{Z}_t$, 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 preserving 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

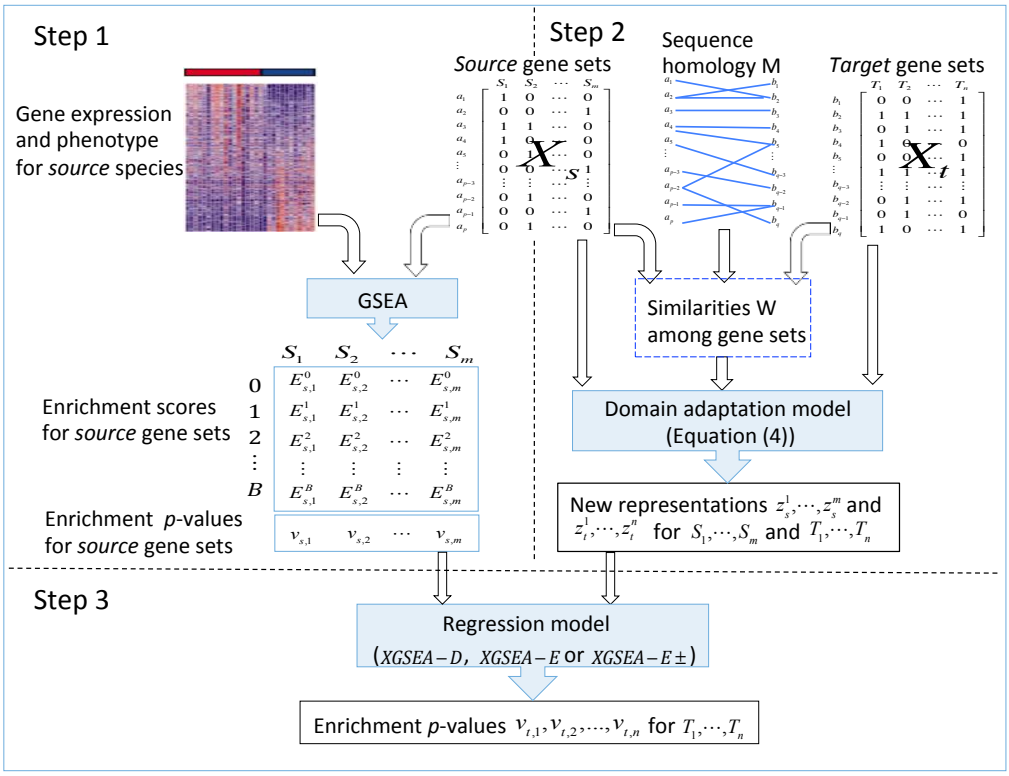


**Fig 2.** Flow chart of XGSEA: we 1) compute $B$+1 enrichment scores and $p$-values for each source gene set by GSEA, where $B$ is the number of permutation, 2) obtain new representations for all source and target gene sets by domain adaptation, and 3) predict enrichment $p$-values for target gene sets by a regression model based on the new representations.

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 XGSEA becomes feasible over regular gene annotation matrices. 3) We empirically validated XGSEA by using four different real phenotypes with expression data. The experimental results showed that XGSEA significantly outperformed three baseline methods under various settings. The advantage of 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

To the best of our knowledge, there are 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. On the other hand, XGSEP estimates enrichment scores through expression data sets obtained under the same phenotype (see Fig 1, though the target expression is assumed to be missing). Thus XGSA is totally different from XGSEP.

## 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 $\boldsymbol{M} \in \mathbb{R}^{p \times q}$ be a binary matrix of sequence homology, where the $(i, j)$-element $\boldsymbol{M}(i, j)$ is 1 if source gene $a_i$ is homologous to target gene $b_j$; otherwise zero. Suppose that we have gene expression matrix $\boldsymbol{G}_s$ with phenotype vector $\boldsymbol{y}_s$ for source genes only, meaning that we can run GSEA over $\boldsymbol{G}_s$ and $\boldsymbol{y}_s$ 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. Thus we have a binary matrix (which we call *annotation matrix*) between $A$ (for rows) and $\mathcal{S}$ (for columns), where 1 means that the corresponding gene is in a gene set; otherwise zero. This can be also for the target side. Let $\boldsymbol{X}_s = [\boldsymbol{x}_s^1, \cdots, \boldsymbol{x}_s^m] \in \mathbb{R}^{p \times m}$ be the annotation matrix for source gene sets $S_1, \cdots, S_m$, where the $i$-th element of $\boldsymbol{x}_s^j$ is 1 if gene $a_i$ is in gene set $S_j$; otherwise zero. Similarly, let $\boldsymbol{X}_t = [\boldsymbol{x}_t^1, \cdots, \boldsymbol{x}_t^n] \in \mathbb{R}^{q \times n}$ be the annotation matrix for target, where the $i$-th element of $\boldsymbol{x}_t^j$ is 1 if gene $b_i$ is in gene set $T_j$; otherwise zero. Then the problem, XGSEP standing for CROSS-species Geneset Enrichment Problem, is, given $\boldsymbol{G}_s$, $\boldsymbol{y}_s$, $\boldsymbol{X}_s$, $\boldsymbol{X}_t$ and $\boldsymbol{M}$, to estimate the enrichment $p$-value of each gene set in $\mathcal{T}$ with respect to the same phenotype of $\boldsymbol{y}_s$. We propose our method XGSEA, standing for CROSS-species Gene Set Enrichment Analysis, to solve XGSEP by using three steps. Fig 2 shows a schematic picture of the three-step procedure of XGSEA. Below we will explain each of these three steps in detail.

## Step 1: Gene set enrichment analysis for source

Since gene expression $\boldsymbol{G}_s$ and phenotype $\boldsymbol{y}_s$ 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 $\mathrm{H}_0^{s,i}$: gene set $S_i$ has no association with phenotype $\boldsymbol{y}_s$ (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 $\boldsymbol{G}_s$ and phenotype $\boldsymbol{y}_s$.

**1b.** Permute the entries in $\boldsymbol{y}_s$ 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: $\mathrm{E}_{\mathrm{NULL}}$ 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 $\mathrm{E}_{\mathrm{NULL}}$ 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.

### Formulating the objective function

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

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

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

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

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

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

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

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

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

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

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

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

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,

$$\begin{aligned} \boldsymbol{W}_{ss}(i,j) = &\quad \min\{\tfrac{|S_i \cap S_j|}{|S_i|}, \tfrac{|S_i \cap S_j|}{|S_j|}\} \quad i,j \in \{1,\cdots,m\}; \\ \boldsymbol{W}_{tt}(i,j) = &\quad \min\{\tfrac{|T_i \cap T_j|}{|T_i|}, \tfrac{|T_i \cap T_j|}{|T_j|}\} \quad i,j \in \{1,\cdots,n\}; \\ \boldsymbol{W}_{st}(i,j) = &\quad \min\{\tfrac{|T_j \cap \widetilde{S_i}|}{|T_j|}, \tfrac{|S_i \cap \widetilde{T_j}|}{|S_i|}\} \quad i \in \{1,\cdots,m\}, \\ &\qquad\qquad\qquad\qquad\qquad\quad j \in \{1,\cdots,n\}, \end{aligned} \tag{3}$$

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 projection should be smooth over homologous similarity matrix $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_{ss} & \boldsymbol{W}_{st} \\ \boldsymbol{W}_{st}^T & \boldsymbol{W}_{tt} \end{bmatrix}$.

Thus entirely divergence $\mathcal{D}$ in (1) should be minimized, being regularized by the smoothness (of the projection) over similarity matrix $\boldsymbol{W}$. Overall the objective function can be given as follows:

$$\begin{aligned} \min_{\boldsymbol{P}_s^T\boldsymbol{P}_s = \boldsymbol{P}_t^T\boldsymbol{P}_t = \boldsymbol{I}} &\mathcal{D}(\boldsymbol{P}_s^T\boldsymbol{X}_s, \boldsymbol{P}_t^T\boldsymbol{X}_t) + \lambda(\frac{1}{2}\sum_{i,j=1}^{m}\boldsymbol{W}_{ss}(i,j)\|\boldsymbol{z}_s^i - \boldsymbol{z}_s^j\|_2^2 \\ &+ \sum_{i,j=1}^{m,n}\boldsymbol{W}_{st}(i,j)\|\boldsymbol{z}_s^i - \boldsymbol{z}_t^j\|_2^2 + \frac{1}{2}\sum_{i,j=1}^{n}\boldsymbol{W}_{tt}(i,j)\|\boldsymbol{z}_t^i - \boldsymbol{z}_t^j\|_2^2) \end{aligned} \tag{4}$$

**Table 1.** Pseudocode of the optimization algorithm in Step 2 of XGSEA

| **Algorithm** |
|---|
| **Inputs.** Source annotation matrix $\boldsymbol{X}_s$, Target annotation matrix $\boldsymbol{X}_t$, Sequence homology $\boldsymbol{M}$ |
| **Parameters.** Regularization $\lambda$ and embedding dimension $d$. |
| **Outputs.** New representations $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_t$ |
| 1. Construct $\boldsymbol{L}$ by (2), $\boldsymbol{W}$ by (3) and $\boldsymbol{X}$; |
| 2. Compute $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{F}\boldsymbol{X}^T$, where $\boldsymbol{F}$ is Laplacian matrix; |
| 3. Solve problem (5) for $\boldsymbol{P}$ with the initial $[\boldsymbol{I}; \boldsymbol{I}]/\sqrt{2}$; |
| 4. Compute $\boldsymbol{Z} = [\boldsymbol{Z_s}\ \ \boldsymbol{Z_t}] = \boldsymbol{P}^T\boldsymbol{X}$. |

### Optimization on Grassman manifold

We can use

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}_s \\ \boldsymbol{P}_t \end{bmatrix} \in \mathbb{R}^{r\times d}, \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_s & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_t \end{bmatrix} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{r\times N},$$

where $r = p + q$, and $N = m + n$ to write $\boldsymbol{Z} = [\boldsymbol{Z}_s\ \ \boldsymbol{Z}_t] = \boldsymbol{P}^T\boldsymbol{X} \in \mathbb{R}^{d\times N}$. Then the first term in (4) can be written as

$$\mathcal{D}(\boldsymbol{P}_s^T\boldsymbol{X}_s, \boldsymbol{P}_t^T\boldsymbol{X}_t) = \text{trace}(\boldsymbol{K}_{\boldsymbol{P}}\boldsymbol{L}),$$

where $\boldsymbol{K}_{\boldsymbol{P}}(i,j) = \exp(-\frac{\|\boldsymbol{P}^T\boldsymbol{x}_i - \boldsymbol{P}^T\boldsymbol{x}_j\|_2^2}{\sigma}), i, j = 1, \cdots, N$, and $\boldsymbol{L}$ is defined in (2). Note that $\boldsymbol{K}_{\boldsymbol{P}}$ depends on $\boldsymbol{P}$.

Also the regularization term in (4) can be written as

$$\lambda\text{trace}(\boldsymbol{Z}^T\boldsymbol{F}\boldsymbol{Z}) = \lambda\text{trace}(\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{F}\boldsymbol{X}^T\boldsymbol{P}) = \lambda\text{trace}(\boldsymbol{P}^T\boldsymbol{G}\boldsymbol{P}),$$

where $\boldsymbol{F} = \boldsymbol{D} - \boldsymbol{W}$ is a Laplacian matrix, $\boldsymbol{D}$ is a diagonal matrix with $\boldsymbol{D}_{ii} = \sum_j \boldsymbol{W}_{ij}$, and $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{F}\boldsymbol{X}^T$.

The constraints can be changed from $\boldsymbol{P}_s^T\boldsymbol{P}_s = \boldsymbol{P}_t^T\boldsymbol{P}_t = \boldsymbol{I}$ to $\boldsymbol{P}_s^T\boldsymbol{P}_s + \boldsymbol{P}_t^T\boldsymbol{P}_t = \boldsymbol{I}$ which can avoid that all samples collapse to the origin. Finally (4) can be transformed into an easily understandable form:

$$\min_{\boldsymbol{P}^T\boldsymbol{P}=\boldsymbol{I}} \text{trace}(\boldsymbol{K}_{\boldsymbol{P}}\boldsymbol{L}) + \lambda\text{trace}(\boldsymbol{P}^T\boldsymbol{G}\boldsymbol{P}). \tag{5}$$

Let $f(\boldsymbol{P}) = \text{trace}(\boldsymbol{K}_P\boldsymbol{L}) + \lambda\text{trace}(\boldsymbol{P}^T\boldsymbol{G}\boldsymbol{P})$. The optimization problem $\min_{\boldsymbol{P}^T\boldsymbol{P}=\boldsymbol{I}} f(\boldsymbol{P})$ can be solved on the Grassmann manifold, with all linear $d$-dimensional subspaces in $\mathbb{R}^p$, since optimizing $f(\boldsymbol{P})$ is not affected by any orthogonal transformation of $\boldsymbol{P}$. We use the conjugate gradient (CG) algorithm on the Grassmann manifold [17] to solve the optimization problem $\min_{\boldsymbol{P}^T\boldsymbol{P}=\boldsymbol{I}} f(\boldsymbol{P})$. The key step is to compute partial derivative $\partial_{\boldsymbol{P}} f(\boldsymbol{P})$, which is used for computing gradient $\nabla_{\boldsymbol{P}} f(\boldsymbol{P})$ of $f$ on the manifold at the current estimate $\boldsymbol{P}$ by $\nabla_{\boldsymbol{P}} f(\boldsymbol{P}) = \partial_{\boldsymbol{P}} f(\boldsymbol{P}) - \boldsymbol{P}\boldsymbol{P}^T\partial_{\boldsymbol{P}} f(\boldsymbol{P})$. The search direction is determined at each step by combining the previous search direction with $\nabla_{\boldsymbol{P}} f(\boldsymbol{P})$, and in this direction, a line search along the geodesic at the current estimated $\boldsymbol{P}$ is performed. Note that partial derivative $\partial_{\boldsymbol{P}} f(\boldsymbol{P})$ at the current $\boldsymbol{P}$ can be obtained as follows

$$\begin{aligned} \partial_{\boldsymbol{P}} f(\boldsymbol{P}) &= \sum_{i,j} \boldsymbol{L}(i,j)\partial_{\boldsymbol{P}}\boldsymbol{K}_{\boldsymbol{P}}(i,j) + 2\lambda\boldsymbol{G}\boldsymbol{P}, \\ &= -2\sum_{i,j} \frac{\boldsymbol{K}_{\boldsymbol{P}}\boldsymbol{P}(i,j)\boldsymbol{L}(i,j)}{\sigma}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T\boldsymbol{P} + 2\lambda\boldsymbol{G}\boldsymbol{P}. \end{aligned}$$

Table 1 shows a pseudocode of the optimization algorithm of Step 2.

**Reducing computational complexity**

The computational complexity of the above algorithm for solving the optimization problem (5) is $O(N^2 + r^2 + Nrd)$. The total number of either human or mouse genes is large, leading to $r(= p + q) \gg N(= m + n)$. This (large $r$) problem can be a bottleneck for our algorithm, and thus we need to reduce the $r$-related part of this complexity. For this purpose we propose an approach, which uses QR decomposition, by which the computational complexity is reduced to $O(N^2)$. Below, we will explain more detailed manner of our approach.

We first use QR decomposition: $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q} \in \mathbb{R}^{r\times N}$ is an orthonormal matrix and $\boldsymbol{R} \in \mathbb{R}^{N\times N}$ is an upper diagonal matrix. By introducing $\tilde{\boldsymbol{P}} = \boldsymbol{Q}^\mathsf{T}\boldsymbol{P} \in \mathbb{R}^{N\times d}$, the objective function in (5) can be transformed as follows:

$$f(\boldsymbol{P}) = g(\tilde{\boldsymbol{P}}) = \text{trace}(\boldsymbol{K}_{\tilde{P}}\boldsymbol{L}) + \lambda\text{trace}(\tilde{\boldsymbol{P}}^\mathsf{T}\tilde{\boldsymbol{G}}\tilde{\boldsymbol{P}}),$$

where $\tilde{\boldsymbol{G}} = \boldsymbol{R}\boldsymbol{F}\boldsymbol{R}^\mathsf{T}$, and $\boldsymbol{K}_{\tilde{P}}$ can be obtained using $\boldsymbol{R}$ since $\boldsymbol{Z} = \tilde{\boldsymbol{P}}^\mathsf{T}\boldsymbol{R} = \boldsymbol{P}^\mathsf{T}\boldsymbol{X}$. Thus we can first solve a small-scale optimization problem of $\tilde{\boldsymbol{P}}$, i.e.

$$\min_{\tilde{\boldsymbol{P}}^T\tilde{\boldsymbol{P}}=I} g(\tilde{\boldsymbol{P}}) \tag{6}$$

and then obtain the projections $\boldsymbol{Z} = \boldsymbol{P}^\mathsf{T}\boldsymbol{X} = \tilde{\boldsymbol{P}}^\mathsf{T}\boldsymbol{R}$. Note that now solving (6) by the above algorithm on the Grassmann manifold needs the computational complexity of only $O(N^2)$.

## Step 3: Enrichment scores and $p$-values for target

The final step in XGSEA is to estimate the $p$-values for the target gene sets, based on the adaptive representations $\mathbf{Z}_s$ and $\mathbf{Z}_t$ for the source and target gene sets obtained in the above step. One idea is to regress $p$-values on the new representations of the gene sets by logistic regression (XGSEA-D). However, the resulting $p$-values may not obey the principle of null hypothesis. By the principle of null hypothesis, $p$-values is defined as the probability of obtaining the same or more extreme statistics than the observation under null hypothesis, and should be determined by the observed enrichment scores and the null distribution of enrichment scores. Thus another idea is to first predict the observed and null enrichment scores, and then determine $p$-values by its definition. This means that we have one more step to reach $p$-values from the enrichment scores $\boldsymbol{E}_t$. In detail, the second idea to predict $p$-value for target gene set $T_j$ is that we first predict the enrichment scores $B + 1$ enrichment scores $E^0_{t,j}, \cdots, E^0_{t,j}$ for target gene set $T_j$ and then estimate $p$-value $v_{t,j}$ by step **1c** in the section of Step1. Based on this idea, we propose to use two regression methods (XGSEA-E and XGSEA-E±). We explain the three methods as follow.

- XGSEA-D: Logistic regression on $p$-values
  We first train the regression parameters $\alpha$ by source gene sets in the following logistic regression model

$$logit(v_{s,i}) := \log(\frac{v_{s,i}}{1 - v_{s,i}}) = \sum_{l=1}^{d} \boldsymbol{\alpha}(l)\boldsymbol{z}^{\boldsymbol{i}}_{\boldsymbol{s}}(l) + \epsilon_i,$$

  for $i = 1, \cdots, m$, and then predict $p$-values for the target gene sets by

$$logit(v_{t,j}) = \sum_{l=1}^{d} \boldsymbol{\alpha}(l)\boldsymbol{z}^{\boldsymbol{j}}_{\boldsymbol{t}}(l), \text{for } j = 1, \cdots, n.$$

  Finally we can obtain the $p$-values of the target species by the following,

$$v_{t,j} = \frac{1}{1 + \exp(-logit(v_{t,j})}, \text{for } j = 1, \cdots, n.$$

- XGSEA-E: linear regression on enrichment scores
  Note that we have computed $E^b_{s,i}$, the enrichment score 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 regress the enrichment scores on the new representations of gene sets. The parameter in the regression model can be learnt based on the source gene sets as follows

$$E^b_{s,i} = \sum_{l=1}^{d} \boldsymbol{\beta}_b(l) \boldsymbol{z^i_s}(l) + \epsilon_i,$$

  for $i = 1, \cdots, m$, and then predict the enrichment scores for target gene sets by

$$E^b_{t,j} = \sum_{l=1}^{d} \boldsymbol{\beta}_b(l) \boldsymbol{z^j_t}(l), \text{for } j = 1, \cdots, n, b = 0, \cdots, B.$$

  Finally, we can compute the enrichment $p$-values for target gene sets by using step **1c** in the section of Step1.

- XGSEA-E±: linear regression on positive and negative enrichment scores, respectively
  Similar to XGSEA-E, we predict $p$-values by first estimating enrichment scores for target source gene sets. Different with XGSEA-E, in XGSEA-E± we learn two linear regression models for positive and negative enrichment scores, separately as follows

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

  for $i = 1, \cdots, m$. The parameters $\boldsymbol{\gamma}^+$ and $\boldsymbol{\gamma}^-$ are learnt by training the source gene sets, and then used to predict enrichment scores for target gene sets by

$$E^b_{t,j} = \begin{cases} \sum_{l=1}^{d} \boldsymbol{\gamma}^+_b(l) \boldsymbol{z^j_t}(l), & \text{if } \|\boldsymbol{z^j_t} - \boldsymbol{z}^+_t\| \leq \|\boldsymbol{z^j_t} - \boldsymbol{z}^-_t\| \\ \sum_{l=1}^{d} \boldsymbol{\gamma}^-_b(l) \boldsymbol{z^j_t}(l), & \text{if } \|\boldsymbol{z^j_t} - \boldsymbol{z}^+_t\| > \|\boldsymbol{z^j_t} - \boldsymbol{z}^-_t\| \end{cases}$$

  where $\boldsymbol{z}^+_t$ and $\boldsymbol{z}^-_t$ are the centers for $\boldsymbol{Z}_s$ with positive and negative enrichment scores respectively, and $j = 1, \cdots, n, b = 0, \cdots, B$.

# Results

## Comparison methods

Since there are no existing methods for XGSEP, we compared XGSEA with three simpler methods, $\text{HM}_1$, $\text{HM}_A$, and $\text{HM}_O$, 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 $p$-values of particular source gene set $S$. However these three baseline methods take different strategies to generate $S$:

$\text{HM}_1$: $S$ has a randomly chosen gene homologous to each gene in $T$ (i.e. $|S| = |T|$).

$\text{HM}_A$: $S$ has all genes homologous to each gene in $T$ (i.e. $|S| \geq |T|$).

$\text{HM}_O$: $S$ has, out of gene sets predefined by biological pathways and GO terms, the set with genes most overlapped with those in $T$.

Since we propose three methods, thus we compared totally six methods : XGSEA-D, XGSEA-E, XGSEA-E±, $\text{HM}_1$, $\text{HM}_A$ and $\text{HM}_O$.

**Table 2.** Statistics of expression data and human gene sets ($\mathcal{T}^{(3)}$, where the cutoff for $p$-values was 0.01 and 0.05 for embryonic development and the others, respectively).

| | | | Gene expression | | Test sets in human | |
|---|---|---|---|---|---|---|
| Data set | Species | #genes | #samples | #labels | #sets | #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 | | |

## Data sets

To evaluate the performance of 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 2 shows the statistics of the four data sets.

- **Embryonic Development** (human and mouse): The two datasets were collected from www.ncbi.nlm.nih.gov/geo with accessing number GSE44183. Both gene expression datasets were obtained from single cell RNA sequencing. In the human dataset, there are 29 samples with 14,766 genes and seven embryonic development stages, 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 sixembryonic development stages, oocytes, pronucleus,2-cell, 4-cell, 8-cell and morula. These datasets were used in a cross-species study [18] already, while this study is not on GSEA.

- **Brain Cancer** (human and mouse): The datasets of the two species were downloaded from GEO with accession number GSE45874 and GSE38591, respectively. Both datasets were measured by microarray. The human dataset has 44,030 genes with six disease and six control samples, while the mouse dataset has 9,653 genes with four disease and four control samples. These datasets were also used in another cross-species study [19], while this study is also not on GSEA at all.

- **Ovarian Cancer** (human and mouse): The two Microarray gene expression datasets were downloaded from GEO with accession number GSE6008 and GSE5987, respectively. The human dataset has 21,188 genes with thirteen mucinous ovarian tumors and four control samples, while the mouse dataset has 45,101 genes with seven disease and four control samples. These datasets were also used in the cross-species study [19].

- **Melanomas** (human and zebrafish): The Microarray gene expression datasets of the two species were downloaded from GEO with accession number GSE83343 and GSE83399, respectively. The human dataset has $42,346$ genes with eight disease and four control samples, while the zebrafish dataset has $13,620$ genes with five disease and three control samples. These datasets were collected from two different studies [20, 21].

We then accessed Ensembl BioMart through http://www.ensembl.org/ [22] 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 [23, 24]. Fig 3 shows two homology matrices between human and mouse (left) and between human and zebrafish (right).

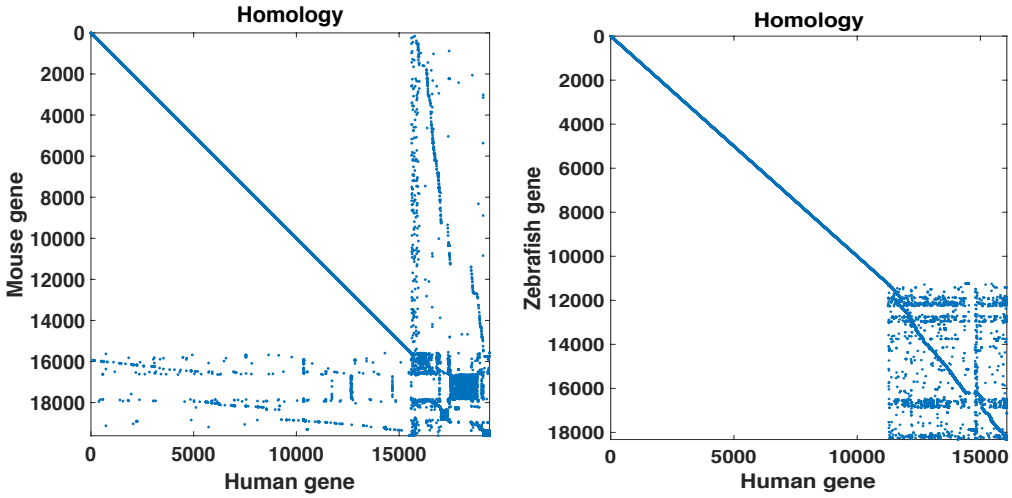


**Fig 3.** Homology relationships for (left) mouse-human and (right) zebrafish-human.

We can see that genes cannot be assigned in a simple one-to-one correspondences manner. We collected 674 human gene sets (pathways) from Reactome in Molecular Signatures Database (MSigDB), 2,250 mouse gene sets from http://baderlab.org/GeneSets and 1,550 zebrafish gene sets from http://bioinformatics.org/go2msig/.

## Experimental setting

In our experiments, we take human species as target species, and take mouse or zebrafish as the target 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 and 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 method, 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 $\mathcal{T}^{(1)} = \{T_1^{(1)}, \cdots, T_n^{(1)}\}$ , 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 have 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 $T_i^{(1)} \subset T_i$ 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 have 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 $g$.

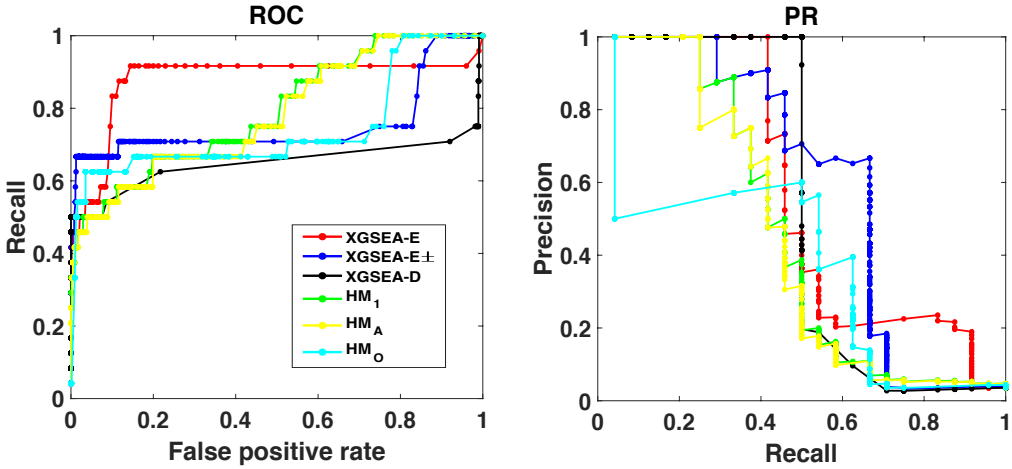


**Fig 4.** ROC and PR curves by XGSEA-D (black), XGSEA-E (red), XGSEA-E± (blue), $\mathrm{HM}_1$ (green), $\mathrm{HM}_A$ (yellow) and $\mathrm{HM}_O$ (light blue) for embryonic deveopment under $\mathcal{T}^{(3)}$.

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

## Evaluating XGSEA by supervised learning

Each target gene set has a ground-truth $p$-value. In evaluation, target gene sets with smaller true $p$-values should be predicted to have smaller $p$-values. In this light, we examined XGSEA and competing methods in a supervised manner: we set a cutoff (significance level) for the ground-truth $p$-values of target gene sets 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 control the number of positives (and negatives) by changing the cutoff. Then once after true positives (and true negatives) are determined by the cutoff for $p$-values in the above manner, we can examine the ROC (receiver operator characteristics) curve (and also precision-recall (PR) curve) by sorting the predicted $p$-values for gene sets in the ascending order. Note that this is regular validation of supervised learning (more precisely binary classification).

The $d$ and $\lambda$ 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.

### Performance on four real data sets

Fig 4 shows sample ROC and PR curves for one of the four data sets, i.e. embryonic development under $\mathcal{T}^{(3)}$ with the cutoff (for $p$-values) of 0.01. These figures shows that XGSEA (red and blue) look outperformed compared naive methods (green, yellow and light blue), except for XGSEA-D (black), indicating that regression of $p$-values on $p$-values directly may perform badly, as we expected. Note that there exist overlaps between XGSEA and naive methods, making the comparison unclear. Thus we checked the performance difference more systematically.

We changed the cutoff for $p$-values: $\{5e\text{-}1, 1e\text{-}1, 5e\text{-}2, 2.5e\text{-}2, 1e\text{-}2, 5e\text{-}3, 2.5e\text{-}3, 1e\text{-}3\}$, resulting in changing the number of true (ground-truth) positives. That is, the number of true positives becomes smaller for smaller cutoff values. Fig 5 (left column) shows, changing the cutoff for $p$-values, the AUC (area under the ROC curve) of all competing methods on all four real data sets under $\mathcal{T}^{(3)}$. The AUC increased as the cutoff was decreasing (the number of true positives was decreasing). For most of the changing cutoff values, XGSEA (black, red and blue)

showed better AUCs than the three baseline methods (green, yellow and light blue).

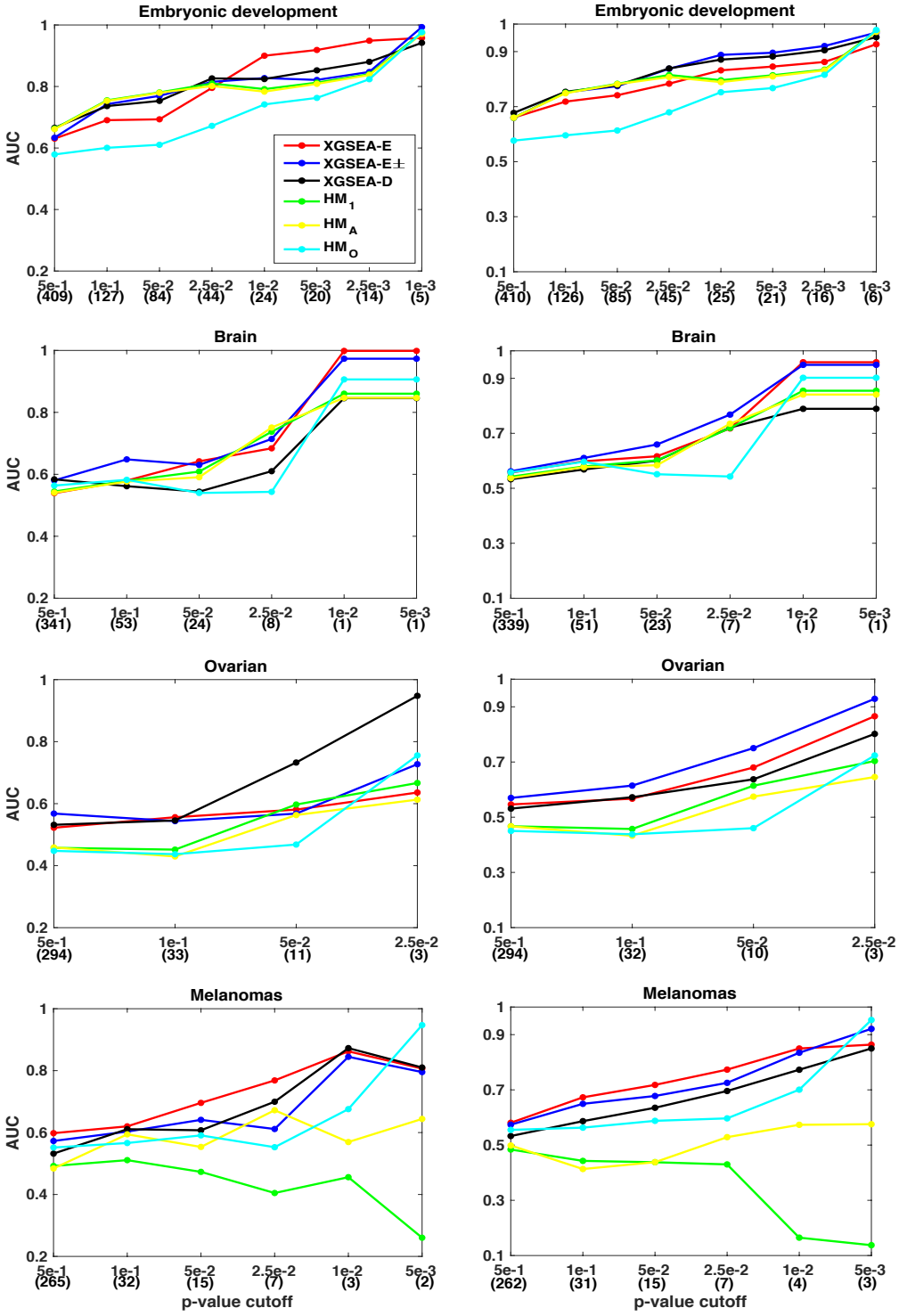


**Fig 5.** (Left column) AUCs on four data sets ($\mathcal{T}^{(3)}$), changing the cutoff for $p$-values. (Right column) Bootstrapped (20 trials) AUCs under the same condition as the left column. Compared methods are XGSEA-D (black), XGSEA-E (red), XGSEA-E± (blue), $\mathrm{HM}_1$ (green), $\mathrm{HM}_A$ (yellow) and $\mathrm{HM}_O$ (light blue).

**Table 3.** AUCs of six competing methods on four data sets and three target gene sets. 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 naive method is shown in brackets.

| 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 |

### Stabilized results by bootstrapping

Smaller cutoff values, such as $5e$-3, resulted in an extremely few number of positives. For example, brain cancer had only one positive for the cutoff of $5e$-3. Also each AUC (in the left column of Fig 5) was obtained by only one trial of training and test. These two aspects made AUCs in the left column of Fig 5 rather unstable. To resolve this issue, we conducted bootstrapping on 674 human gene sets of $\mathcal{T}^{(3)}$ by repeating sampling with replacement 20 times, resulting in 20 AUCs, which were averaged. Fig 5 (right column) shows the averaged AUCs (over 20 trials) of all methods on all four real data sets, under $\mathcal{T}^{(3)}$, changing the cutoffs for $p$-values. Comparing with the left column, the results were stabilized, clarifying the performance advantage of XGSEA (black, red and blue) over the three baseline methods (green, yellow and light blue). In particular, even the difference between the three proposed methods became clearer.

We then, fixing the cutoff value, examined the performance of the competing methods. 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 XGSEA significantly outperformed the baseline methods. For example, XGSEA-E± achieved the best in nine out of all 12 cases, followed by 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.

### Robustness against parameter value change

We examined the performance robustness of XGSEA, regarding parameter ($\lambda$) variation. Fig 6 shows AUCs of XGSEA-E under three gene sets ($\mathcal{T}^{(1)}$ (red), $\mathcal{T}^{(2)}$ (blue) and $\mathcal{T}^{(3)}$ (black)) on embryonic development and melanomas, when $\lambda$ is one of $\{1e\text{-}4, 1e\text{-}3, 1e\text{-}2, 1e\text{-}1\}$. This figure shows that AUC of XGSEA-E was rather stable within the given range, implying that the advantage over the baseline methods will be kept constantly.

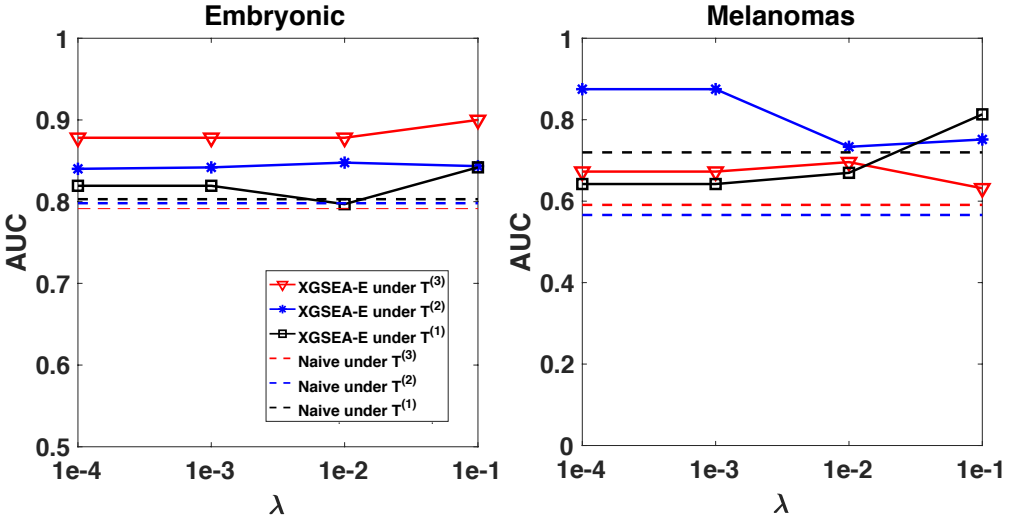


**Fig 6.** AUCs of XGSEA-E (solid line) and the best of naive methods (dotted line) under $\mathcal{T}^{(1)}$ (red), $\mathcal{T}^{(2)}$ (blue) and $\mathcal{T}^{(3)}$ (black) on (right) embryonic development and (left) melanomas.

**Table 4.** AUCs of XGSEA in variants and transferabilities, respectively, in embryonic development under gene set $\mathcal{T}^{(3)}$.

| | | XGSEA-D | XGSEA-E | XGSEA-E± |
|---|---|---|---|---|
| Similarity | MMD | 0.62 | 0.88 | 0.71 |
| | MMD+W | 0.61 | 0.88 | 0.74 |
| | MMD+B | 0.62 | 0.80 | 0.73 |
| | MMD+WB | 0.60 | 0.90 | 0.75 |
| Homology | $\boldsymbol{M}_{50}$ | 0.72 | 0.77 | 0.51 |
| | $\boldsymbol{M}_{500}$ | 0.73 | 0.78 | 0.53 |
| | $\boldsymbol{M}_{5000}$ | 0.75 | 0.84 | 0.73 |
| | $\boldsymbol{M}$ | 0.75 | 0.84 | 0.74 |

### 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 XGSEA, by modifying the objective function in the formulation of XGSEA. The objective function of XGSEA is given by (4), 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 (4), 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. (4).

We applied these four variants to embryonic development data with target gene set $\mathcal{T}^{(3)}$. Table 4 shows AUCs obtained with the cutoff (for $p$-values) of 0.01. From Table 4, MMD+WB (i.e. original (4)) achieved the best result for XGSEA-E and 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.

**Table 5.** 11 human pathways (with $p$-values) identified by 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 |

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 5,000 for this number of selecting 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 XGSEA over embryonic development data under gene set $\mathcal{T}^{(3)}$ to predict enrichment $p$-values.

Table 4 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 5,000, the performance was almost consistent with that of using the original $\boldsymbol{M}$, implying that interestingly 5,000 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 [25]. 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 [25], we ran XGSEA-E ($B$ = 100,000, $\lambda$=0.01 and $d$=5) to identify human pathways out of 1,960 Reactome pathways (downloaded from https://reactome.org/download-data).

Table 5 shows 11 human pathways identified by 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 [25].

The remaining 9 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 [26]. 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 [27]. Furthermore, the Notch pathway

**Table 6.** 20 human pathways (with $p$-values) identified by $\text{HM}_A$ for T cell dysfunction and reprogramming.

| Pathway | $p$-value |
|---|---|
| Assembly Of The HIV Virion | 4e-5 |
| Membrane binding and targetting of GAG proteins | 2e-4 |
| Mineralocorticoid biosynthesis | 4e-4 |
| Type II Na+/Pi cotransporters | 8e-4 |
| Interactions of Tat with host cellular proteins | 3e-3 |
| Biogenic amines are oxidatively deaminated to aldehydes | 5e-3 |
| SDK interactions | 0.01 |
| Cohesin Loading onto Chromatin | 0.01 |
| XAV939 inhibits tankyrase, stabilizing AXIN | 0.01 |
| Mitotic Telophase/Cytokinesis | 0.02 |
| Interleukin-9 signaling | 0.02 |
| TWIK-releated acid-sensitive K+ channel (TASK) | 0.02 |
| Defective Mismatch Repair Associated With MLH1 | 0.02 |
| Budding and maturation of HIV virion | 0.02 |
| CREB phosphorylation through the activation of CaMKK | 0.04 |
| Synthesis of 5-eicosatetraenoic acids | 0.04 |
| Apoptotic execution phase | 0.04 |
| NOTCH2 intracellular domain regulates transcription | 0.05 |
| Toxicity of botulinum toxin type C (BoNT/C) | 0.05 |
| CD209 (DC-SIGN) signaling | 0.05 |

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 [28]. 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 [29]. These facts support the reliability of the pathways identified by XGSEA.

On the other hand, we ran a naive approach, $\text{HM}_A$, over the same data, under the cutoff of 0.05, resulting in 20 pathways showed in Table 6. Although the number of pathways is larger than Table 5, these 20 pathways were diverse and less connected to the chromatin states, such as only two being related to Notch signaling pathways. This result implies that XGSEA-E would be more convincing than $\text{HM}_A$.

# Conclusion

We have defined XGSEP for promoting GSEA on species with scarce expression data, and proposed XGSEA with three steps, which can be simply: 1) GSEA, 2) domain adaptation, and 3) regression. Our empirical supervised validation over four real data sets revealed that XGSEA outperformed three naive approaches in AUC under various settings, particularly the advantage being proved statistically by bootstrapping and $t$-test. In the case study, mouse ATAC-Seq expression data is used to identify significant human pathways for T cell dysfunction and reprogramming. 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.

Improvement of XGSEA would be definitely interesting future work. It would be worth working on exploring a better variation on each of the three steps of 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 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.

# References

1. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545–15550.

2. Stuart, M J. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science. 2003;302(5643):249–255.

3. Zheng-Bradley X, Rung J, Parkinson H, Brazma A. Large scale comparison of global gene expression patterns in human and mouse. Genome Biology. 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):6987–6992.

6. Mestas J, Hughes CCW. Of Mice and Not Men: Differences between Mouse and Human Immunology. Journal of Immunology. 2004;172(5):2731–2738.

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, Schenkel JM, Odumade OA, Casey KA, et al. Normalizing the environment recapitulates adult human immune traits in laboratory mice. Nature. 2016;532:512–516.

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, Thomas. The storm has cleared: lessons from the CD28 superagonist TGN1412 trial. Nature Reviews Immunology. 2012;12:317–318.

11. Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Transaction on Knowledge and Data Engineering. 2010;22(10):1345–1359. doi:10.1109/TKDE.2009.191.

12. Huang J, Gretton A, Borgwardt KM, Schölkopf B, Smola AJ. Correcting Sample Selection Bias by Unlabeled Data. In: Schölkopf B, Platt J, Hoffman T, editors. Advances in Neural Information Processing Systems 19. Cambridge, MA: MIT Press; 2006. p. 601–608.

13. Pan SJ, Kwok JT, Yang Q. Transfer Learning via Dimensionality Reduction. In: AAAI 2008; 2008. p. 677–682.

14. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A Kernel Two-sample Test. Journal of Machine Learning Research. 2012;13:723–773.

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–i628. doi:10.1093/bioinformatics/btw428.

17. Edelman A, Arias TA, Smith ST. The Geometry of Algorithms with Orthogonality Constraints. SIAM Journal on Matrix Analysis and Applications. 1998;20(2):303–353.

18. Sun J, Jiang Z, Tian X, Bi J. A cross-species bi-clustering approach to identifying conserved co-regulated genes. Bioinformatics. 2016;32(12):i137–i146. doi:10.1093/bioinformatics/btw278.

19. Normand R, Du W, et al. Found In Translation: a machine learning model for mouse-to-human inference. Nature methods. 2018;15:1067–1073.

20. Filipp F, Li C, Boiko A. CD271 is a molecular switch with divergent roles in melanoma and melanocyte development. Sci Rep. 2019;9(1):7696.

21. Venkatesan A, Vyas R, Gramann A, Dresser K, et al. Ligand-activated BMP signaling inhibits cell differentiation and death to promote melanoma. J Clin Invest. 2018;128(1):294–308.

22. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols. 2009;4(8):1184–1191.

23. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, et al. Ensembl 2002: accommodating comparative genomics. Nucleic Acids Res. 2002;31:38–54.

24. Flicek P, et al. Ensembl 2014. Nucleic Acids Res. 2014;42(Issue D1):D749–D755.

25. Philip M, Fairchild L, Sun L, Horste EL, Camara S, Shakiba M, et al. Chromatin states define tumour-specific T cell dysfunction and reprogramming. Nature. 2017;545(7655):452–456.

26. Freddy R, H Robson M, Fabienne TC. Regulation of innate and adaptive immunity by Notch. Nature Reviews Immunology. 2013;13(6):427–437.

27. Tsukumo Si, Yasutomo K. Regulation of CD8+ T Cells and Antitumor Immunity by Notch Signaling. Frontiers in Immunology. 2018;9:101. doi:10.3389/fimmu.2018.00101.

28. Ende Z, Tomasz M, Ilona K, et al. Cancer mediates effector T cell dysfunction by targeting microRNAs and EZH2 via glycolysis restriction. Nat Immunol. 2016;17(1):95–103.

29. Catarina M, Vania NS, Pires AR, Paula M, Rui V M M, Sousa AE, 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. Journal of Immunology. 2014;193(12):5854–5862.